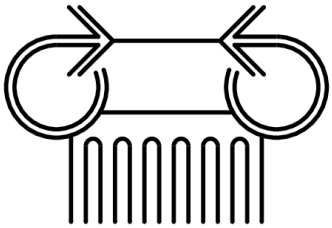
Tietojenkäsittelytieteen laitoksen julkaisuja
Department of Computer Science Series of Publications B
Raportti/Report B-2012-1

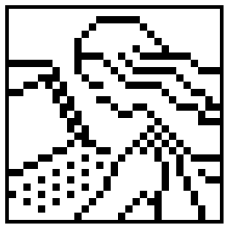
Yhdistetyt tietojenkäsittelyn päivät 2012

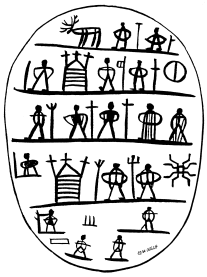
Federated Computer Science Event 2012

28-29.5.2012 Kumpula, Helsinki, Finland

Sasu Tarkoma, Joni-Kristian Kämäräinen, Tapio Pahikkala (toim./eds.)







Tukija / Supported by

NOKIA

Yhteystiedot / Contact information

Postiosoite / Postal address:
Department of Computer Science
P.O. Box 68 (Gustaf Hällströmin katu 2 b) FIN-00014 University of Helsinki
Finland

Sähköposti / Email address: info@cs.Helsinki.FI

URL: http://www.cs.Helsinki.FI/
Puhelin / Telephone: +358 9 1911
Faksi / Telefax: +358 9 191 51120

Tietojenkäsittelytieteen laitoksen julkaisuja
Department of Computer Science Series of Publications B
Raportti/Report B-2012-1

ISSN 1458-4786
ISBN 978-952-10-8024-1 (nid. / paperback)
ISBN 978-952-10-8025-8 (PDF)
Helsinki 2012
Unigrafia Oy

Esipuhe

Yhdistetyt tietojenkäsittelyn päivät (YTP) juhlistavat tietojenkäsittelytieteen ja tietotekniikan pitkää historiaa Suomessa. Päivät yhdistävät kolme alueen seuraa: Tietojenkäsittelytieteen seuran, Hahmontunnistuksen seuran, sekä Suomen tekoälyseuran. Tietojenkäsittelytieteen seuran osalta kuluvana vuonna tulee täyteen 30 vuotta seuran perustamisesta ja viidennettoista seuran järjestämät päivät. Tänä vuonna tulee täyteen myös 100 vuotta Turingin syntymästä. Päivien teemana onkin "Tietojenkäsittelyn historia ja tulevaisuus".

Päivien tavoitteena on tuoda yhteen kaikki tietojenkäsittelyn alueella toimivat ja siitä kiinnostuneet. Nyt yhdistetyssä muodossa päivät tarjoavat entistä paremman mahdollisuuden tietojenkäsittelyn eri alojen kohtaamiselle. Tietojenkäsittelytiede on varsin kansainvälinen tieteenala ja tama näkyy päivien järjestelyissä. Valtaosa esityksistä ja artikkeleista on englanninkielisiä. Tilaisuus pidetään Exactum-rakennuksella Kumpulan kampuksella, jossa Tietojenkäsittelytieteen seuran päivät pidettiin viimeksi 2006.

YTP:n ohjelma sisältää merkittävien tietojenkäsittelytieteilijöiden kutsuesitelmiä, väitöskirjakilvan ja pro gradu –kilvan voittajien palkintoesitelmät, tieteellisten artikkelien esitelmät, ja monipuolisen posteritilaisuuden. Posteritilaisuus järjestetään yhdessä HeCSE –tutkijakoulun kanssa. Tieteelliset artikkelit hyväksyttiin vertaisarviointimenettelyllä.

Kohokohtiin kuuluvat myös päivien illallinen Kalastajatorpalla, seurojen vuosikokoukset, ja keskustelut kahvitauoilla ja posteritilaisuuden aikana.

Kiitän kaikkia järjestelytoimikunnan jäseniä ja muita järjestelyyn osallistuneita. Toivotamme teidät lämpimästi tervetulleiksi Kumpulan kampukselle ja osallistumaan seurojen toimintaan.

Foreword

The Federated Computer Science Event is a two-day conference that combines the traditional annual events of three societies in Finland: The Finnish Society for Computer Science, the Pattern Recognition Society of Finland, and the Finnish Artificial Intelligence Society. The theme of the event is "History and Future of Computer Science". The event celebrates the 30th anniversary of the Finnish Society for Computer Science and the 100th birthday of Turing.

The event is held at the Kumpula campus of the University of Helsinki (UH), and it is organized by the Department of Computer Science of UH. The aim of the event is to get together people representing different areas of computer science. The federated nature of the event allows the meeting of people across the subdisciplines of computer science. The programme of the event consists of keynote presentations by distinguished scientists, presentations based on peer-viewed articles, poster and demo sessions, and the social event at Kalastajatorppa.

I thank all the members of the organizing committee and people who have helped in the organization. We welcome you to the event at Kumpula campus, and to participate in the societies.

May 2nd, 2012 in Helsinki
Sasu Tarkoma

Tietojenkäsittelytieteen seuran esipuhe

Tietojenkäsittelytieteen Seura ry perustettiin vuonna 1982 toimimaan tietojenkäsittelytieteen tutkijoiden ja alan tutkimustuloksista kiinnostuneiden henkilöiden yhdyssiteenä sekä edistämään tietojenkäsittelytieteen tutkimusta, tutkimustulosten soveltamista ja tutkimuksesta tiedottamista. Tarkoituksensa toteuttamiseksi Seura järjestää keskustelu- ja koulutustilaisuuksia jäsenilleen, harjoittaa tiedotustoimintaa ja on yhteistyössä muiden alan kotimaisten ja ulkomaisten järjestöjen kanssa. Lisäksi Seura voi osallistua kansainvälisten tapahtumien ja konferenssien järjestämiseen, tehdä esityksiä ja aloitteita sekä antaa lausuntoja ja jakaa stipendejä. Keskeisiä toimintakanavia ovat Tietojenkäsittelytieteen päivät sekä julkaisutoiminta, jota Seura toteuttaa Tietojenkäsittelytiede-lehden muodossa.

Seuran toiminnassa on viime vuosina ollut pyrkimyksenä kehittää tieteidenvälistä yhteistyötä. Nyt järjestettävien Tietojenkäsittelytieteen päivien tavoitteena on tuoda yhteen kaikki tietojenkäsittelyn alueella toimivat ja siitä kiinnostuneet. Yhteistyön laajentaminen tietojenkäsittelytieteen sisältöalueiden välille (ACM & IEEE 2008 linjaus) on myös tärkeä toiminnan tavoite, jota perinteisesti on vaalittu Seuran SIG-toiminnan kautta. Tieteidenvälisyys toteutuu nyt ensimmäistä kertaa yhdistetyssä muodossa toteutettavilla päivillä (YTP 2012), jotka järjestetään yhteistyössä Suomen hahmontunnistustutkimuksen seuran ja Suomen tekoälyseuran kanssa. Tämän yhteistyön tarkoituksena on tarjota mahdollisuus koota yhteen erityisesti tietotekniikan, tietoliikenteen ja tietojenkäsittelytieteen tutkijoita. Toivottavasti tulevina vuosina tämäntyyppistä yhteistyötä voidaan järjestää myös muiden tietojenkäsittelytieteen alaan sisällöllisesti kuuluvien seurojen kanssa.

YTP 2012 -päivät ovat myös esimerkki Seuran toiminnan täydentämisestä kansainvälistyvillä toimintatavoilla. Tämänkertaisten Tietojenkäsittelytieteen päivien esitelmistä ja artikkeleista suuri osa on englanninkielisiä. Päivien tarkoituksena on kuitenkin toimia suomalaisten yliopistojen tietojenkäsittelytieteen eri alojen tutkijoiden yhteisenä foorumina, jolla uusi tieto leviää tehokkaasti ja kansallinen alan yhteistyö lujittuu. Kiitän Tietojenkäsittelytieteen seuran puolesta Suomen hahmontunnistustutkimuksen seuraa ja Suomen tekoälyseuraa mainiosta yhteistyön avauksesta.

5.5.2012 Jyväskylässä

Tutkimusjohtaja Hannakaisa Isomäki
TKTS ry:n hallituksen puheenjohtaja 2011-2012

Hahmontunnistuksen seuran esipuhe

Suomen hahmontunnistustutkimuksen seuran (Hatutus) historia alkaa vuodesta 1977, jolloin yksi Suomen tunnetuimpia ja viitatuimpia tietotekniikan tutkijoita, akateemikko Teuvo Kohonen, tiedusteli eri tahoilta seuran perustamista silloin uuden alan, hahmontunnistuksen (engl. pattern recognition), ympärille. Allekirjoittanut, seuran nykyinen puheenjohtaja, oli silloin kolmevuotias ja käytti hahmontunnistusta lähinnä löytääkseen leluja hiekkalaatikolta. Lähes neljänkymmenen vuoden aikana ala on vain kasvattanut merkitystään.

Mitä insinöörit kutsuvat hahmontunnistukseksi, sitä usein tietojenkäsittelyn piirissä kutsutaan koneoppimiseksi (engl. machine learning) ja sen rooli on keskeinen sekä tekoälyssä (engl. artificial intelligence), tietokonenäössä (engl. computer vision), signaalinkäsittelyssä (engl. signal processing) ja monella muulla tietotekniikan/tietojenkäsittelytieteen osa-alueella. Luulisi että näiden alojen kasvun myötä meistä tietojenkäsittelyn ja tietotekniikan tutkijoista olisi tullut Suomen vahvin tutkijayhteistö, joka edistäisi alaamme kansallisilla ja kansainvälisillä foorumeilla - varsinkin kun Suomi on pieni maa. Mutta tässä yhdistysten luvatussa maassa me olemmekin pirstaloituneet kukin omien otsikoidemme alle. Voinkin ilolla todeta, että nyt päättäessäni puheenjohtajakauteni yksi tärkeimmistä tavoitteistani, seurojen ja tutkijoiden saattaminen yhteen ja toimimaan yhden tieteenalan puolesta, on ottanut ensimmäisen ja tärkeän askeleen Yhdistettyjen tietojenkäsittelyn päivien (YTP 2012) muodossa. Tästä suuri kiitos kahdelle muulle seuralle, Tietojenkäsittelytieteen seuralle ja Suomen tekoälyseuralle, jotka vastasivat kutsuun ja joiden aktiivit tekivät paljon minua suuremman työn asian eteen. Toivottavasti tämä johtaa entistä suurempaan ja vahvempaan yhteisöön, jossa tutkijat saavat uusia ideoita ja yhteistyökumppaneita myös oman erikoisalansa ulkopuolelta. Joukossa on voimaa ja toivottavasti YTP 2012 -konferenssi muistetaan uuden aikakauden aloittajana!

2/5/2012 Lappeenrannassa

Prof. Joni-Kristian Kämäräinen

Foreword of the Finnish Artificial Intelligence Society

Welcome to the Federated Computer Science Event (YTP 2012) that combines the traditional annual events of the Finnish Society for Computer Science, the Pattern Recognition Society of Finland, and the Finnish Artificial Intelligence Society. This is also the 15th Finnish Artificial Intelligence Conference (Suomen tekoälytutkimuksen päivät, STeP 2012) event. The first STeP 1984 was held at the Helsinki University of Technology (TKK) in August 20-23, 1984. It has been organized regularly every two years ever since. Federated Computer Science Event will be held at the Kumpula campus of the University of Helsinki (UH) on May 28-29, 2012, and it is organized by the Department of Computer Science of UH.

In this event Finnish Artificial Intelligence Society has decided to give the dissertation award for Arto Klami for his thesis "Modeling of Mutual Dependencies", and the grant for António Gusmão for his master's thesis "Reinforcement Learning In Real-Time Strategy Games".
I am grateful to all the active researchers who have submitted their contributions to the conference. I thank the Pattern Recognition Society of Finland for inviting us for the joined event; Tapio Pahikkala, Antti Airola, and Jaakko Väyrynen of the organizing committee for arranging the event from Finnish Artificial Intelligence Society side; Pentti Haikonen for his great XCR-1 robot demonstrations; and Tapani Raiko for his constant help with the arrangements.

Helsinki, May 6th, 2012

Jukka Kortela

Chairman, Finnish Artificial Intelligence Society

Sisällys

Contents

Ohjelmatoimikunta / Program committee

Puheenjohtaja (chair) prof. Sasu Tarkoma, Helsingin yliopisto
Prof. Pekka Orponen, Aalto-yliopisto
Prof. Joni-Kristian Kämäräinen, Lappeenrannan teknillinen yliopisto
Dr. Lea Kutvonen, Helsingin yliopisto
Dr. Pirjo Moen, Helsingin yliopisto
Dr. Tapio Pahikkala, Turun yliopisto
Fil. lis. Tiina Niklander, Helsingin yliopisto
Dr. Sini Ruohomaa, Helsingin yliopisto

Hecse–tutkijakoulun yhteyshenkilöt

Prof. Petri Myllymäki, Helsingin yliopisto
Dr. Petri Kontkanen, Helsingin yliopisto

Toimikunnan lisäksi tieteelliseen arviointiin osallistuneet / Additional reviewers

Antti Airola, Turun yliopisto, TUCS
Pentti Haikonen, University of Illinois at Springfield
Markus Koskela, Aalto-yliopisto
Esa Rahtu, Oulun yliopisto
Tapani Raiko, Aalto-yliopisto
Jaakko Väyrynen, Aalto-yliopisto

Juhlaesitykset

Heikki Mannila, professori, Suomen Akatemian pääjohtaja
Jorma Rissanen, prof. emer.

Väitöskirjapalkinnot

Tietojenkäsittelytieteen seura
Satu Jumisko-Pyykkö, Tampereen teknillinen yliopisto

Hahmontunnistuksen seura
Antti Airola, Turun yliopisto

Suomen tekoälyseura
Arto Klami, Aalto-yliopisto

Pro gradu –palkinnot

Tietojenkäsittelytieteen seura
Joel Rybicki, Helsingin yliopisto
Ville Kangas, Lappeenrannan teknillinen yliopisto (kunniamaininta)
Joris Kinable, Aalto-yliopisto (kunniamaininta)
Suomen tekoälyseura
António Gusmão, Aalto-yliopisto

Keynotes

Heikki Mannila, professor, President of the Academy of Finland
Jorma Rissanen, prof. emer.

Doctoral Dissertation Awards

The Finnish Society for Computer Science
Satu Jumisko-Pyykkö, Tampere University of Technology

The Pattern Recognition Society of Finland
Antti Airola, University of Turku

The Finnish Artificial Intelligence Society
Arto Klami, Aalto University

MSc Thesis Awards

The Finnish Society for Computer Science
Joel Rybicki, University of Helsinki
Ville Kangas, Lappeenranta University of Technology (honorable mention)
Joris Kinable, Aalto University (honorable mention)
The Finnish Artificial Intelligence Society
António Gusmão, Aalto University

Award Summary Articles

User-Centered Quality of Experience and Its Evaluation Methods for Mobile Television

Summary of PhD thesis [1]

Satu Jumisko-Pyykkö
Tampere University of Technology
satu.jumisko-pyykko@iki.fi

ABSTRACT

In order to verify the user's satisfaction of the quality of a system or its components under development it is essential to evaluate quality of experience. The aim of this thesis is two-fold; 1) To understand what the components of experienced quality are and how these components affect experienced quality, and 2) To develop user-centered quality evaluation methods for examining experiences of multimedia quality. The thesis contains eleven extensive quality evaluation experiments and a literature review. The experiments were carried out for mobile television and mobile three-dimensional television with a relatively low quality level at a time when the systems were not available on the consumer market. More than 500 naïve evaluators (mostly non-students) participated in the experiments carried out in controlled laboratory and quasi-experimental field circumstances using hybrid data-collection methods containing quantitative quality excellence evaluation, qualitative descriptions of quality, observation and advanced techniques for situational data-capture. The audiovisual system parameters varied with respect to the level of content, media and transmission. The systematic literature review of over 100 high-quality papers clarified the components of use contexts for mobile-human-computer interaction.

The descriptive model of User-Centered Quality of Experience (UC-QoE) and the evaluation methods developed summarize the outcome of the work. UC-QoE is constructed from four main components: the user's characteristics, the system's characteristics, the context of use and the experiential dimensions. According to the results, contrary to earlier understanding, quality of experience is a broader phenomenon than sensorial excellence of a system component, and therefore its evaluation and design needs to consider the components surrounding it. The methodological contribution has five parts: 1) a holistic framework for User-Centered Quality of Experience evaluation, 2) Bidimensional method for assessing quantitatively the domain-specific acceptance threshold, 3) Experienced quality factors - interview-based descriptive method, 4) Open Profiling of Quality, as an advanced mixed method, that combines quantitative quality evaluation and qualitative descriptive quality evaluation based on an individual's own vocabulary, and 5) Hybrid method for quality evaluation in the context of use. These methods are concrete tools for practitioners to conduct quality evaluation experiments within the framework presented. Beyond this fundamental and applied research contribution, this thesis supports user-centered development of novel mobile multimedia systems for providing a better user experience in the long term.

1. INTRODUCTION

Television has a significant role in everyday life. On average more than 2.5 hours are spent daily by viewing moving pictures via different devices (e.g. [2]). To provide a more and more pleasurable viewing experience, the television technology has gone through an evolution since it was established in 1926. There has been an evolution of quality from black-and-white to color images, an increase in screen sizes and digitalization. This evolution is expected to continue towards improvements in depth (3D). To measure the excellence of video quality of television, International Telecommunication Union (ITU) has provided well-validated test methodologies for more than 30 years [3].

The revolution of personal and mobile computing also had its effect on the emergence of television and video. Broadcast television on mobile devices became a dream of the mass medium and the system providers to offer ubiquitous viewing possibilities for customers. New challenges were set to video quality not only because of the small display size but also because of the necessity to combine a huge amount of data with a wireless transmission channel and wireless reception, computational power and battery life time. This requires a high-level of optimization in the multiple stages of the system. The further development from 2D to 3D mobile television and video is a highly expected next step in the development. To create value for the end-users, their needs and requirements for quality have to be fulfilled. The change in the technological context also requires new ways of evaluating quality and taking into account the challenges of ubiquitous usage.

Videolization - the dramatic change in the availability and consumption of video - has taken place during the course of this study (2005-2011). It covers the shift from analogue to digital television, the introduction of TV over the internet (IPTV), and videos as a part of online newspaper editing. Parallel to the accessibility of professionally created content via different devices, user -created content has become available. For example, since the opening of YouTube in 2005 it has been subscribed by millions of people. These videos introduced highly compressed, impaired and low video quality to the users. Furthermore, video captures become a basic function of digital cameras and multimedia mobile phones within the course of this study. Videolization has made the consumers familiar with the range of the different digital video qualities presented on different devices as a part of their video consumption.

Related work - To quantify the experienced quality of certain system components, and to optimize them or predict their quality automatically, subjective evaluation experiments are conducted (e.g.[4]). The mainstream view on the concept of experienced quality is strongly formulated through the recommendations of International Telecommunication Union, which are widely spread among the engineering quality evaluation society. The current

mainstream approach to quality is the following: 1) Perceived quality is examined only quantitatively on the sensorial level, favoring studies of one modality or on a certain piece of system at a time. 2) Assessment is conducted in a highly controlled environment (e.g. the requirements for non-functional system components are derived from perceptually perfect conditions), even though the final application is assumed to be used in heterogeneous mobile contexts. 3) The background of evaluators or users does not have or has only a small impact on quality evaluations. 4) Quality evaluation is not connected to the use of the final multimedia application. Although the current approach has benefits in aiming at maximizing the high-level control in the examination of the causal effects and in serving the needs of identifying trade-offs between a limited set of system components in their development, its view on experienced quality is limited and strongly system-centric.

The principles of the existing approaches are highly in contradiction to what is known about perception in psychology and human-computer-interaction e.g. [5]-[9]: 1) Human perception includes high-level cognitive processing in which emotions, attitudes, knowledge and the context are part of the active interpretation of perception. 2) Multimodal perception is adaptive, flexible, and different from a simple sum derived from two perceptual channels separately. 3) The final user experience of an application is characterized by factors from the user, the system or service and its context of use and is described by the different type of experiential influences and consequences. These approaches emphasize a broad or holistic understanding of human perception and experiences, and a pragmatic view when utilizing this information at different stages of design and evaluation processes.

2. OBJECTIVES AND SCOPE

This thesis has two main research goals**.** The first aim of this thesis is **to understand what the components of experienced quality are and how these components impact on experienced quality.** The outcome is a descriptive model of User-Centered Quality of Experience. The second aim is **to develop user-centered quality evaluation methods for examining experienced quality.** The outcome is a research methodology for user-centered multimodal quality evaluation for video on mobile devices. Within the methodology, emphasis is given to quality evaluation in the context of use, descriptive quality, and measurements of minimum quality levels that are useful.

Scope. Nature of this thesis is multidisciplinary. It primarily belongs to the research field of human-computer interaction (HCI) and secondarily to field of multimedia which covers the various aspects of multimedia systems and technology, signal processing and applications. Empirical work has been conducted in multidisciplinary research teams. In more detail, the scope of this thesis is to evaluate the low produced qualities of critical system components in the next generation multimedia services under the viewing task on mobile devices. At the time of conducting the studies, 2D/3D mobile video and television were considered as next generation products. There were no similar systems available on the market, they were not adopted by the users, and the related technologies and standards were still highly maturing. The term critical system component refers to the part of the whole system that can have a negative impact or prohibit the utility of the whole system from user's point of view. Mobile (3D) TV is a service that is capable of receiving, reproducing and distributing (stereoscopic) video and audio content through different networks and that can be used via a pocket sized mobile device (adapted from [10]). In mobile 2D/3D television under the broadcasting scenario, multimedia processing is extremely demanding requiring a high-level optimization in multiple stages of the system from capturing content, coding, transmission, presentation on display. This can result in independent or jointly occurring noticeable impairments or artefacts in the presentation of content (Overview, [11]). The term low quality characterizes a multimedia presentation which can contain perceived noticeable impairments and the viewing or listening conditions are limited (e.g. small screen size), and the term makes a distinction to perceptually impairment-free high-qualities (e.g. top-end multi channel audio, or high-definition visual presentation). One aim is at ensuring that the experienced quality of critical system components, developed in isolation from the other components of product constitutes no obstacle to the wide audience acceptance of a product or service. From the system perspective, non-functional system-components are the focus. Furthermore, this thesis focuses on the user assessment of the quality while viewing content because viewing is the most important phase in video content use. The user's interactive tasks prior to and during the viewing with a device are out of the scope of this thesis.

In the thesis, I understand quality to contain three different characteristics: It is 1) an integrated set of perceptions of overall excellence, 2) composed of distinctive perceptual attributes, 3) user's perception of the degree to which the user's requirements have been fulfilled.

3. RESEARCH METHODS

The thesis contains twelve extensive quality evaluation experiments and a literature review. The experiments were carried out for mobile television and for mobile three-dimensional television with a relatively low quality level. Each of the experiments has 30-75 naive participants (non-students) forming a broad pool of data from over 500 participants. The experiments were conducted in the controlled laboratory and field circumstances using hybrid data-collection methods containing quantitative quality excellence evaluation, qualitative quality descriptions, and advanced techniques for situational data-capture. The audiovisual system parameters varied on the level of content (content types), media (presentation modes, bitrate, framerate, error concealments) and transmission (MFER error rates). The literature review defined the framework for the problem of the thesis and the central concept of context of use for quality evaluation studies in the field.

The results of these studies are published in 12 scientific publications (5 in journals, the rest in the conferences). The candidate is the first author in 10 publications and has a significant contribution in all papers. In addition, the candidate has 16 supplementary publications in the theme of her thesis.

4. RESULTS AND CONTRIBUTION

This thesis provides both fundamental and applied research contributions. The descriptive Model of User-Centered Quality of Experience (UC-QoE) and the evaluation methods developed summarize the main outcomes of this thesis.

User-Centered Quality of Experience

Based on the literature review and the results of the studies conducted for the thesis, the descriptive model of User-Centered Quality of Experience (UC-QoE) was constructed. It contains four

main components: the user's characteristics, the system's characteristics, the context of use, and the experiential dimensions (overview to model Figure 1).

The user's influence on the quality of experience was characterized by several demographic and psychographic variables underlining the active nature of human perception at the sensorial, emotional, attitudinal and cognitive levels.

The influence of the system quality factors to experienced quality depend on their perceptual characteristics, modalities and the overall quality level. Multimodal quality of experience is composed of both audio and video quality, and their relative importance can vary between quality levels, content and impairment types. For example, on the good quality level, influent audio is found annoying. Experienced quality is also unequally influenced by impairment types; temporally dominating accountable and detectable cut offs in audio and/or video have a strongly interruptive nature towards the user's viewing task. Experienced quality between monoscopic and stereoscopic video reflects a hierarcial structure. Experienced quality of 3D video on a small screen can improve the viewing experience if the level of visible impairments is low; otherwise, a monoscopic presentation mode can provide better experience. The ease of viewing (ability to maintain optimal viewing conditions and focus on content), visibility of objects, good spatial quality, natural and impairment-free depth are central requirements for 3D video on a small screen and, visual discomfort can be part of the experienced quality of stereoscopic video.

According to the descriptive attributes, experienced quality is constructed of the interpreted characteristics of video (audio, visual, audiovisual, content) and the components of viewing experience and use (e.g. the task, ease of viewing, visual comfort and user's relation to content). These confirm that quality perception is an active process which goes beyond apparent features of produced quality and is tightly related to action-related properties [6]. The descriptive quality model for mobile 3D video, containing the attributes and a vocabulary, provides a timely guide for the development and evaluation of upcoming systems.

Finally, the quality requirements drawn in conventional controlled conditions were more easily detected and less appreciated compared to the requirements in the natural context of use with variable physical and social distractions and actively divided attention. These studies also highlight use-related aspects, not only quality.

Taken together, quality of experience is a broader phenomenon than the sensorial excellence of the system component, as it was earlier understood, and therefore its evaluation and design need to consider the components surrounding it.

Further research 1) needs to clarify the influence of the user's characteristics on quality domain more specifically and over several studies, 2) examine the joint influence of the independent components by aiming at maximizing the several independent components (users, system, context of use) in comparison to the conventional quality evaluation approach. In this way, 3) the utility of the approach presented can be estimated compared to an existing one, and 4) the relation between the experience of system components and a holistic user experience can be addressed. Further work also needs to consider novel ways of 5) modeling the quality of experience utilizing the components presented and a descriptive quality model, as well as 6) designing context-aware solutions utilizing the level of quality and multimodality, visual presentation modes and context characteristics. Finally, long-term consequences of quality of experience are worth sketching.

Evaluation methods

The methodological contribution of this thesis has five parts:

1) The holistic framework was developed to give an overview of the factors and techniques essential to its evaluation. It underlines the external validity in the selection of the users, the system parameters and contents, the context of evaluation as well as a multi-methodological assessment to connect quality evaluation to the expected use. The increase of realism and a deeper understanding of experienced quality can result in more expensive studies (i.e. more complex designs, increased time for planning and analyzing), underline the locality and external validity of the results over high-level of control (e.g. test materials, circumstances) of the conventional approach.

2) Bidimensional research method of acceptance was developed for identification of minimum useful level of quality for use of a certain application as a part of quantitative quality evaluation. The method is beneficial when connecting quality to the expected use of the (novel) system and it is applicable to stimuli on the variable quality levels. The use of the method is estimated to slightly increase the cost of the study in the phase of analysis compared to existing retrospective one-dimensional measures.

3) Experienced quality factors – the interview-based descriptive method – is a flexible method that provides a fast data-collection procedure combined with psychoperceptual excellence evaluation in different circumstances. The main benefits of the method are an ability to explain the results of excellence evaluation and build fundamental understanding of phenomenon (e.g. components of quality), but it may also help to identify design ideas. Accuracy of the results, acquired using only the free-description task, is limited with regard to a certain individual stimulus. The method can be applied in the evaluation of novel and heterogeneous stimuli with naïve participants.

4) Open Profiling of Quality is an advanced mixed method which combines quantitative quality evaluation and qualitative descriptive quality evaluation based on an individual's own vocabulary in a multi-step data-collection procedure. It is applicable in evaluation of novel and heterogeneous stimuli with naïve participants. The method requires a multi-step data-collection procedure as vocabulary-based methods in general. It provides rich and tightly stimuli connected data that enables use of broad set of techniques of analysis, and can offer complementation and convergence between the quantitative and qualitative results.

The methods 3-4 stress the understanding of descriptive quality attributes as a part of the evaluation of complex and heterogeneous stimuli.

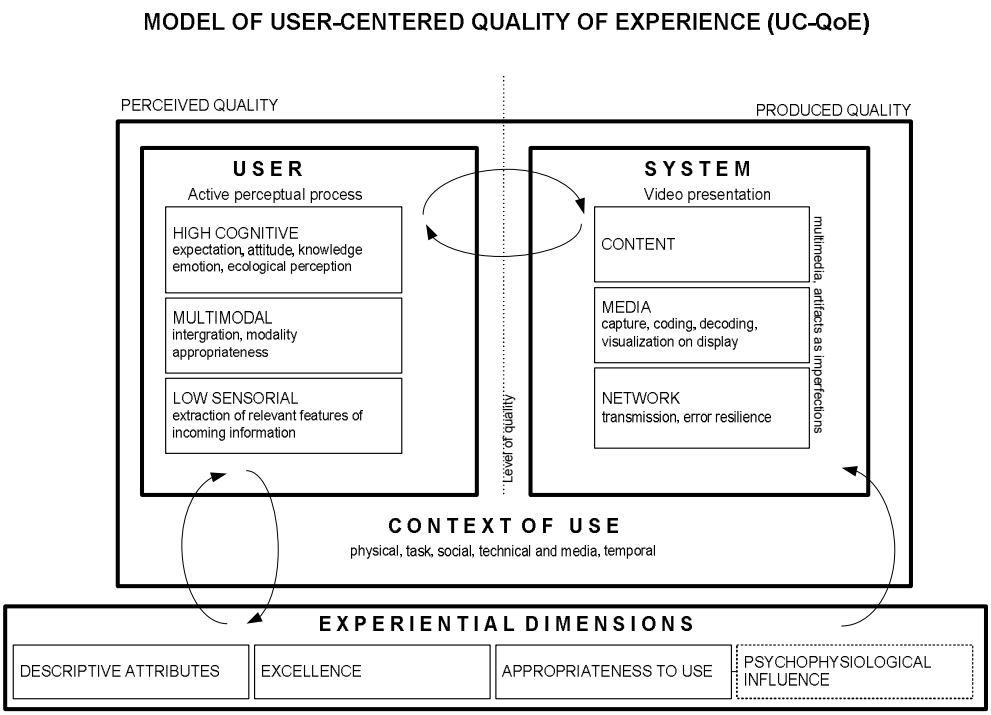


Figure 1 The model of User-Centered Quality of Experience for Mobile (2D/3D) Television

5) Hybrid method of quality evaluation in the context of use is a tool for quasi-experiments conducted in natural circumstances (e.g. viewing mobile television while travelling by bus). It contains a procedure for planning, data-collection and analysis, an identification of the situational characteristics surrounding quality evaluation on the macro and micro levels, the use of several techniques through the study. The main benefits of the method are the ability to characterize the contextual quality requirements, extend the quality evaluation towards the use, and improved ecological validity. As the quasi-experiments in general are relatively demanding to design and carry out, it is currently proposed to conduct these experiments to complement the laboratory experiments in a sequential workflow with a limited set of stimuli.

The methods presented vary in the levels of details and they are partly related. **The methods are concrete tools for practitioners to conduct quality evaluation experiments within the framework presented and they have also contributed to the standardization activities of the quality of experience evaluation** [12][13].

Beside this main contribution, the model of context of use for mobile HCI was developed to clarify the central concept of the context of use, its components, subcomponents and properties, based on the systematic literature review of over 100 high-quality papers. **The model can help both practitioners and academics to identify broadly relevant contextual factors when designing, experimenting with, and evaluating in mobile contexts of use.**

Future work for methodological development needs to focus on extensive between-method comparisons for qualitative and mixed methods to increase the awareness of their benefits, the applicability and limitations to guide practitioners to use them, and finally to support safe long-term development of these methods. Secondly, future work needs to create a collection of well validated tools to quantify the user's relation to content, multimodal quality and a system or service. In the long term, the use of these tools can help to understand the most influential individual differences and build up user profiles over the studies. Thirdly, to build up a more complete picture of the experiential aspects of quality of experience, further work needs to examine the relation between subjective and objective (psycho-physiological) quality evaluation methods.

REFERENCES

[1] Jumisko-Pyykkö, S. 2011. User-Centered Quality of Experience and Its Evaluation Methods for Mobile Television. PhD Thesis. Tampere University of Technology. http://satujumiskopyykko.net/wb/media/PhDThesis/PhDThesisJumiskoPyykko_User-CenteredQualityOfExperience.pdf

[2] Nielsen. (2010). How People Watch: A Global Nielsen Consumer Report. August 2010, http://no.nielsen.com/site/documents/Nielsen_HowPeopleWatch_August2010.pdf, retrieved: 26.11.2010.

[3] ITU-R BT.500-11 Recommendation. (2002). Methodology for the subjective assessment of the quality of television pictures. International Telecommunications Union (ITU) – Radiocommunication sector.

[4] Engeldrum, P. (2000). *Psychometric Scaling: A Toolkit for Imaging Systems Development*. Winchester, Mass: Imcotek Press.

[5] Neisser, U. (1976). *Cognition and Reality, Principles and Implications of Cognitive Psychology*. San Francisco: W.H. Freeman and Company.

[6] Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin, Lawrence Eribaum.

[7] Stein, B. E., London, N., Wilkinson, L. K., & Price, D. D. (1996). Enhancement of perceived visual intensity by auditory stimuli: A psychophysical analysis. *Journal of Cognitive Neuroscience, 8*, 497–506.

[8] Mahlke, S. & Thüring, M. (2007). Studying antecedents of emotional experiences in interactive contexts. *Proceedings CHI 2007*, 915–918.

[9] Hassenzahl, M., & Tractinsky, N. (2006). User experience – A research agenda. *Behaviour and Information Technology, 25(2)*, 91–97.

[10] Oksman, V., Ollikainen, V., Noppari, E., Herrero, C., & Tammela, A. (2008). 'Podracing': Experimenting with mobile TV content consumption and delivery methods. Multimedia Systems, 14(2), 105–114.

[11] Boev, A., Hollosi, D., Gotchev, A., & Egiazarian, K. (2009). Classification and simulation of stereoscopic artefacts in mobile 3DTV content. *Electronic Imaging Symposium 2009, Stereoscopic Displays and Applications*.

[12] Jumisko-Pyykkö, S., & Utriainen, T. (2011). Hybrid Method for Multimedia Quality Evaluation in the Context of Use Contribution*, International Telecommunication Union, Q13/12, Study group 12*

[13] Strohmeier, D. & Jumisko-Pyykkö, S. Proposal on Open Profiling of Quality as a mixed method evaluation approach for audiovisual quality assessment Contribution, *International Telecommunication Union, Q13/12, Study group 12*

Machine learning and performance estimation methods for ranking problems

Antti Airola*
Department of Information Technology
University of Turku

Abstract

The task of learning to rank refers to the machine learning problem, where the aim is to infer from past observations a ranking model that can order new objects according to how well they match some underlying criterion. Ranking problems are commonly encountered in applications such as document retrieval, game playing, information extraction and recommender systems. While learning to rank has been a topic of active research for more than a decade, developing scalable learning methods, and reliable and efficient validation methods has proven to be challenging.

The doctoral thesis of the author, summarized in this article, provides the following main contributions towards solving these issues. First, novel training algorithms based on optimizing a pairwise criterion in the regularized risk minimization framework are derived. Previously, the most well established method of this type is the ranking support vector machine (RankSVM). The introduced RankRLS method, as well as the proposed improvements to RankSVM, lead to orders of magnitude gains in efficiency, without decrease in predictive performance. Second, novel cross-validation approaches are proposed in order to account for the data dependencies and multivariate performance measures characteristic of ranking tasks. Computational short-cuts allow the efficient computation of these estimates for the RankRLS method. Finally, an application study introducing a novel method for information extraction from biomedical text combines several key ideas of the thesis, resulting in a state-of-the-art solution to the problem.

Keywords: cross-validation, information extraction, kernel methods, learning to rank, machine learning, regularized least-squares, regularized risk minimization, support vector machine

1 Introduction

In learning to rank, the aim is to infer from previously collected data a ranking function, that is able to order new sets of objects according to how well they match the underlying ranking criterion. Learning is necessary in such applications, where we do not know the true underlying ranking function, but rather have access to previous judgements made by some actor, typically a human being. Two typical examples of such applications are search engines that rank documents according to their match to user queries [Joachims 2002] and recommender systems [Minkov et al. 2010]. For an overview of the problem domain and related work we refer to [Liu 2009; Fürnkranz and Hüllermeier 2010; Airola 2011].

We assume the availability of training data, which contains both feature representations and judgements related to objects from the application domain of interest. These previous judgements may be categorical (e.g. good/bad, 1-5 stars) or supplied as real-valued utility scores, where a higher score indicates higher rank than a lower one. More generally, the information may be provided in terms of pairwise comparisons, indicating that certain objects are preferred over other ones. Based on the training data we learn a scoring function, that maps the feature representation of any given object to a predicted utility score. When ranking a new set of objects, the ranking is constructed by sorting the objects according to predicted scores. An accurate ranking function is such that the produced rankings match well the true rankings also for such sets of objects that were not observed in the training set. Finding such a function requires striking a balance between the phenomena of underfitting and overfitting, as learning may fail either due to considering too simple hypotheses (e.g. linear models for a highly non-linear concept) or due to allowing too rich set of hypotheses and ending up simply modeling the noise in the training data.

The family of regularized kernel methods embodies one of the mainstream approaches to machine learning [Schölkopf and Smola 2002; Shawe-Taylor and Cristianini 2004]. These methods allow the use of structured data and non-linear modeling, and offer principled ways to dealing with both the underfitting and overfitting phenomena, while still leading to convex optimization problems, where globally optimal solutions can be found. Widely used kernel methods include the support vector machine (SVM) [Vapnik 1995] and the regularized least-squares (RLS) [Poggio and Smale 2003] algorithms. The ranking support vector machine (RankSVM) method extends standard SVMs to learning to rank by casting the problem as a binary classification problem over pairs of objects. While the method has been demonstrated to achieve excellent ranking performance, the training methods proposed in [Herbrich et al. 1999; Joachims 2002] unfortunately lead to solving optimization problems whose size may depend quadratically rather than linearly on the size of the training set. For linear RankSVM more efficient training methods are known [Joachims 2006; Chapelle and Keerthi 2010], but these are limited to settings where the number of possible ranks in the data can be assumed a small constant.

Cross-validation is one of the most widely used techniques in machine learning for estimating the predictive power of the learned models. However, standard cross-validation approaches such as the leave-one-out method turn out to be highly unreliable in many ranking settings. The assumption that the training data is sampled independently is routinely broken, leading to biased estimates (see e.g. [Pahikkala et al. 2012b]). The use of multivariate ranking performance measures, such as the area under the ROC curve (AUC) and its generalizations leads to problems when predictions made on different rounds of cross-validation are combined together [Parker et al. 2007; Forman and Scholz 2010]. Further, straightforward implementations of cross-validation procedures also incur high computational costs, due to the necessity to re-train a learning algorithm multiple times.

The thesis [Airola 2011] summarized in this article provides a number of contributions towards solving the aforementioned problems. RankRLS [Pahikkala et al. 2009] is a novel learning to rank method, that combines regularized risk minimization with a pairwise least-squares loss function. This choice leads to a closed-from solution expressed as a system of linear equations, that can be solved efficiently even though the loss is implicitly computed over all pairs in the training set. Further, using matrix update formulas, the regularization parameter can be selected, and exact cross-validation estimates can be computed, at the same asymptotic cost as training RankRLS once. [Airola et al. 2011b] introduces an improved version of the linear RankSVM training method of [Joachims 2006] reducing the worst-case quadratic computational cost to $O(m \log(m))$ scaling by applying self-balancing search

*e-mail:antti.airola@utu.fi

trees. [Airola et al. 2011a] considers the use of Nyström approximation for generating low-dimensional feature representations for training kernel machines, and how to efficiently re-compute these during cross-validation, in cases when there are dependencies present in the data. In [Airola et al. 2011c] we consider the problem of cross-validation when applying pairwise performance measures, demonstrating that the proposed leave-pair-out procedure provides almost unbiased performance estimates with low variance. In [Airola et al. 2008] an application in protein-protein interaction extraction from scientific articles combines a number of ideas, later improved and refined in the aforementioned works, in order to achieve computational efficiency and reliable performance estimates.

2 Learning to rank

2.1 Regularized risk minimization

Let the input space $\mathcal{X}$, and output space $\mathcal{Y}$ be sets. We are supplied with a training set Z containing inputs, and associated label information, defined as $Z = (X, Y) \in \mathcal{X}^m \times \mathcal{Y}$. By $X = (x_1, \ldots, x_m) \in \mathcal{X}^m$ we denote the set of m inputs belonging to the training set. By $Y \in \mathcal{Y}$ we denote a structured object containing the label information associated with X. In learning to rank, in the simplest case $\mathcal{Y} = \mathbb{R}^m$, meaning that each input is associated with one utility score. More generally, the labels may have dependencies between them, or be associated with pairs of inputs rather than with individual inputs.

The learning algorithm takes as input a finite training set Z, and outputs a scoring function $f : \mathcal{X} \to \mathbb{R}$, which aims to model the dependency between the inputs and the labels. Let $X \in \mathcal{X}^m$, $m \in \mathbb{N}$ be a sequence of inputs. Then by $f(X) \in \mathbb{R}^m$ we denote the vector of predictions for this sample. A loss function

$$l : \bigcup_{m \in \mathbb{N}} \mathbb{R}^m \times \mathcal{Y} \mapsto [0, \infty)$$

measures how well the predicted labels and true labels for a data set match. The goal of learning is to find such a scoring function that would incur minimal expected loss on data drawn from the same distribution from which the training data was sampled from. In practice we can never compute the expected loss, but are rather limited to using the estimate known as the empirical risk

$$R(f) = l(f(X), Y),$$

that is simply the loss computed on the training set.

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a a finitely positive semi-definite kernel function. In the kernel methods framework, we consider hypotheses of the type

$$f(x) = \sum_{i=1}^{m} \alpha_i k(x, x_i),$$

where $\alpha_i \in \mathbb{R}$. In the special case where $\mathcal{X} = \mathbb{R}^n$ and $k = \langle \cdot, \cdot \rangle$ is the standard inner product in $\mathbb{R}^n$ this setting reduces to standard linear models of the type $f(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{w}$. We denote the hypothesis space as $\mathcal{H}$.

The regularized risk minimization problem (see e.g. [Evgeniou et al. 2000]) can be defined as

$$\operatorname*{argmin}_{f \in \mathcal{H}} R(f) + \lambda \|f\|^2, \tag{1}$$

where the first term measures how well f fits the training data, and the second term called the regularizer measures the complexity of the hypothesis, and $\lambda > 0$ is the regularization parameter.

2.2 Ranking problem

In the following, we assume that the label information in the training set is supplied in terms of pairwise preferences. These may be collected directly from pairwise comparisons. For example, [Joachims 2002], collected such preferences from clickthrough data from search engine, by considering clicked links to be preferred over those that were not chosen by a user. When supplied with scored data, preferences can be constructed by considering objects with higher scores being preferred over those with lower ones. In such cases not all objects may be comparable, for example the standard approach to modeling document retrieval problems [Joachims 2002; Liu 2009] results in a setting where data consists of query-document pairs, and preferences are constructed only between pairs such as are associated with same query. The concept of preference graph allows us to unify all these settings, though for computational reasons we might often in practice want to avoid its explicit construction.

A set of pairwise preferences can be encoded as a directed preference graph, where input points serve as vertices, and the edges encode preferences between the vertices. By an edge $e_i = (h, j)$, where $h \neq j$, we encode that x_h is preferred over x_j. We denote a preference graph drawn from the underlying distribution as

$$E = (e_1, \ldots, e_l).$$

In addition to pairwise preferences, we may in some settings have access to preference magnitudes, that denote to which degree an object is preferred over another. For scored data, preference magnitude can be defined as $y - y'$. If such information is not available and magnitudes are required, we may assume that each preference has a magnitude 1. In the following, we use E_M to denote a set of pairwise preferences augmented with preference magnitudes, meaning that each $e_i = (h, j, w_i) \in E_M$ contains a magnitude w_i encoding the degree, to which x_h is preferred over x_j.

Following [Herbrich et al. 1999], we measure the discrepancy between predicted and true rankings using the pairwise ranking error, defined as

$$l(f(X), E) = \sum_{(i,j) \in E} H(f(x_j) - f(x_i)).$$

where H is the Heaviside step function defined as

$$H(a) = \begin{cases} 1, & \text{if } a > 0 \\ 1/2, & \text{if } a = 0 \\ 0, & \text{if } a < 0 \end{cases}.$$

Intuitively, the loss can be considered as an estimate of the probability that the function is able to correctly predict, which of two randomly drawn examples is preferred over another. No polynomial time algorithm is known for minimizing this loss, which motivates the convex approximations introduced next.

2.3 RankRLS

The magnitude preserving pairwise ranking loss is defined as

$$l(f(X), E_M) = \sum_{(h,j,w_i) \in E} (w_i - f(x_h) + f(x_j))^2.$$

By inserting this loss in to (1), we recover the RankRLS method. The method extends the RLS regression method [Poggio and Smale 2003] by casting the problem of ranking into a pairwise regression

framework. In [Pahikkala et al. 2009] we proved that a global minimizer of the RankRLS risk functional can be found by solving a system of linear equations.

Let us denote by m the number of training inputs, by l the number of pairwise preferences in the training set, and by n the dimensionality of the feature space[1]. The complexity of training RankRLS using the algorithm described in [Pahikkala et al. 2009] is $O(m^3)$, which is based on solving a $m \times m$ linear system, using matrix factorization algorithms. This is a significant improvement compared to the straightforward approach of training RankRLS using a black-box RLS solver trained directly on the pairwise preferences, as this would result in highly impractical $O(l^3)$ worst case complexity (note that in many problems $l \approx m^2$). When using the linear kernel, RankRLS can be solved in $O(n^3 + \min(n^2m + m^2n + l, n^2l))$ time. If $n << m$ this can be quite efficient. Thus using basic dense linear algebra techniques based on matrix factorization, RankRLS can be trained in a time that is either cubic in the number of training examples, or cubic in the dimensionality of feature space.

Perhaps the main advantage of the RankRLS approach is the number of computational shortcuts made possible by the closed form solution. First, it can be shown that solutions for different regularization parameter values λ can subsequently be computed by re-using computations needed for RankRLS training in quadratic time. This is quite useful, since one rarely knows in advance the suitable value for λ, rather it is typically chosen by grid searching. Second, based on low-rank matrix update operations, one can develop computationally efficient cross-validation algorithms for RankRLS. These methods in effect allow a trained RankRLS model with a minimal number of operations to "unlearn" the effects of a hold-out set of examples. In [Pahikkala et al. 2009] we introduce exact methods for leave-pair-out cross-validation and leave-query-out cross-validation, and prove that these estimates can be computed with no additional asymptotic cost compared to training RankRLS once.

When using kernels, reduced set approximation can be used to scale RankRLS training beyond a few thousand training examples. This approach is considered in detail in [Pahikkala et al. 2009], and can be seen as a special case of the Nyström approximation scheme studied in [Airola et al. 2011a]. Finally, let us consider application domains where the data is sparse, meaning that the data matrix is filled mostly with zeroes. Using the linear kernel, it is possible to make use of this sparsity, avoiding explicitly constructing dense $m \times m$ or $n \times n$ matrices. Using the conjugate gradient method, the RankRLS optimization can rather be formalized in terms of sparse matrix - vector products. The basic technique is described in [Pahikkala et al. 2009], more detailed analysis and further experimental results are presented in [Airola et al. 2010]. Let $\overline{n}$ be the average number of non-zero features per example, and t the number of iterations that conjugate gradient optimization needs to converge. Then linear RankRLS can be trained with $O(tm\overline{n} + tl)$ cost.

2.4 RankSVM

The pairwise hinge loss is defined as

$$l(f(X), E) = \sum_{(i,j) \in E} \max(1 - f(x_i) + f(x_j), 0).$$

By inserting this loss in to (1), we recover the RankSVM method. The method extends SVMs [Vapnik 1995] by casting the problem of ranking into a pairwise classification framework. The approach was first considered in [Herbrich et al. 1999].

[1] Learning from scored data is more efficient than from pairwise preferences. In this setting, the terms containing l can be removed from all the following RankRLS complexities

	Running times					
Inputs	200	500	1000	2000	2500	4000
RankRLS	1	3	10	48	83	280
RankSVM	2	150	1740	13707	20055	-

Table 1: *Runtime comparisons of training kernel RankRLS and RankSVM in CPU seconds. The number of inputs ranges from 200 to 4000, the runtimes are measured in seconds.*

In theory, any standard SVM solver can be used to solve also the RankSVM problem by training on pairwise preferences directly. This approach was originally adapted in [Herbrich et al. 1999]. Further, the popular kernel RankSVM solver included in the $\mathrm{SVM}^{\mathrm{light}}$ software package uses a standard SVM solver trained on pairwise preferences for training the RankSVM [Joachims 2002]. The downside of this approach is that the computational complexity of these solvers becomes dependent not on the number of examples, but on the number of pairwise preferences, leading to $O(m^4)$ or worse scaling. For scored data and linear kernel, [Joachims 2006] introduce a method with $O(m\overline{n} + m\log(m) + rm)$, where r is the number of different utility levels in the data. If the number of allowed scores is not restricted, at worst case $r = m$ with the resulting complexity $O(m\overline{n} + m^2)$, meaning quadratic behavior.

[Airola et al. 2011b] presents a technique for removing this dependence on r from the complexity. The method uses self-balancing binary search trees to speed up loss and subgradient computations, allowing $O(m\overline{n} + m\log(m))$ worst case behavior for linear RankSVM training. On large enough data sets this can make a substantial difference in training times, reducing days of training time to minutes. Efficient kernel RankSVM training can be achieved using the empirical kernel map corresponding to the Nyström approximation, explored especially in [Airola et al. 2011a], in order to convert the dual RankSVM problem to the primal problem. This idea was introduced already in [Pahikkala et al. 2009]. The use of this approach for kernel RankSVM training has subsequently been independently considered by [Chapelle and Keerthi 2010]. Briefly put, using the efficient linear RankSVM training method this approach leads to $O(mk^2 + m\log m)$ RankSVM training complexity, where $k << m$ is a parameter that controls the amount of approximation.

2.5 Experimental results

In [Pahikkala et al. 2009], we report results for an experimental comparison of RankRLS and RankSVM, as well as RLS regression as baseline. Considered problems include collaborative filtering, document retrieval, n-best re-ranking of syntactic parses and text categorization. The main conclusion of the experiments is that there are rarely significant differences in ranking performance between the RankRLS and RankSVM approaches, while both tend to outperform the baseline method. However, as discussed before, there are major differences in computational efficiency.

In Table 1, we re-produce subset of the runtime comparison between RankRLS and RankSVM presented in [Pahikkala et al. 2009]. The RankRLS implementation is based on an early version of the RLScore software, whereas the RankSVM implementation is from $\mathrm{SVM}^{\mathrm{light}}$. As can be expected from the computational complexity considerations, RankRLS has much better scalability than the standard RankSVM implementation. When one further considers the costs of performing regularization parameter selection and cross-validation the difference becomes much larger, since these

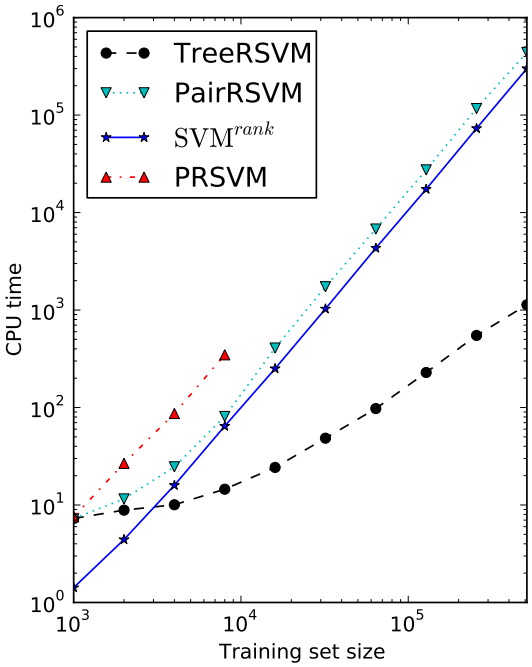


Figure 1: *Linear RankSVM runtime comparison.*

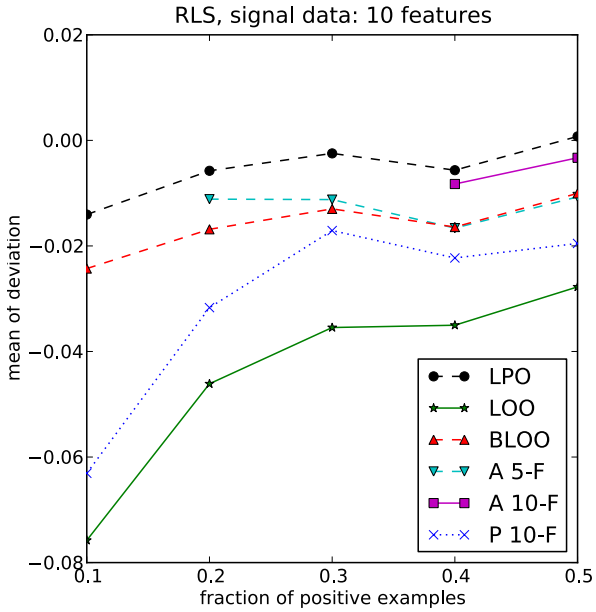


Figure 2: *Comparison of different cross-validation strategies.*

procedures can be performed essentially for free for RankRLS.

For the linear kernel, both RankRLS and RankSVM can be scaled to much larger problem sizes. Next, we consider how the RankSVM training algorithm that was introduced in [Airola et al. 2011b] for linear models and scored data, compares to the best previously known methods. The proposed method (TreeRSVM) is compared to the methods of [Joachims 2006] (SVM$^{\text{rank}}$ and PairRSVM) as well as to that of [Chapelle and Keerthi 2010] (PRSVM). With 512000 training examples, training SVM^{rank} took 83 hours, whereas training TreeRSVM took only 18 minutes in the same setting. Both methods reach the same solution. These results also suggest quite simple approach to improving nonlinear RankSVM training methods, since as discussed previously, for regularized kernel methods the kernelized learning problem can always be cast into a linear learning problem. Comparing results in [Airola et al. 2011b] and [Airola et al. 2010], it can be established that the fastest linear RankSVM and RankRLS training methods have quite similar scalability.

3 Cross-validation

Cross-validation is one of the most widely used methods in machine learning for model selection and performance evaluation. In cross-validation, one repeatedly splits the data set into two parts, a training set and a holdout set. The model is trained on the training set, after which it is used to make predictions on the holdout set. This procedure is repeated a number of times, after which a final estimate of the performance is computed over all the holdout sets on which predictions were made on. One major challenge in applying cross-validation is the computational cost, due to having to train a learner multiple times; our contributions towards solving this issue were briefly discussed in the previous section. In typical ranking problems further challenges are encountered with regards to reliability of cross-validation results, these are discussed next.

First, there exist two general strategies for aggregating cross-validation results together. [Bradley 1997] who considered the specific problem of AUC estimation referred to these alternatives as pooling and averaging. In pooling all the predictions are combined together and the performance measure is then computed over the combined predictions. In averaging, the performance is computed separately for each holdout set, and finally the average over these is computed. For classical univariate performance measures such as classification accuracy, or squared error, it makes little difference which strategy is used. However, for pairwise ranking measures, there is a clear difference. Briefly put, in pooling most of the compared pairs are formed from predictions made on different rounds of cross-validation. It can be shown that this can lead to substantial biases in the results. The averaging strategy avoids comparing predictions from different rounds. However, for typical approaches this leads to increased variability in the estimates, since most of the pairs are in this case ignored during cross-validation. These issues are further discussed in [Airola et al. 2011c].

Next, we consider one of the simulation studies reported in [Airola et al. 2011c], where a number of cross-validation strategies, are applied for AUC estimation. The problem is a bipartite ranking task (or equivalently, binary classification), with 30 training examples and 10 features. Instances from both classes are first drawn from normal distributions with unit variance and no covariance between the features. Nine of the features have mean zero for both classes, the tenth has mean 0.5 for the positive, and -0.5 for the negative class. We compute the cross-validation estimates over 10000 repetitions of the experiments, and compare these to the true expected AUC computed on a simulated test set of 10000 examples. The deviation measures the mean difference between estimated and true AUC. For an unbiased estimator this should be 0, positive values indicate optimistic, and negative values pessimistic bias.

The compared standard approaches are averaged 10-fold cross-

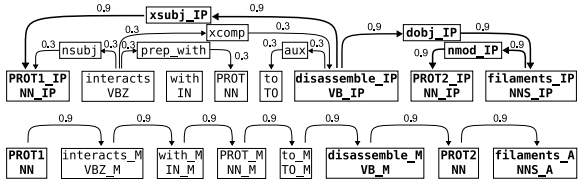


Figure 3: *Graph representation of a sentence.*

validation (A 10-F), averaged 5-fold cross-validation (A 5-F), pooled 10-fold cross-validation (P 10-F), and leave-one-out (LOO). Further, we consider the balanced leave-one-out variant proposed by [Parker et al. 2007] (BLOO). Finally, we test the leave-pair-out method (LPO) where pairs of data points are used as holdout sets, which we argue to be the most natural choice for pairwise performance measures such as AUC. The results, presented in Figure 2 demonstrate substantial biases in all the pooled methods. The averaged methods have less bias, with LPO providing the least biased estimates. Further experiments verify this trend over a large number of settings, while experiments that consider the variability of the approaches show LPO to be more reliable than the other averaging methods in this respect. The conclusion of the study is that for AUC estimation LPO should be preferred over other approaches, in cases where it can be computationally afforded.

Further studies in cross-validation can be found in [Airola et al. 2011a] where we consider how to correctly and efficiently deal with hold-out basis vectors when using the Nyström approximation to speed up training of kernel methods. In [Airola et al. 2008], we consider in an information extraction study the substantial biases that can arise due to the fact that training examples generated from the same sentence are much more similar to each other, than those generated from different sentences. The findings support the notion that for data where strong dependencies occur between the examples, dependent data points should never be split between the training and test sets, in order to ensure reliable performance estimation.

4 Biomedical information extraction

The task of protein-protein interaction (PPI) extraction from scientific literature is one of the major tasks considered in the field of biomedical natural language processing. Online resources, such as PubMed offer researchers access to millions of research articles in the biomedical domain, making manual search for stated results about interactions impractical. Rather, automated information extraction systems are needed. In the past, both rule-based approaches and methods based on machine learning have been proposed for solving the problem (see references in [Airola et al. 2008; Airola 2011]). In [Airola et al. 2008], we proposed and evaluated a novel approach for PPI extraction. The approach combines a novel graph kernel approach to learning from syntactic parses of sentences with an RLS based learning method, and efficient and reliable cross-validation strategies for model selection.

In Figure 3 we see an example of a sentence talking about PPI-interactions. In addition to the sentence itself, the figure presents a dependency parse for the graph, that was generated using an automated syntactic parser. Based on both the syntactic parse, as well as the linear ordering of the words in the sentence, the system has to decide which of the protein pairs appearing in a sentence are stated to interact, and which are not. Using kernel methods learning from this type of structured data becomes possible once we can define a suitable kernel function between the graph representations. In [Airola et al. 2008] we propose such an approach that is based on random walks in a graph, extending the earlier work of [Gärtner et al. 2003; Pahikkala et al. 2006].

The final method combines a wide range of approaches considered in other articles related to the thesis. Learning and parameter optimization is done by optimizing a RLS based objective function, using the Nyström approximation to scale the method. The fast leave-document-out cross-validation approach is very closely related to the computational shortcuts presented in [Pahikkala et al. 2009; Airola et al. 2011a], while reliable AUC-estimation requires accounting for the issues considered in [Airola et al. 2011c].

The experimental results presented in [Airola et al. 2008] demonstrated, that the method reached state-of-the-art performance compared to methods that had been previously proposed. Since then [Miwa et al. 2009] have proposed an improved solution to the same problem, that incorporates the proposed graph kernel as one of its main components. [Tikk et al. 2010] have further recently conducted a large-scale benchmark study of different kernel-based approaches to PPI-extraction, and reported the graph kernel to be among the most competitive approaches. As a further development, Turku Event Extraction System [Björne et al. 2011] provides a solution to the more challenging task of extracting complex structured interactions, incorporating many of the same syntactic features as used by the graph kernel.

5 Open source software

The importance of sharing open source implementations of published methods has recently been advocated in the machine learning community [Sonnenburg et al. 2007]. Accordingly, we are currently working on developing the RLScore machine learning open source software framework, which is made publicly available[2] under the MIT open source license. The package contains the RankRLS algorithms, as well as a wide variety of other learning methods. Further, the All-paths graph kernel -software package[3] [Airola et al. 2008], as well as the TreeRankSVM package [Airola et al. 2011b] are made publicly available under open source license.

6 Conclusion

In this article we have summarized the main contributions of the thesis [Airola 2011]. One of the major themes of research in this work was the development of computationally efficient algorithms for training and cross-validation, especially through the use of matrix algebra based techniques. In our related research, not included in the thesis, similar ideas have been applied for example in order to learn preference relations over relational graphs [Pahikkala et al. 2010], extend feature selection methods to genome wide scale [Pahikkala et al. 2012a], and for speeding up unsupervised and semi-supervised RLS training [Gieseke et al. 2012]. Most of the developed methods are or will be made freely available as part of the RLScore open source software.

Acknowledgements

I would like to thank the supervisors of my thesis, Adjunct Prof. Tapio Pahikkala and Prof. Tapio Salakoski, and the other co-authors. My thanks go also to the reviewers of the thesis, Prof. Eyke Hüllermeier and Associate Prof. Juho Rousu as well as to my opponent Adjunct Prof. Timo Honkela, for their constructive criticisms

[2] http://tucs.fi/rlscore

[3] http://mars.cs.utu.fi/PPICorpora/GraphKernel.html

and encouragement. The thesis work was done at the department of Information Technology at the University of Turku, as a member of the graduate programme of Turku Centre for Computer Science. I would further like to acknowledge the financial support provided by the Academy of Finland, Nokia Foundation, Sofia and Wilhelm Fagerholm Scholarship Fund and the Turku University Foundation for the thesis work. Finally, I would like to thank the Pattern Recognition Society of Finland for the dissertation award and the invitation to present this work.

References

AIROLA, A., PYYSALO, S., BJÖRNE, J., PAHIKKALA, T., GINTER, F., AND SALAKOSKI, T. 2008. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics 9 Suppl 11*.

AIROLA, A., PAHIKKALA, T., AND SALAKOSKI, T. 2010. Large scale training methods for linear RankRLS. In *Proceedings of the ECML/PKDD-Workshop on Preference Learning (PL-10)*, E. Hüllermeier and J. Fürnkranz, Eds.

AIROLA, A., PAHIKKALA, T., AND SALAKOSKI, T. 2011a. On learning and cross-validation with decomposed Nyström approximation of kernel matrix. *Neural Processing Letters 33*, 1, 17–30.

AIROLA, A., PAHIKKALA, T., AND SALAKOSKI, T. 2011b. Training linear ranking SVMs in linearithmic time using red-black trees. *Pattern Recognition Letters 32*, 9, 1328–1336.

AIROLA, A., PAHIKKALA, T., WAEGEMAN, W., DE BAETS, B., AND SALAKOSKI, T. 2011c. An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Computational Statistics & Data Analysis 55*, 4, 1828–1844.

AIROLA, A. 2011. *Kernel-Based Ranking: Methods for Learning and Performance Estimation*. Doctoral thesis, Turku Centre for Computer Science.

BJÖRNE, J., HEIMONEN, J., GINTER, F., AIROLA, A., PAHIKKALA, T., AND SALAKOSKI, T. 2011. Extracting contextualized complex biological events with rich graph-based feature sets. *Computational Intelligence 27*, 4, 541–557.

BRADLEY, A. P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition 30*, 7, 1145–1159.

CHAPELLE, O., AND KEERTHI, S. S. 2010. Efficient algorithms for ranking with SVMs. *Information Retrieval 13*, 3, 201–215.

EVGENIOU, T., PONTIL, M., AND POGGIO, T. 2000. Regularization networks and support vector machines. *Advances in Computational Mathematics 13*, 1–50.

FORMAN, G., AND SCHOLZ, M. 2010. Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *SIGKDD Explorations 12*, 1, 49–57.

FÜRNKRANZ, J., AND HÜLLERMEIER, E. 2010. Preference learning. In *Encyclopedia of Machine Learning*. 789–795.

GÄRTNER, T., FLACH, P. A., AND WROBEL, S. 2003. On graph kernels: Hardness results and efficient alternatives. In *Proceedings of the Sixteenth Annual Conference on Learning Theory and Seventh Annual Workshop on Kernel Machines (COLT/Kernel 2003)*, Springer, B. Schölkopf and M. K. Warmuth, Eds., vol. 2777 of *Lecture Notes in Artificial Intelligence*, 129–143.

GIESEKE, F., KRAMER, O., AIROLA, A., AND PAHIKKALA, T. 2012. Efficient recurrent local search strategies for semi- and unsupervised regularized least-squares classification. *Evolutionary Intelligence*, 1–17. Accepted for publication.

HERBRICH, R., GRAEPEL, T., AND OBERMAYER, K. 1999. Support vector learning for ordinal regression. In *Proceedings of the Ninth International Conference on Articial Neural Networks (ICANN 1999)*, Institute of Electrical Engineers, London, 97–102.

JOACHIMS, T. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2002)*, ACM Press, New York, NY, USA, D. Hand, D. Keim, and R. Ng, Eds., 133–142.

JOACHIMS, T. 2006. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2006)*, ACM Press, New York, NY, USA, T. Eliassi-Rad, L. H. Ungar, M. Craven, and D. Gunopulos, Eds., 217–226.

LIU, T.-Y. 2009. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval 3*, 3, 225–331.

MINKOV, E., CHARROW, B., LEDLIE, J., TELLER, S., AND JAAKKOLA, T. 2010. Collaborative future event recommendation. In *Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM 2010)*, ACM, New York, NY, USA, J. Huang, N. Koudas, G. J. F. Jones, X. Wu, K. Collins-Thompson, and A. An, Eds., 819–828.

MIWA, M., SÆTRE, R., MIYAO, Y., AND TSUJII, J. 2009. Protein-protein interaction extraction by leveraging multiple kernels and parsers. *International Journal of Medical Informatics 78*, e39–e46.

PAHIKKALA, T., TSIVTSIVADZE, E., BOBERG, J., AND SALAKOSKI, T. 2006. Graph kernels versus graph representations: a case study in parse ranking. In *Proceedings of the ECML/PKDD 2006 workshop on Mining and Learning with Graphs (MLG 2006)*, T. Gärtner, G. C. Garriga, and T. Meinl, Eds.

PAHIKKALA, T., TSIVTSIVADZE, E., AIROLA, A., JÄRVINEN, J., AND BOBERG, J. 2009. An efficient algorithm for learning to rank from preference graphs. *Machine Learning 75*, 1, 129–165.

PAHIKKALA, T., WAEGEMAN, W., AIROLA, A., SALAKOSKI, T., AND DE BAETS, B. 2010. Conditional ranking on relational data. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2010)*, Springer, J. L. Balcázar, F. Bonchi, A. Gionis, and M. Sebag, Eds., vol. 6322 of *Lecture Notes in Computer Science*, 499–514.

PAHIKKALA, T., OKSER, S., AIROLA, A., SALAKOSKI, T., AND AITTOKALLIO, T. 2012a. Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations. *Algorithms for Molecular Biology 7*, 1, 11.

PAHIKKALA, T., SUOMINEN, H., AND BOBERG, J. 2012b. Efficient cross-validation for kernelized least-squares regression with sparse basis expansions. *Machine Learning 87*, 3, 381–407.

PARKER, B. J., GUNTER, S., AND BEDO, J. 2007. Stratification bias in low signal microarray studies. *BMC Bioinformatics 8*, 326.

POGGIO, T., AND SMALE, S. 2003. The mathematics of learning: Dealing with data. *Notices of the American Mathematical Society (AMS) 50*, 5, 537–544.

SCHÖLKOPF, B., AND SMOLA, A. J. 2002. *Learning with kernels*. MIT Press, Cambridge, Massachusetts, USA.

SHAWE-TAYLOR, J., AND CRISTIANINI, N. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge.

SONNENBURG, S., BRAUN, M. L., ONG, C. S., BENGIO, S., BOTTOU, L., HOLMES, G., LECUN, Y., MÜLLER, K. R., PEREIRA, F., RASMUSSEN, C. E., RÄTSCH, G., SCHÖLKOPF, B., SMOLA, A., VINCENT, P., WESTON, J., AND WILLIAMSON, R. 2007. The need for open source software in machine learning. *Journal of Machine Learning Research 8*, 2443–2466.

TIKK, D., THOMAS, P., PALAGA, P., HAKENBERG, J., AND LESER, U. 2010. A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS Computational Biology 6*, 7, e1000837+.

VAPNIK, V. N. 1995. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Multi-View Factorizations and Modeling of Mutual Dependencies

Arto Klami*
Helsinki Institute for Information Technology HIIT
Department of Information and Computer Science
Aalto University

Abstract

Data analysis studies the task of applying computational methods for extracting understandable summaries of data collections, such as textual corpora or measurements of gene activities in cells. One of the central challenges in such analysis is in separating relevant or interesting information from noise. The task is readily solved in supervised learning, where the variation independent of the prediction target can be called noise, but remains open for unsupervised analysis. The thesis summarized in this article studied one approach for separating the relevant signal from noise also for unsupervised learning, by extending the learning setup to scenarios with multiple co-occurring data sets. In such multi-view learning setups we can define the relevant signals as the information shared by the multiple views, and noise as the information independent of the other views. We introduced several machine learning models for finding the dependencies between the views, and and in particular showed how Bayesian generative models can be used for finding the dependencies and view-specific noise signals also from small samples of high-dimensional data.

Keywords: Bayesian modeling, canonical correlation analysis, machine learning, multi-view learning, mutual dependency

1 Introduction

Data analysis is a sub-field of computer science, studying the problem of understanding data collections with computational methods, including statistical models, computational algorithms, and machine learning tools. The primary applications are in analysis of data collections that are too large for manual interpretation and for which the underlying mechanisms are typically unknown. As a practical example, it is nowadays fairly easy to measure the activity of all human genes with microarrays, resulting in a data matrix with tens of thousands of observations. The cellular processes causing the observed data are largely unknown, and hence data analysis tools are needed for digging out interesting information. Frequently the tools seek to summarize the data collection in a more understandable form, for example by clustering the genes into groups or by reducing the dimensionality of the data with linear or non-linear mappings.

One of the central challenges in data analysis is in defining what kind of properties are interesting, and in separating those from the noise inherent in all complex data collections. Application-specific information will often help here; for example, we might know the nature of the noise induced by a specific measurement device and it is often easy to build that knowledge into the model. For more complex systems, such as the functionality of a whole cell, the problem becomes much more challenging. No simple summary can contain all of the information, and the task becomes to pick interesting or relevant information. This extends the definition of noise: Any variation that is not interesting for the current focus of the analysis is noise, regardless of how strong structure it shows and whether we know the process or not.

*e-mail:arto.klami@aalto.fi

This article is a brief summary of the PhD thesis of the author [Klami 2008], which studied one approach for automatically separating such structured noise from the interesting signal. The fundamental hypothesis is that by recording multiple co-occurring representations of the same data, we can find the interesting information by studying the relationships between these representations. The variation that is reflected in all sources of data is likely to be more relevant, whereas the variation independent of all other sources can typically be assumed to be noise for the purpose of the analysis. In essence, this generalizes the same implicit assumption made in supervised analysis, where all the variation independent of the target variable to be predicted is considered noise.

The analysis of multiple co-occurring data sources, here represented as matrices $X_m \in \mathbb{R}^{N \times D_m}$ for $m \in [1, ...M]$, is called multi-view analysis; each of the M matrices provides a different view (of D_m dimensions or features) to the same set of N objects. Given this representation, we can formalize the idea presented above as the task of factorizing the variation in the whole collection into effects shared by the views and those specific to each view. Assuming the views have been chosen with care, the shared variation can be interpreted as the interesting signal. The setup is illustrated in Figure 1 (left).

In our work, the main focus has been in formalizing the above problem definition and in developing computational methods for solving the it. The main task has been to study methods applicable for cases with small sample size N and high dimensionalities D_i. Such data collections are challenging for data analysis due to lack of evidence for determining the true effects, yet common in many application fields, especially in life sciences where making several measurements would be expensive of even impossible (in case of rare diseases). For tackling the challenge of limited data we have focused on the probabilistic modeling framework, which enables quantifying the amount of uncertainty in the results through Bayesian inference. As a main contribution of the thesis, we introduced novel Bayesian models for factorizing the variation in multi-view setups to effects shared by the views and to those specific to each view.

2 Applications

To motivate the rest of the work, we start by introducing some prototypical examples of multi-view learning setups. The main application studied in the thesis was molecular cellular biology, where the task is to understand, for instance, how cancer influences the cells. Clearly the underlying mechanisms are very complex and largely unknown, and it is unreasonable to assume that any single measurement such as activity of the genes would correctly reveal all the relevant effects. However, by measuring multiple aspects of the cell we can create a multi-view collection with either the genes or patients as co-occurring samples. In [Klami and Kaski 2007] we studied the regulation of heat shock in yeast by searching dependencies between gene expression and transcription factor binding measurements, and in [Tripathi et al. 2008] we combined multiple gene expression time series to extract cell-cycle regulated genes. Later similar ideas have been used also for inferring post-translational regulation by factorizing the variation in gene expression and protein concentration [Rogers et al. 2010].

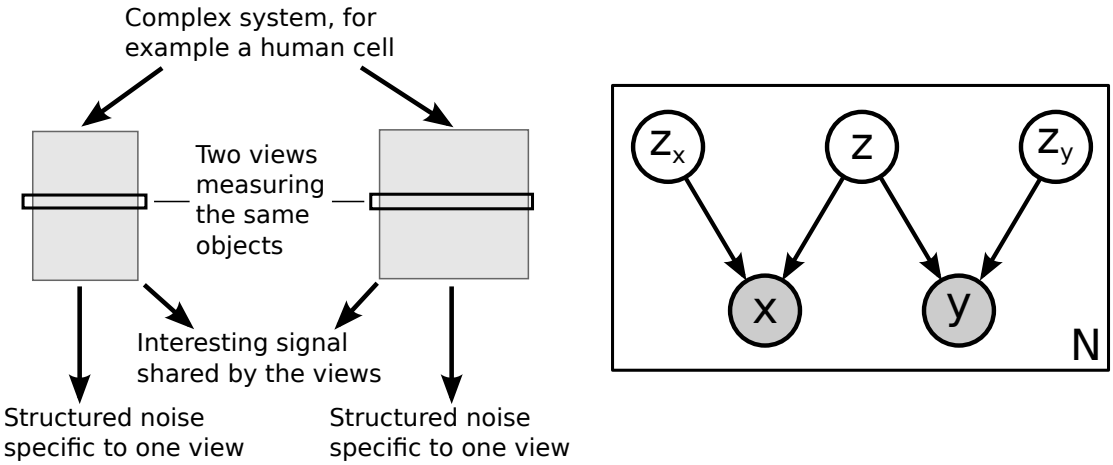


Figure 1: Left: *The basic concept of factorizing the variation in a multi-view data collection for extracting the relevant signal. The different vies (here two) are different measures or representations of a underlying complex system, given for the same set of objects. By modeling statistical dependencies between the views we can extract the relevant signal as the information shared by both views, whereas the residual variation in either view can be considered as structured noise not relevant for the application.* **Right:** *The idea presented on the left formulated as a Bayesian latent variable model. The shaded nodes* $\mathbf{x}$ *and* $\mathbf{y}$ *correspond to the two vectorial views, and the rest of the nodes are latent unobserved variables denoting the shared (*$\mathbf{z}$*) and view-specific variation (*$\mathbf{z}_x$ *and* $\mathbf{z}_y$*). The full Bayesian formulation is obtained by specifying prior distributions for the latent variables and determining the functional form of the mappings (which are also random variables) from the latent signals to the observations.*

Another interesting application field in life sciences is in computational neuroscience or brain signal analysis. In [Savia et al. 2009] we applied an extension of the DeCA model of [Klami and Kaski 2005] to understand brain activation caused by naturalistic stimuli, by learning dependencies between fMRI measurements and feature representations of the complex stimuli; for better overview of the setup, see [Ylipaavalniemi et al. 2009]. Others have also applied similar computational methods for understanding functional connectivity in brain [Deleus and Hulle 2011] by using different regions of brain as views, or reconstructing images from brain activity [Fujiwara et al. 2009] through multi-view analysis of the image content and fMRI.

Multi-view applications are also found in multi-lingual document analysis and analysis of multimedia content (e.g. text documents containing also images). In [Tripathi et al. 2010] we learned relationships between documents written in English and Finnish, and e.g. [Blei and Jordan 2003] have studied the problem of modeling paired text and images.

In general, multi-view learning models are useful for all application scenarios where collecting multiple different representations or measurements for the same objects is possible and it is reasonable to assume that the variation common to all (or many) of such views is likely to be more interesting. The basic idea applies to most physical measurements where the view-specific information is often sensor noise, as well as to scenarios where the task is to find a consensus amongst multiple subjects; brain response shared by several subjects viewing the same video is likely to be caused directly by the stimuli, whereas signal specific to an individual subject might well raise from some unrelated brain process.

3 Modeling dependencies between data sets

As described earlier, one of the core problems in multi-view learning is factorizing the variation in the data collection into signals shared by the views and into those specific to each view (the noise). A central element in solving the problem is in finding the shared signals, by modeling statistical dependencies between the views. Perhaps the easiest solution builds on the classical method of canonical correlation analysis (CCA) [Hotelling 1936]. It is a method for finding linear projections of the two data sources, $\mathbf{X}$ and $\mathbf{Y}$, so that the projections are maximally correlated. That is, the task is to find projection vectors $\mathbf{u}$ and $\mathbf{v}$ to maximize the quantity

$$\rho = \text{cor}\left(\mathbf{X}\mathbf{u}^T, \mathbf{Y}\mathbf{v}^T\right). \quad (1)$$

In practice we can find several such components by solving a generalized eigenvalue problem. By re-formulating the problem as a principal component analysis (PCA) of suitably transformed views, we were also able to generalize the formulation to more than two views [Tripathi et al. 2008].

The CCA model is limited in the sense that it only searches for correlations, whereas the true dependencies may be non-linear and more complex in nature. In [Klami and Kaski 2005] we introduced an extension called dependent component analysis (DeCA) that replaces the cost (1) with a non-parametric estimate of mutual information $\hat{I}(\mathbf{X}\mathbf{u}^T, \mathbf{Y}\mathbf{v}^T)$. It is no longer possible to find an analytic solution, but by using a Parzen kernel estimator we can still learn the projections that maximize the dependency with an iterative algorithm.

In effect, both CCA and DeCA directly provide a representation for the variation that is shared between the views. While this is not yet a full solution to the whole factorization problem, the residuals can be analyzed separately for further interpretation. These approaches are hence suitable for solving the factorization problem for the cases where linear component summaries are sufficient and there are enough training samples for reliably estimating the correlation or mutual information. However, for small sample sizes they severely overfit to the data.

4 Bayesian modeling of dependencies

Probabilistic modeling is one of the frequently applied frameworks for learning from data; the task is to build a generative description, formulated through probability distributions, for the data. One of the advantages of this framework is that it allows justified treatment of uncertainty. Most real-world measurements are noisy, typically on several levels, and probabilistic models allow proper treatment of such uncertain effects. As a result, the uncertainty in the observed data will be properly reflected in the results, revealing to which extent we can rely on the interpretations. This is particularly critical for small sample sizes.

In [Klami and Kaski 2006; Klami and Kaski 2008] we introduced the general guidelines for finding the multi-view factorizations with probabilistic models. The basic idea is that we build a model that has three separate parts (for two data sources): One that models the dependencies between the two, and one for modeling the variation specific to each source. For the case of linear-Gaussian models this corresponds to the model

$$\begin{aligned} \mathbf{x} &= \mathbf{W}_x\mathbf{z} + \mathbf{V}_x\mathbf{z}_x + \boldsymbol{\epsilon}_x \\ \mathbf{y} &= \mathbf{W}_y\mathbf{z} + \mathbf{V}_y\mathbf{z}_y + \boldsymbol{\epsilon}_y, \end{aligned} \tag{2}$$

where the parts $\mathbf{W}_x\mathbf{z}$ and $\mathbf{W}_y\mathbf{z}$ model the dependencies through a shared latent (non-observed) variable $\mathbf{z}$, and $\mathbf{V}_x\mathbf{z}_x$ and $\mathbf{V}_y\mathbf{z}_y$ model the variation specific to each source; $\boldsymbol{\epsilon}$ correspond to simple additive noise with no structure. This generative formulation is depicted in Figure 1 (right).

Building on the above observation, we introduced Bayesian variant of CCA [Klami and Kaski 2007]. From the perspective of a probabilistic model, CCA should model all of the correlations with the shared part, while making as little assumptions on the view-specific variation. For linear-Gaussian models this corresponds to assuming that the view-specific variation comes in form of arbitrary multivariate Gaussian. We obtain that by letting the dimensionality of $\mathbf{z}_x$ and $\mathbf{z}_y$ to match the ranks of the data matrices and by integrating them out. This results in the Bayesian CCA model

$$\begin{aligned} \mathbf{x} &\sim \mathrm{N}(\mathbf{W}_x\mathbf{z}, \mathbf{V}_x\mathbf{V}_x^T + \sigma_x^2\mathbf{I}) \\ \mathbf{y} &\sim \mathrm{N}(\mathbf{W}_y\mathbf{z}, \mathbf{V}_y\mathbf{V}_y^T + \sigma_y^2\mathbf{I}), \end{aligned}$$

where we can change the parameterization to $\boldsymbol{\Psi}_x = \mathbf{V}_x\mathbf{V}_x^T + \sigma_x^2\mathbf{I}$ and $\boldsymbol{\Psi}_y = \mathbf{V}_y\mathbf{V}_y^T + \sigma_y^2\mathbf{I}$ without loss of generality.

For full Bayesian treatment, the above model needs to be complemented with prior distributions and suitable inference algorithm for estimating the posterior distribution of the parameters given the data (which cannot be computed analytically). The interesting detail in the choice of priors is that by specifying an automatic relevance determination (ARD) prior

$$\begin{aligned} p(\alpha_k^{-1}) &\sim \mathrm{Gamma}(\alpha_0, \beta_0) \\ p(\mathbf{W}_x|\boldsymbol{\alpha}) &= \prod_{k=1}^{K} \mathrm{N}(\mathbf{W}_x(k)|0, \alpha_k^{-1}) \end{aligned}$$

for the projection matrices $\mathbf{W}_x$ and $\mathbf{W}_y$, we obtain a model that automatically learns the true number of correlating components. For inference, we have applied both deterministic variational approximation [Klami and Kaski 2008] and stochastic Markov chain Monte Carlo techniques [Klami and Kaski 2007].

One of the key advantages of the Bayesian framework is that simple changes in the generative process will result to novel models more suitable for specific data types. Already in [Klami and Kaski 2007] we introduced a Dirichlet process mixture of Bayesian CCA models and in [Klami and Kaski 2006; Klami and Kaski 2008] we considered clustering models following the same ideas.

5 Discussion

The above sections described the basic approach for factorizing the variation in multiple co-occurring data sources into the dependencies between the sources and to the structured noise independent of all the other sources, and summarized the main methodological contributions of the thesis [Klami and Kaski 2008]. Since then, we have worked on several extensions enabled by the groundwork done in thesis for specifying the general form of Bayesian multi-view factorizations. In [Rogers et al. 2010] we created non-parametric coupled clustering models, [Virtanen et al. 2011] gave improved optimization that makes Bayesian CCA a feasible model for very high-dimensional data, [Virtanen et al. 2012] extended the models to multiple views with more complex structure, and [Klami et al. 2010; Viinikanoja et al. 2010] introduced Bayesian CCA models with more flexible noise distributions. Other authors have also worked on extensions of the same ideas: [Archambeau and Bach 2009; Fujiwara et al. 2009; Rai and Daumé III 2009] have created sparse variants of Bayesian CCA, [van der Linde 2011] introduced Bayesian CCA for functional data, and [Huopaniemi et al. 2010; Nakano et al. 2011] have integrated Bayesian CCA as part of bigger hierarchical models.

Overall, the task of factorizing the variation in multi-view collections has hence became an active topic of study in the machine learning community. In particular, the latest works solving the remaining computational issues [Virtanen et al. 2011; Virtanen et al. 2012] have taken the models to the level that they can be reliably applied on real-world applications, including setups with tens or hundreds of high-dimensional views. As an example of a recent application success stemming from the methodological development, we managed to identify segments of speech based on single-trial MEG brain activity recordings using a mixture of Bayesian CCA models [Koskinen et al. 2012].

6 Acknowledgments

I would like to thank the supervisor of the thesis, Prof. Samuel Kaski, as well as the co-authors of the included articles, PhDs Janne Sinkkonen, Abhishek Tripathi, and Jaakko Peltonen. The thesis work was done in the AIRC Finnish Center of Excellence in the Computer and Information Science laboratory of Helsinki University of Technology, and funded primarily by the graduate school HeCSE.

References

ARCHAMBEAU, C., AND BACH, F. 2009. Sparse probabilistic projections. In *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. MIT Press, 73–80.

BLEI, D. M., AND JORDAN, M. I. 2003. Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, New York, NY, USA, 127–134.

DELEUS, F., AND HULLE, M. V. 2011. Functional connectivity analysis of fmri data based on regularized multiset canonical correlation analysis. *Journal of Neuroscience methods 197*, 1, 143–157.

FUJIWARA, Y., MIYAWAKI, Y., AND KAMITANI, Y. 2009. Estimating image bases for visual image reconstruction from hu-

man brain activity. In *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, Eds., 576–584.

HOTELLING, H. 1936. Relations between two sets of variates. *Biometrika 28*, 321–377.

HUOPANIEMI, I., SUVITAIVAL, T., NIKKILÄ, J., OREŠIČ, M., AND KASKI, S. 2010. Multivariate multi-way analysis of multi-source data. *Bioinformatics 26*, i391–i398. (ISMB 2010).

KLAMI, A., AND KASKI, S. 2005. Non-parametric dependent components. In *Proceedings of ICASSP 2005, IEEE International Conference on Acoustics, Speech, and Signal Processing*, IEEE, V–209–V–212.

KLAMI, A., AND KASKI, S. 2006. Generative models that discover dependencies between data sets. In *Proceedings of MLSP'06, IEEE International Workshop on Machine Learning for Signal Processing*, IEEE, 123–128.

KLAMI, A., AND KASKI, S. 2007. Local dependent components. In *Proceedings of ICML 2007, the 24th International Conference on Machine Learning*, Omnipress, Z. Ghahramani, Ed., 425–432.

KLAMI, A., AND KASKI, S. 2008. Probabilistic approach to detecting dependencies between data sets. *Neurocomputing 72*, 39–46.

KLAMI, A., VIRTANEN, S., AND KASKI, S. 2010. Bayesian exponential family projections for coupled data sources. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (2010)*, AUAI Press, Corvallis, Oregon, P. Grunwald and P. Spirtes, Eds., 286–293.

KLAMI, A. 2008. *Modeling of mutual dependencies*. PhD thesis, Helsinki University of Technology.

KOSKINEN, M., VIINIKANOJA, J., KURIMO, M., KLAMI, A., KASKI, S., AND HARI, R. 2012. Identifying fragments of natural speech from the listener's MEG signals. *Human Brain Mapping*.

NAKANO, T., KIMURA, A., KAMEOKA, H., MIYABE, S., SAGAYAMA, S., ONO, N., KASHINO, K., AND NISHIMOTO, T. 2011. Automatic video annotation via hierarchical topic trajectory model considering cross-model correlations. In *Proceedings of ICASSP*. 2011.

RAI, P., AND DAUMÉ III, H. 2009. Multi-label prediction via sparse infinite CCA. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, Eds., 1518–1526.

ROGERS, S., KLAMI, A., SINKKONEN, J., GIROLAMI, M., AND KASKI, S. 2010. Infinite factorization of multiple non-parametric views. *Machine Learning 79*, 1-2, 201–226.

SAVIA, E., KLAMI, A., AND KASKI, S. 2009. Fast dependent components for fMRI analysis. In *Proceedings of ICASSP 09, the International Conference on Acoustics, Speech, and Signal Processing*, IEEE, 1737–1740.

TRIPATHI, A., KLAMI, A., AND KASKI, S. 2008. Simple integrative preprocessing preserves what is shared in data sources. *BMC Bioinformatics 9*, 111.

TRIPATHI, A., KLAMI, A., AND VIRPIOJA, S. 2010. Bilingual sentence matching using kernel CCA. In *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, IEEE, S. Kaski, D. J. Miller, E. Oja, and A. Honkela, Eds., 130–135.

VAN DER LINDE, A. 2011. Reduced rank regression models with latent variables in Bayesian functional data analysis. *Bayesian Analysis 6*, 1, 77–126.

VIINIKANOJA, J., KLAMI, A., AND KASKI, S. 2010. Variational Bayesian mixture of robust CCA models. In *Machine Learning and Knowledge Discovery in Databases. Proceedings of European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010*, Springer, Berlin, A. G. José Luis Balcázar, Francesco Bonchi and M. Sebag, Eds., vol. III, 370–385.

VIRTANEN, S., KLAMI, A., AND KASKI, S. 2011. Bayesian cca via group sparsity. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ACM, New York, NY, USA, L. Getoor and T. Scheffer, Eds., ICML '11, 457–464.

VIRTANEN, S., KLAMI, A., KHAN, S. A., AND KASKI, S. 2012. Bayesian group factor analysis. In *Proceedings of AISTATS'12*.

YLIPAAVALNIEMI, J., SAVIA, E., MALINEN, S., HARI, R., VIGÁRIO, R., AND KASKI, S. 2009. Dependencies between stimuli and spatially independent fMRI sources: Towards brain correlates of natural stimuli. *NeuroImage 48*, 176–185.

Exact bounds for distributed graph colouring

Joel Rybicki*
Department of Computer Science, University of Helsinki, Finland

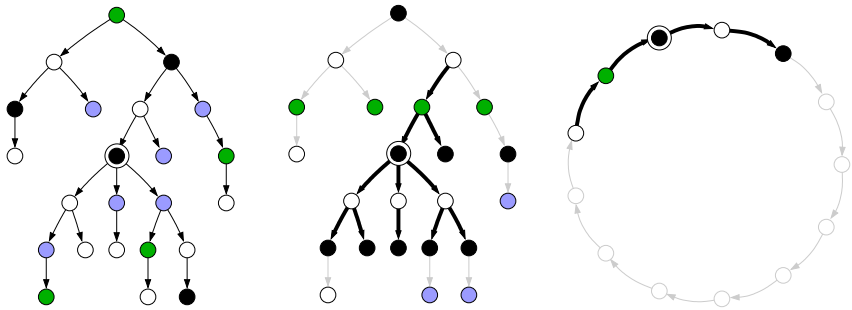


Figure 1: *Distributed colouring of directed rooted trees can be reduced to colouring directed cycles.*

Abstract

A distributed system is a collection of networked processing units which have to work in a cooperative manner. Such systems are often modelled as graphs where nodes represent the processors and edges denote communication links between the processors. One fundamental problem in distributed computing is the graph colouring problem. In this problem, the task is to assign colours (numerical labels) for the nodes in the network such that no adjacent nodes share the same colour. This paper summarizes new refined upper and lower bounds for distributed graph colouring in directed cycles and directed rooted trees given in the author's Master's Thesis [2011]. The new bounds were obtained using computational methods, as the existence of distributed colouring algorithms corresponds to the colourability of so-called neighborhood graphs.

CR Categories: C.2.4 [Computer-communication networks]: Distributed systems; F.2.2 [Analysis of algorithms and problem complexity]: Nonnumerical algorithms and problems—Computations on discrete structures.

Keywords: distributed algorithms, graph colouring, colour reduction

1 Introduction

A distributed system consists of networked processors that solve a computational task cooperatively. The processors act simultaneously and may communicate with each other using communication links. The fundamental algorithmic problems in distributed computing often relate to network management; examples of such tasks include scheduling the activity of processors or finding routing schemes in the network. Thus, the network itself is often considered to be a part of the input for a distributed algorithm.

The structure of the communication network is modelled as a graph $\mathcal{G} = (V, E)$ where each node $v \in V$ represents a processor and an edge $\{u, v\} \in E$ denotes a communication link between two processors. Each processor may be given additional input which is not initially available to other processors. This local input may represent a unique identifier, resource constraints or other information related to the problem at hand. The processors may share their local information using the communication links. When the computation finishes, each processor outputs its own part of the solution, such as a new colour for the node or an activity schedule for the node.

A central theme in distributed computing is locality and efficient coordination of processors even in very large networks. A key question is how well can the distributed system be managed even if the processors know very little about the structure of the network. In a centralized setting, a single processor sees all of the input and can then sequentally compute the output. However, in a distributed system, a single processor sees only a small portion of the input. Therefore, the goal is to find algorithms that require no global information about the network: it should be sufficient to share information only with relatively near-by processors.

As an example, consider a scheduling problem in wireless sensor networks where the task is to allocate time slots for radio transmissions in such a way that near-by transmission do not interfere with each other. The processors must agree on time periods when each processor may be active without disrupting any other transmission. Of course, if global information about the network is available, allocating non-conflicting time slots is easy. On the other hand, it is non-trivial to develop algorithms for which strictly local information suffices.

1.1 Graph colouring

This work concentrates on the graph colouring problem. The objective of the graph colouring problem is to assign a colour to each node in such a way that no two nodes connected by an edge share the same colour. Formally, given a graph $\mathcal{G} = (V, E)$ and a positive integer k, the objective is to find a mapping $\varphi\colon V \to \{0, 1, \ldots, k-1\}$ such that for all nodes $u, v \in V$ it holds that $\{u, v\} \in E \implies \varphi(u) \neq \varphi(v)$. A graph that has been coloured using at most k different colours is said to be k-coloured. Figures 1 and 2 illustrate examples of coloured graphs.

Essentially, a k-colouring divides the nodes into k independent sets which can then complete tasks in a conflict-free manner, as each colour class can act in sequence: nodes with colour 0 act first, after which nodes with colour 1 act, and so on. Thus, in order to minimize running times of algorithms that rely on a graph colouring, it is preferable to use as few colours as possible.

*e-mail: joel.rybicki@cs.helsinki.fi

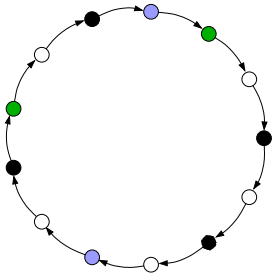


Figure 2: *A 4-coloured directed cycle graph.*

1.2 Model of distributed computation

This work uses the synchronous message-passing model of distributed computation: every node runs the same algorithm, and the system operates in discrete synchronous communication rounds. During each round, a node can communicate with its neighbours and perform local computation. The time complexity of an algorithm is the number of communication rounds required to run the algorithm.

It is assumed that the graph is already properly coloured with at most k colours. If each node is given a unique identifier, then these identifiers form a colouring. The task of the algorithm is to significantly reduce the number of used colours. For cycles and trees, the goal is to achieve a 3-colouring. Finding an optimal colouring is rarely possible with a fast distributed algorithm, since an optimal colouring may depend on global properties of the graph.

In the message-passing model, it is feasible to state the *exact* running time of an algorithm without resorting to asymptotics. Similarly, it is possible to find matching lower bounds on the number of communication rounds required to distributively solve a problem. In the extreme case, it can be shown that $f(k)$ communication rounds suffices to solve a problem as well as prove that it is impossible to solve in $f(k) - 1$ rounds.

However, only a few examples of such tight results are known so far; for many problems, the gaps between positive and negative results remain quite large. In this work, the focus is on the exact complexity of distributed graph colouring in directed cycles and directed trees. In particular, the goal is to narrow the gap between the previously known lower bound results and upper bound results.

2 Previous work

A directed cycle is a 2-regular graph with a globally consistent orientation of the edges: each edge is assigned a label denoting the direction of the edge. A directed rooted tree is a tree with edges oriented from the root towards the children as illustrated in Figure 1. Algorithms designed for more general graphs often apply colouring algorithms for these simpler graphs as subroutines.

Previous research has established that complexity of 3-colouring a k-coloured directed cycle is asymptotically $\Theta(\log^* k)$ communication rounds where $\log^* k = \min\{i\colon \log^{(i)} k \leq 1\}$ and $\log^{(i)}$ stands for i repeated applications of the base-2 logarithm. The $\log^*$ function is an extremely slow growing function. For all practical values of k, the value $\log^* k$ is at most 7.

The original positive results were provided by Cole and Vishkin [1986] who designed an algorithm that can be adapted for 3-colouring k-coloured cycles in the message-passing model. A careful analysis of the algorithm shows that a directed cycle can always be 3-coloured in $\frac{1}{2}(\log^* k + 7)$ communication rounds.

A corresponding lower bound was later established by Linial [1992] who showed that the Cole–Vishkin algorithm is asymptotically optimal: a directed cycle cannot be 3-coloured in less than $\frac{1}{2}(\log^* k - 3)$ communication rounds. Thus, reducing the k-colouring to a 3-colouring in directed cycles is always possible in exactly $\frac{1}{2}(\log^* k + c)$ rounds for some constant c.

3 Results and methods

In addition to a review of positive and negative results of distributed graph colouring, the thesis [2011] shows refined positive and negative results for the complexity of distributed graph colouring in directed cycles and trees. These results are derived using computational methods, as the existence of distributed colouring algorithms corresponds to the colourability of so-called neighbourhood graphs.

Originally, the colourability of these graphs was analytically studied by Linial. In the thesis, this approach is extended by analysing the chromatic number of various neighbourhood graphs. The computational analysis was done using Boolean satisfiability (SAT) solvers.

The analysis yields new bounds for the constant c. For any $k \geq 3$, colouring a k-coloured directed cycle with at most three colours is possible in $\frac{1}{2}(\log^* k + 3)$ rounds. Furthermore, it is also shown that for some values of k, colouring a directed cycle with at most three colours requires at least $\frac{1}{2}(\log^* k + 1)$ communication rounds.

Moreover, it is observed that distributed colouring of trees is twice as hard as colouring cycles: a k-coloured directed rooted tree can be 3-coloured in $2T(k)$ rounds if and only if a k-coloured directed cycle can be 3-coloured in $T(k)$ rounds. Therefore, in the case of directed rooted trees, reducing a k-colouring into a 3-colouring requires at least $\log^* k + 1$ rounds for some k and is possible in $\log^* k + 3$ rounds for all $k \geq 3$.

Since the $\log^*$ function is a very slowly-growing function, the constant additive term significantly contributes to the total running time. For example, suppose we have a cycle network where each processor is given a unique identifier from the set $\{0, 1, \ldots, 2^{128} - 1\}$. In this case, the identifiers can be considered as 128-bit IPv6 addresses which also form a colouring, as no node is given the same address. With the previous algorithms, 3-colouring would take 5 communication rounds, but this can be done in only three rounds.

Finally, it is shown that these types of computational methods are not only limited to graph colouring problems. The thesis explores extensions to neighbourhood graph constructions which can be used to capture the existence of distributed algorithms for other problems. As an example, it is shown how the existence of distributed algorithms for maximal matching can be formulated as a Boolean satisfiability problem.

Acknowledgements. I am grateful to my thesis advisors Jukka Suomela and Petteri Kaski invaluable discussions and comments. Moreover, I wish to thank Veli Mäkinen for additional feedback and supervising the thesis.

References

COLE, R., AND VISHKIN, U. 1986. Deterministic coin tossing with applications to optimal parallel list ranking. *Information and Control 70*, 1, 32–53.

LINIAL, N. 1992. Locality in distributed graph algorithms. *SIAM J. Comput. 21*, 1, 193–201.

RYBICKI, J. 2011. *Exact bounds for distributed graph colouring*. Master's thesis, University of Helsinki.

Reinforcement Learning In Real-Time Strategy Games

António Gusmão and Tapani Raiko
Aalto School of Science

Abstract

We consider the problem of effective and automated decision-making in modern real-time strategy (RTS) games through the use of reinforcement learning techniques. RTS games constitute environments with large, high-dimensional and continuous state and action spaces with temporally-extended actions. To operate under such environments we propose Exlos, a stable, model-based Monte-Carlo method. Contrary to existing model-based algorithms, Exlos assumes models are imperfect, reducing their influence in the decision-making process. Its effectiveness is further improved by including a novel online search procedure in the control policy. Experimental results in a testing environment show the superiority of Exlos in discrete state spaces when compared to traditional reinforcement learning methods such as Q-learning and Sarsa. Furthermore, Exlos is shown to be effective and efficient on an environment with a large continuous state and action space. This work is a summary of [Gusmao 2011].

Keywords: reinforcement learning, real-time strategy, games, artificial intelligence, UCT, planning, continuous reinforcement learning

1 Introduction

In the last decades, the video and computer game industry has been growing at an amazing rate, already being a multi-billion dollar industry, one of the biggest in the entertainment sector. In 2008, the computer and video game sales grew to 11.7 billion dollars in the United States alone, a growing trend consolidated over a 12 year-period [Entertainment Software Association 2009]. This work will focus on one game genre, real-time strategy (RTS) games, but the algorithms discussed herein have a much broader applicability, including other game genres and various real-world problems. RTS games are most usually related to warlike simulations, involving both economic development and military tactics. Opposing teams must collect resources in order to build armies and defeat their opponents. Players are faced with many tactical decisions which must be taken quickly and, in most cases, under uncertainty since enemy location and units are unknown to the player. Recent RTS games have rich worlds that can involve thousands of units deployed in massive game scenarios. Game designers stand much to gain from incorporating state-of-the-art AI techniques into their games. Perhaps more importantly, academic AI researchers have in video-games an important step between purely theoretical work and solving complex real-world problems. This is not a novel concept (see e.g. [Buro 2004]) but has yet to catch the necessary attention from the AI academic community.

RTS games are essentially complex simulators with the power to model a large diversity of real-world problems. They force researchers to come up with new, robust and efficient algorithms that can be tested in a virtual environment without requiring expensive hardware (e.g. robots) or cryptic, unintuitive simulators. With the increase in computation power in the last decades, machine-learning methodologies become valid candidates to tackle the complexity of RTS games. Several authors have proposed learning systems for RTS games [Hsieh and Sun 2008; Weber et al. 2010; Molineaux et al. 2009] but most are either not effective or operate under very limited conditions. From available techniques, reinforcement learning (RL) stands out as a method that seamlessly combines planning and learning in a framework with a wide scope of applicability. In addition, reinforcement learning is an established field with significant theoretical grounding and is currently a popular and active research topic, stimulated by recent discoveries of convergent algorithms for RL with function approximation [Maei et al. 2009; Maei et al. 2010].

Currently, there are few success cases of RL in high-dimensional and continuous state or action spaces, such as the ones encountered in RTS games. This article introduces a novel method which combines ideas from several existing algorithms, resulting in a robust learning system for stochastic environments with temporally-extended actions and continuous state and action spaces.

2 Background

2.1 Reinforcement Learning

In reinforcement learning (RL), a positive reward is awarded to a decision maker if something good happens and negative reinforcement is given if something bad happens. The decision maker attempts to adjust his decisions in order to receive increased positive rewards. A RL agent must know how to act at each state it might visit by discovering which of its current available actions lead to the largest long-term rewards, essentially learning a mapping between states and actions. The optimal actions are discovered from interaction with the environment, may that interactions be real or simulated. For an extensive introduction to reinforcement learning the reader is referred to [Sutton and Barto 1998].

2.2 UCT Algorithm

UCT [Kocsis and Szepesvári] stands for upper confidence trees and is a Monte-Carlo tree search (MCTS) method that sets up a multi-armed bandit problem for each state.

Consider a reinforcement learning agent. Denote $Q(s,a)$ as the value of action a in state s and $\mathcal{A}_s$ as the set of available action at state s. Let $\beta(s,a)$ be a real-valued bias factor. At each state s UCT takes the action with largest:

$$Q_{UCT}(s,a) = Q(s,a) + \beta(s,a) \tag{1}$$

The bias factor is:

$$\beta(s,a) = \sqrt{\frac{2b^2 \ln(\sum_{a'} N(s,a'))}{N(s,a)}}$$

where $N(s,a)$ is the number of times action a was taken in state s.

Essentially, UCT adds a bonus to each action value, the bonus being determined by the bias factor $\beta(s,a)$. Taking action a at state s results in a decrease of the bonus given to a and an increase to the one given to all other actions that could have been taken, i.e. $\beta(s,a)$ is decreased and $\beta(s,a')$ is increased for all $a' \neq a, a' \in \mathcal{A}_s$. UCT will never stop exploring. It follows directly from Equation (1) that, in an infinite number of action selections at state s, any action will have its value increased to infinity if it is not picked infinitely often.

3 Exlos

3.1 Speeding Up Learning: Simulated Experience

In our problem setting we assume action models, $P(s'|s,a)$, are available but inaccurate. Simulating experience is a way of compensating for lack of real experience, but it can become a nuisance if real experience is not lacking. A simple way to solve this issue is to combine simulated experience and real experience in a way that the former becomes less important as the latter is accumulated. The resulting algorithm possesses the advantages of model-based learning and no significant drawback. A general purpose manner of doing this is to define the value of a state as a weighted average of a simulation-based value and a value based on real experience, enforcing the weight of the simulation-value to drop to zero as the number of episodes experienced goes to infinity. Let $V_{sim}(s)$ represent the value of state s obtained from simulated experience, henceforth known as simulated value. Let $V_{real}(s)$ denote the value obtained from real experience, henceforth known as real value. Then value $V(s)$ for state s is:

$$V(s) = \frac{w_{real}V_{real}(s) + w_{sim}V_{sim}(s)}{w_{real} + w_{sim}} \quad (2)$$

where w_{real}, w_{sim} are the weights of real value and simulated value, respectively. If action models or the learning process of V_{sim} are not reliable, weights should be updated so that $V(s)$ approximates V_{real} as the number of experienced episodes goes to infinity.

For a state s_t, we compute the value $V_{sim}(s_t)$ as the average return obtained by starting at state s_t, and simulating partial episodes, $\{s_t, s_{t+1}, ..., s_{t+d}\}$, until either a terminal state is found or a threshold episode length is reached. At the final simulated state, s_{t+d}, the value of $V_{real}(s_{t+d})$ is taken as the reward for that episode and is propagated through the simulated episode back to s_t, the state where the simulation started. The simulation value function is dependent on the real value function. Hence, it makes sense to update the former only when the latter changes. Exlos updates simulated values after experiencing an episode and updating the real value function.

In essence, the simulation value function is a way to propagate real experience to larger regions of state-space, effectively extracting more information from each experienced episode. Smooth function approximators perform a similar generalization but loosely based on the assumption that state proximity corresponds to a proximity in values. However, this is not necessarily a good assumption. The simulation value function, however, generalizes based on the structure of the reinforcement learning problem, considering how actions affect the environment.

3.2 Incorporating Heuristic Estimates

In large state-spaces that result in lengthy episodes, a prior state-value function is essential. When no experience has been collected, algorithms can do no better than to select actions at random. Agents acting randomly can take a prohibitively long time to reach terminal states. In addition, it might take many episodes for a reasonable region of state space to be evaluated. Another problem comes from function approximation, which when lacking experience, generates erroneous, non-zero values for unvisited regions of state-space. A prior state-value function guides the agent in the initial stages of learning and effectively reduces the influence of poor generalization due to lack of experience. Consequently, Exlos supports the inclusion of a prior state-value function which is combined with both real and simulated value of a state according to:

$$V(s) = \frac{w_{real}V_{real}(s) + w_{sim}V_{sim}(s) + w_{prior}V_{prior}(s)}{w_{real} + w_{sim} + w_{prior}} \quad (3)$$

The weight w_{prior} may be a function of state but for simplicity, we consider it a constant. The contribution of V_{prior} to $V(s)$ should approach zero as experience is collected. This can be achieved by having w_{real} represent the number of visits to state s. In that situation a constant w_{prior} represents the experience contained in the the prior value function.

3.3 Online Search

In computer chess, the minimax planning algorithm assumes that the value function is not optimal, and therefore creates a search process at each decision state, attempting to improve its action-value estimates. In reinforcement learning the focus is on learning the value function and, in most cases, no online search is performed. This is acceptable if one assumes the value function converges to the target value function, which is the optimal value function for off-policy learning, and may or may not be for on-policy learning. However, value functions represented by function approximators converge to an approximation of the target value function. Often, the estimated value function exhibits artifacts related to the functional family of the approximator. For example, multi-layer perceptrons with logistic neurons typically exhibit long and narrow ridges, whereas radial-basis function (RBF) networks generate bumpy surfaces. An online search process can look past those artifacts and correct local inaccuracies, greatly alleviating the need for a locally accurate function approximator. In turn this allows the use of smoother function approximators that are likely to generalize better. Hence, online search should be an essential part of any RL algorithm that relies on function approximation.
Searching is a non-trivial process in MDPs and increasingly so in continuous environments. The approach we propose considers each search as a reinforcement learning problem in a local environment extracted from the original environment. At decision state s, Exlos starts a reinforcement learning problem with starting state s and with modified terminal conditions. For the map environment defined in the beginning of this chapter Exlos creates a smaller squared map, centered in respect to the current position of the RL agent. The set of terminal positions for this new environment includes all terminal positions of the original environment that lie inside the squared region, plus all positions that lie outside the squared region. In other words, a simulated episode is terminated if the agent either reaches a terminal position or reaches a position outside the squared region. For positions outside the squared region, the experience value V_{real} is taken as the reward for that position.

Computational efficiency is improved by limiting the situations where the online search procedure is executed. Instead of determining an ephemeral local value function at each decision step, the value function is stored and queried in future decisions. For simplicity, the validity of the local value function is considered to be the state-space where it was computed. When the agent leaves the local environment, the local value function is discarded and will not be available even if the agent returns to states belonging to the local environment. Whereas traditional planning methods output sequences of actions, Exlos's online search outputs a policy in the form of a local value function. Therefore, Exlos is able to adapt to stochastic events whereas traditional planning is bound to follow the sequence of actions in the plan, irrespective of the actual state transitions that occur. In addition to reusing the value function, Exlos only performs online search where it is most beneficial - when the algorithm reaches a local maxima of the value function.

Since hill-climbing based on one-step lookups of the global value function is ineffective at escaping local maxima, Exlos invokes the search procedure, broadening the region of state-space effectively considered when deciding the next action to take.

3.4 Exploration

It is common for reinforcement learning agents to use ϵ-greedy or a softmax distribution to enforce exploration of the state-action space. We argue that such policies lack an adequate system that promotes exploration of rarely-visited states. Policies which rely on random behavior to explore the state-space are not appropriate for problems where the environment has a tendency to transition to low-reward states. This is the case in the map environment and in competitive games in general, where opponents constantly exploit sub-optimal decisions made by the reinforcement learning agent. We have analysed an algorithm which encourages exploration through a different mechanism - UCT. UCT explores state-action space by acting greedily in respect to an action-value function that increases the value of actions which are not taken regularly. This is a more effective method of exploration because it forces the algorithm to direct itself to states and actions which are rarely taken. We propose a mechanism inspired by UCT which encourages exploration by increasing or decreasing the value of states, resulting in a corresponding increase or decrease in action values.

Exlos reduces the values of visited states and increases the value of non-visited states. The bias, $\beta_{offline}(s)$, is determined by:

$$\beta_{offline}(s) = c\sqrt{\frac{\ln(N)}{T(s)}} \tag{4}$$

where N is the number of episodes experienced and $T(s)$ is the number of episodes in which state s was visited. The value of a state becomes:

$$V_{explor}(s) = V(s) + \beta_{offline}(s) \tag{5}$$

The final ingredient that completes Exlos is an additional bias value that increases exploration within an episode (which is already guaranteed by the stochastic policy). Denote the bias as β_{online}. The value for a state s becomes:

$$V_{explor}(s) = V(s) + \beta_{offline}(s) + \beta_{online}(s) \tag{6}$$

β_{online} is an ephemeral function which is reset at the end of each episode. It is the sum of two components, one that incites exploration by devaluing visited states during the episode and one that devalues local maxima found.

3.5 Overview of Exlos

Figure 1 summarizes the structure of Exlos.

Let us go through the decision and learning process of Exlos. A weighted average (3) combines a prior value function, V_{prior}, an experience-based value function, V_{real} and a simulation-based value function, V_{sim}:

$$V_1(s) = \frac{w_{real}V_{real}(s) + w_{sim}V_{sim}(s) + w_{prior}V_{prior}(s)}{w_{real} + w_{sim} + w_{prior}}$$

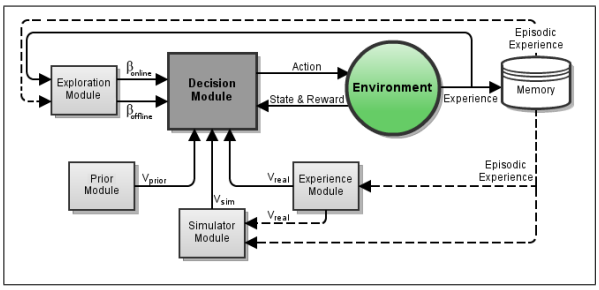


Figure 1: *Exlos structure. Dashed lines represent offline connections that are not active during execution of an episode.*

The simulation value training algorithm updates V_{sim} from values of V_{real} and both V_{real} and V_{sim} are estimated using function approximators. Visit counters are used to determined w_{real}. The exploration module encourages exploration across episodes through an added bias signal, $\beta_{offline}$ and exploration within an episode through β_{online}:

$$V(s) = V_1(s) + \beta_{offline}(s) + \beta_{online}(s)$$

Given a state-value function and knowledge of action models, action-values are computed:

$$Q(s,a) = E[V(s') \mid s,a] = \sum_{s'} P(s'|s,a)V(s')$$

With action-values calculated, the decision module picks actions following a softmax policy. In addition, an online search procedure is invoked whenever a (local) maximum of the state-value function is found. The search solves a reinforcement learning problem in a simulated local environment with $V_{real}(s)$ as the source of rewards as well as a prior for the local value function. The value function outputted by the search is used as replacement for $V_{real}(s)$.

4 Testing Environment

We implemented a set of two-dimensional rectangular maps, each map being an independent environment. The environments are episodic with fixed terminal states and each state is a position $\{x, y\}$ of the map. There are two versions of the map environments. A discrete version (a tile-based map) where x and y must be integers, and a continuous version where x and y are real-valued. Certain states cannot be visited and no action will generate them. They are represented as black squares in the map. For all purposes, they are not part of the state-space. Denote these states *walls*. Rewards are zero in all states except terminal states.There are two types of terminal states. Winning states and losing states. In a winning state the reward given is $+1$. Conversely, in a losing state the reward is -1. At each state there are four actions: move up, move down, move left and move right. The moves are not deterministic. With a certain probability a move in any direction will result in moving one tile in the direction of the closest losing state. This mimics an adversarial environment where the opponent leads the reinforcement learning agent to a losing state. In the case of continuous state maps, the set of actions is augmented with temporally-extended actions that perform movements across greater distances than the base four actions. These actions are sampled during the execution of each episode and consist of movements to locations sampled randomly from a circular area centered at the state at which the actions are available. The reinforcement learning agent solves a discounted reward problem.

5 Results

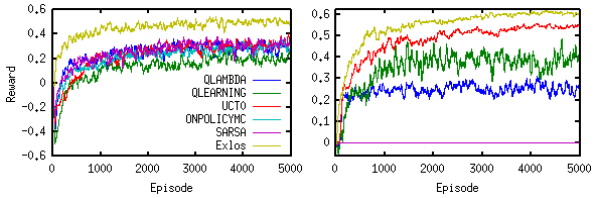


Figure 2: *Rewards obtained by popular reinforcement learning algorithms on two different, discrete map environments of small dimensions.*

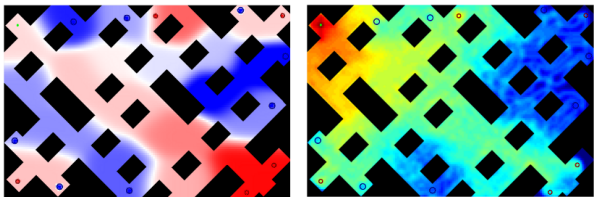


Figure 3: *A map with 1200x800 tiles. Black positions are walls, states that the agent cannot visit. The start location is in the top left. Small circles are winning locations, larger circles are losing locations. On the left the state-value function is shown, and on the right an heat map of the states visited by the agent during the learning process. Learning was performed for 200 episodes.*

6 Conclusions

We have motivated the research of real-time strategy game AI as a simulation environment for the development of novel and efficient learning algorithms that operate in complex, large-scale, stochastic environments with continuous state-action spaces and temporally-extended actions. In addition, we introduced Exlos, an effective on-policy Monte-Carlo algorithm. Exlos is a model-based algorithm that learns a state-value function. It encourages exploration of state-space by having two exploration signals, one driving exploration across episodes and the other enforcing exploration within an episode. The former is a a state-space version of the exploration bias found in UCT whereas the latter aids the algorithm to escape (local) maxima of the value function. To extract the most knowledge possible from each experienced episode, an additional state-value function is learned, the simulation value-function. It is determined by simulating short-duration partial episodes that bootstrap from the actual value-function learned from real experience. Exlos decision making is extended by an online search process. Online search is adapted to continuous state-action spaces by considering it as a reinforcement learning problem defined for a local environment analogous to the original environment. Thus, any algorithm that operates in the original environment can potentially be used for online search. Online search is computationally expensive and must be seldom relied upon. In Exlos, computational efficiency is improved by invoking online search only when it is most crucial - when a (local) maxima at a non-terminal state.

Exlos operating over a tabular representation of the state-value function was compared to Q-learning, Q(λ), on-policy Monte-Carlo and Sarsa(λ). In small-scale maps, Exlos learned approximately optimal value functions faster than the remaining algorithms. Exlos was shown to learn effectively on a large map with continuous state and action spaces and temporally-extended actions. In that map, near-optimal policies may produce plans with a thousand or more actions. No other algorithms were tested in that environment but, to our knowledge, very few algorithms in existence can cope with such complexity.

Exlos was not implemented in an actual RTS-game environment; this does not detract from the developed methods and tests performed since the testing environment mimics most of what is expected from a RTS game.

References

BURO, M. 2004. Call for AI research in RTS games. In *Proceedings of the AAAI Workshop on AI in Games*, AAAI Press, 139–141.

ENTERTAINMENT SOFTWARE ASSOCIATION, 2009. 2009 sales, demographics and usage data. Essential facts about the computer and video game industry.

GUSMAO, A. 2011. *Reinforcement Learning in Real-Time Strategy Games*. Master's thesis, Aalto School of Science, Department of Information and Computer Science.

HSIEH, J.-L., AND SUN, C.-T. 2008. Building a player strategy model by analyzing replays of real-time strategy games. In *Proceedings of IEEE International Joint Conference on Neural Networks*, 3106–3111.

KOCSIS, L., AND SZEPESVÁRI, C. Bandit based Monte-Carlo planning. In *Proceedings of ECML 2006*, vol. 4212 of *LNCS*.

MAEI, H., SZEPESVÁRI, C., BHATNAGAR, S., SILVER, D., PRECUP, D., AND SUTTON, R. 2009. Convergent temporal-difference learning with arbitrary smooth function approximation. In *NIPS*, 1204–1212.

MAEI, H., SZEPESVÁRI, C., BHATNAGAR, S., AND SUTTON, R. 2010. Toward off-policy learning control with function approximation. In *ICML*, Omnipress, J. Fürnkranz and T. Joachims, Eds., 719–726.

MOLINEAUX, M., AHA, D. W., AND MOORE, P. 2009. Learning continuous action models in a real-time strategy environment. In *Proceedings of the Twenty-First International Florida Artificial Intelligence Research Society Conference (FLAIRS 2009)*.

SUTTON, R. S., AND BARTO, A. G. 1998. *Reinforcement Learning: An Introduction*. MIT Press.

WEBER, B., MAWHORTER, P., MATEAS, M., AND JHALA, A. 2010. Reactive planning idioms for multi-scale game AI. In *IEEE Symposium on Computational Intelligence and Games (CIG)*, 115–122.

Articles and Posters from the Open Call

Articles

Posters

Bag-of-features approach to unsupervised visual object categorisation

Teemu Kinnunen* Media technology, Aalto University

Abstract

The large and growing number of digital images is making manual image search laborious. Only a fraction of the images contain metadata that can be used to search for a particular type of image. Thus, a Ph.D. thesis by Kinnunen [2011] studied if it is possible to learn visual object categories directly from images. In this paper, the most interesting aspects of the thesis are summarised. There are various methods introduced in the literature to extract low-level image features and also approaches to connect these low-level features with high-level semantics. One of these approaches is called Bag-of-Features which is studied in the work. The goal is to find groups of similar images, e.g., images that contain an object from the same category. The standard Bag-of-Features approach is improved by using spatial information and visual saliency. It was found that the performance of the visual object categorisation can be improved by using spatial information of local features to verify the matches. The visual object categorisation performance was improved by using foreground segmentation based on saliency information, especially when the background could be considered as clutter.

Keywords: bag-of-features, self-organizing map, local feature, unsupervised visual object categorisation, spatial verification, saliency detection, computer vision

1 Introduction

The number of digital images has increased dramatically during the last decade. This originates from the popularity of digital cameras and the fact that nearly all mobile phones contain a built-in camera. The increasing number of images has lead to many image sharing services such as Flickr and Picasa, and also digital art sharing services such as DigitalArt and devianART. Nowadays, these image sharing services contain billions of images, e.g., Flickr alone already contains more than 6 billion images [Kremerskothen 2011]. Despite the many image sharing services, only a fraction of the images are stored on the image sharing services; the majority of the images are stored in the photographers' personal computers and mobile phones.

Because of such high number of images, it is not possible to manually browse through all the images to find a particular type of image. Therefore, the image sharing services provide an image search for the users, to search for images from the massive image collections by typing in keywords. However, all of these services have one serious limitation. The content of each image must be described using meta data, i.e., by giving tags as in Flickr, or by giving a representative name as in DigitalArt and devianArt, and uploading images to the correct predefined category. This causes two problems: i) Images need to be described manually which is laborious; ii) Descriptions of the images might vary significantly which makes the search impractical without intelligent cross-referencing keywords or use of taxonomies. For example, one might give the same tag for different kinds of images or give different tags for the same image.

One obvious solution to the manual image search problem is to use computers to organise and find a particular type of images because the computers offer a great amount of computational power and they never get exhausted. However, it is not a straightforward matter to use computers to search images because the computers store and process images as a long list of pixels which do not have a clear connection to high-level concepts which could be used to assist users to search images.

Smeulders et al. [2000] made a comprehensive study on Content Based Image Retrieval systems (CBIR) prior 2000. One of their contributions was that they divided the problem of recognising real world objects using visual information into two problems: the sensory gap and the semantic gap. The sensory gap was defined as the gap between the object in the real 3-D world and the captured 2-D image. When a real world object is captured into a 2-D image, some of the information is lost, e.g., we cannot be sure what is behind the object because of occlusions. The semantic gap was defined as the difficulty of connecting extracted low-level features with the high-level concepts. There are many methods of extracting low-level features such as edges [Canny 1986], lines and curves [Duda and Hart 1972], blobs [Lowe 2004], colour histograms, etc., but it is not self-evident how these low-level features should be connected to the high-level concepts. However, by defining these two gaps, researchers can concentrate on closing one of the gaps in their research. In this work, the focus is on the semantic gap, i.e., we have a set of images that we want to organise based on high-level concepts.

In Visual Object Categorisation (VOC), low-level features are extracted from the images and then connected to the high-level concepts. Many of the current VOC methods [Bar-Hillel and Weinshall 2008; Bart et al. 2008; Chum and Matas 2010; Csurka et al. 2004; Deng et al. 2010; Dueck and Frey 2007; Fulkerson et al. 2008; Grauman and Darrell 2006; van Gemert et al. 2008] are based on local features, particularly in Scale Invariant Feature Transform (SIFT) [Lowe 2004]. A local feature is a description of a detected region in the image. Description can be an $N \times N$ grey-level patch [Leibe et al. 2008] of a detected region or it can be a histogram of gradients [Lowe 2004]. The idea is that one can use local features to find similar regions from different images. The most trivial way to compute the similarity between the images is to compute the number of similar regions in the images using the local features [Grauman and Darrell 2006]. One popular approach using these local features to describe the content of the whole image originates from text document search, where documents are described as occurrences of a predefined vocabulary, i.e., a set words. This approach is called the Bag-of-Words approach [Blei et al. 2003; Lewis 1998]. In the VOC, visual words, i.e., local feature descriptors, are used instead of textual words. This approach is called the Bag-of-Features approach [Sivic and Zisserman 2003; Csurka et al. 2004] which is presented in Fig. 1.

2 Background

The most important works related to this work are [Csurka et al. 2004; Tuytelaars et al. 2010]. Csurka et al. [2004] introduce the BoF approach for VOC which is extended to UVOC in this work. In this work, we utilise the performance measure from Tuytelaars et al. [2010] and compare our results to the method in [Tuytelaars et al. 2010].

The BoF approach [Sivic and Zisserman 2003; Csurka et al. 2004] is illustrated in Fig. 1 and discussed more in detail in the next section. First, regions of local features are detected from images. Sec-

*teemu.kinnunen@aalto.fi

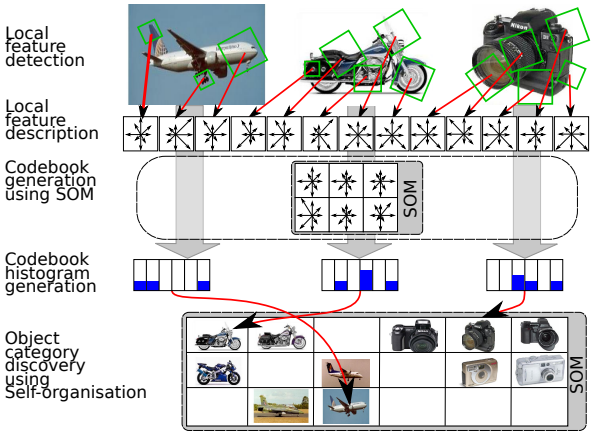


Figure 1: *Unsupervised image categorisation using the bag-of-features approach*

ond, these regions are converted into scale and rotation invariant descriptors in the local feature description step [Lowe 2004]. In the third step, a codebook is constructed using the descriptors of local features. In the study by Csurka et al. [2004], the codebook generation was performed during the training phase using the k-Means clustering algorithm. In the feature generation step, the extracted local features are matched against the generated codebook. A standard feature is the frequency vector over the codebook codes – "a bag of features". Finally, the given images are organised in the groups of similar images, e.g. using the Self-organising Map [Kohonen 1990] as in Fig. 1.

Albeit, the supervised VOC methods have been evolving rapidly and the performance of the state-of-the-art VOC method has increased dramatically in the annual Pascal VOC competitions [Everingham et al. 2010], the supervised VOC is now facing problems, especially when the number of classes is increased from tens to thousands. It is laborious to obtain training data for a large number of categories and also the supervised VOC has scalability issues as was shown by Deng et al. [2010]. Thus this work focuses on Unsupervised Visual Object Categorisation (UVOC). In this work, unsupervised learning methods, especially self-organisation, are studied in order to develop a new method for UVOC. The benefit of UVOC is that it does not need training images which can be too laborious to obtain. In UVOC, the goal is to find images that belong to the same group or category, i.e. images that contain an object from the same category.

Tuytelaars et al. [2010] made a comprehensive study about UVOC based on the BoF approach. They compared local feature detectors, normalisation methods, categorisation methods and different sizes of visual codebooks. They also introduced a new method for evaluating the performance of a UVOC which is also used in this work in addition to the method introduced by Sivic et al. [2008].

In this work, the BoF approach is used because in the supervised VOC it has shown superior performance [Song et al. 2011] and it is scalable. Moreover, Tuytelaars et al. [2010] used BoF in their UVOC experiments and showed that the baseline methods achieve state-of-the-art results. BoF contains some weaknesses, e.g. spatial information is not used in the basic method. The thesis [Kinnunen 2011] revisited and revised the standard parts of the BoF approach and the most important and interesting aspects are presented in this work.

3 Bag-of-features approach

Csurka et al. [2004] demonstrated how visual object categories can be learned using local features, clustering and supervised learning using the BoF approach. The BoF approach in UVOC is illustrated in Fig. 1, where given images are categorised using unsupervised learning. As in BoF, the first step for the VOC is a local feature detection where important local features are detected. Subsequently, these detected local features are described by using a local feature descriptor. A codebook is constructed using cluster centroids produced by a clustering method or using SOM [Kohonen 1990] node vectors as in our case. After this, the given image is described by matching extracted local features with the codebook and computing frequency of how many times each code has a match.

3.1 Local features

In this section, a few of the most popular local feature extraction methods are discussed and their characteristics are summarised at the end of the chapter. In the local feature extraction, local features are detected using a local feature detector and then described using a local feature descriptor. Thus, the result of local feature extraction is the spatial location of the detected region and description of the region. The local feature extraction is one of the key elements in visual object categorisation using the BoF approach because the descriptions of the detected regions are used to describe the appearance of the image. The selection of the local feature extractor has significant impact on categorisation accuracy [Mikolajczyk et al. 2005a; Mikolajczyk and Schmid 2005; Mikolajczyk et al. 2005b].

As mentioned earlier, local feature extraction consist of two steps: local feature detection (i.e. interest point detection) and local feature description (i.e. key-point description) [Lowe 2004]. In the first step, important regions are detected from an input image which are then described using a descriptor in the second step. The output of local feature extraction is a combination of spatial location information (x, y, scale (or scale-u, scale-v if an affine detector is used) and orientation) and region description of the appearance of the detected region [Mikolajczyk et al. 2005b].

A number of local feature detectors and descriptors have been proposed in the literature. A survey and comparison of different detectors can be found in the work by Mikolajczyk et al. [2005b] and for the descriptors in [Mikolajczyk and Schmid 2005]. These comparisons, however, are based on the repeatability and matching performances over different views of the same scenes. More explicit VOC evaluations have been carried out by Zhang et al. [2007] and Mikolajczyk et al. [2004]. Their main conclusions were that detector combinations performed better than any single detector, and that the extended versions of Scale Invariant Feature Transform (SIFT) [Lowe 2004] descriptor, the Gradient Location and Orientation Histogram (GLOH) [Mikolajczyk and Schmid 2005], is slightly superior to others in VOC. The better performance using the detector combinations can also be explained by the increased number of detected features. The drawback of GLOH is that it requires training data to estimate eigenvectors for the required PCA dimensionality reduction step – proper selection of the PCA data can explained the slightly better performance compared with the original SIFT.

3.2 Codebook generation

In the codebook generation step, extracted local features are used to form a codebook which is used to generate BoF histograms to describe the input images. In the original BoF study by Sivic et al. [2003] the k-Means clustering algorithm was used to cluster extracted local features and the cluster centroids were used as the

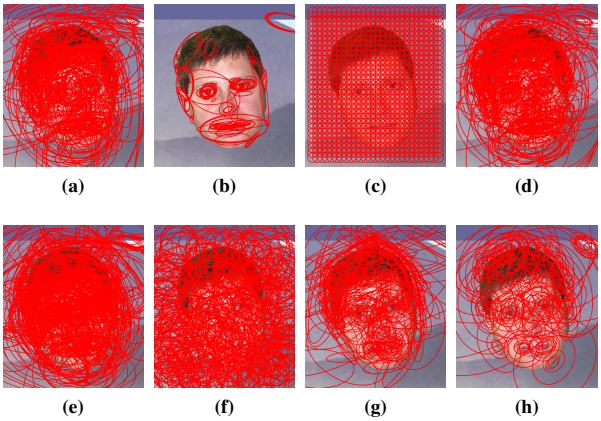


Figure 2: *Detected interest regions by using several methods and their different implementations: (a) Harris-Affine (FS); (b) MSER (FS); (c) Dense sampling; (d) Harris-Laplace (FS); (e) Hessian-Affine (FS); (f) DoG, i.e. SIFT (LV); (g) Harris-Laplace (LV); (h) Hessian-Laplace, basically Hessian-Affine, (LV). FS: implementation from the Feature Space web-site:* `http://www.featurespace.org`*; LV: implementation from the Lip-Vireo web-site:* `http://vireo.cs.cityu.edu.hk`*.*

codebook. In this section, a few alternative approaches for the codebook generation are discussed.

One family of algorithms for codebook generation are the ones typically used for data visualisation and exploration, such as the Multi-Dimensional Scaling (MDS) [Borg and Groenen 2005], Kohonen's Self-Organising Map (SOM) [Kohonen 1990], Isomap [Tenenbaum et al. 2000], and Locally Linear Embedding (LLE) [Roweis and Saul 2000]. These methods have similar properties, and therefore, in the work the one that can find a topological grouping of data points effectively is selected: the self-organising map and its public implementation, the SOM Toolbox [Alhoniemi et al. 2000]. The self-organising map has been successful compared with the k-Means algorithm in the experiments [Kinnunen et al. 2009].

3.3 Codebook histogram generation and normalisation

In the Bag-of-Features approach, images are described by matching extracted local features with the codebook in the feature generation step that is illustrated in Fig. 3. The process of describing the images with codebook histograms in the BoF approach [Csurka et al. 2004] can be described as follows: Let $\mathbf{D}$ be a set of local feature descriptors which are detected from an image using a local feature detector such as the Hessian-Affine [Mikolajczyk and Schmid 2002] and described using a local feature descriptor such as SIFT [Lowe 2004], and let $\mathbf{CB}$ be a codebook which contains N_{cb} codes. In practice, codes in the $\mathbf{CB}$ are clusters' centroids. Let N_{lf} be the number of local feature descriptors extracted from the image. After this, a BoF histogram f is generated according to the Bag-of-Features approach which is defined in Algorithm 1. The $Dist$ function calculates the Euclidean distance between two vectors. The smaller the distance, the greater the similarity is between the two vectors. Hence, a code that minimises the distance from a descriptor is chosen as the best match which has an index of bm.

Tuytelaars et al. [2010] made a comprehensive study of the effects of normalisation methods in UVOC. Their conclusion was that the

Algorithm 1 Codebook histogram generation using the Bag-of-Features approach

Require: $\mathbf{D}$, $\mathbf{CB}$
  $N_{cb} \leftarrow length(\mathbf{CB})$
  $\mathbf{f}_{1,\ldots,N_{cb}} \leftarrow 0$
  $N_{lf} \leftarrow length(\mathbf{D})$
  for all $\mathbf{d}_i = \mathbf{d}_1, \ldots, \mathbf{d}_{N_{lf}} \in \mathbf{D}$ **do**
    $bm \leftarrow \arg\min_j \ Dist(\mathbf{d}_i, \mathbf{CB}_j)$
    $\mathbf{f}_{bm} \leftarrow \mathbf{f}_{bm} + 1$
  end for
  return $\mathbf{f}$

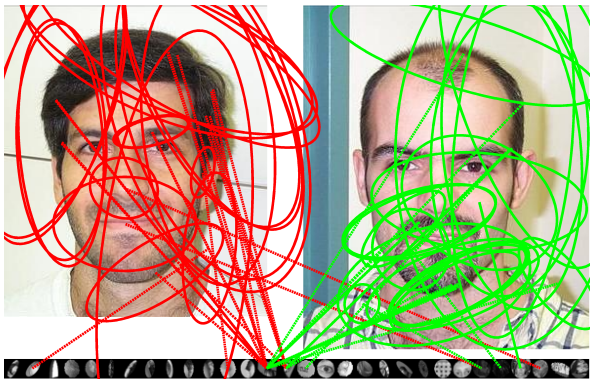


Figure 3: *Feature generation by matching local features with the codebook.*

L2-norm normalisation produces the best performance followed by the binarised-BoF. Their results show that the normalisation has a significant impact on the categorisation performance, thus it is also an important topic for discussion in this work.

In the code-wise normalisation, all bins of a certain code are normalised. In the binarised-BoF, the median of occurrences of each code is computed, and all bins below the median are set to zero, and all above to one. By binarising the BoF histograms, the BoF histograms should be more stable because small differences diminish in the normalisation. [Tuytelaars et al. 2010] In the Term Frequency - Inverse Document Frequency (TF-IDF) normalisation [Jones 1972], the number of occurrences of a code in an image (Term Frequency) is divided by the number of images containing the code (Inverse Document Frequency).

3.4 Image categorisation

In the final step of the BoF approach, the input images are categorised. In the unsupervised visual object categorisation, the codebook feature histograms (i.e. images) can be categorised by using any clustering method. One of the most popular methods is the k-Means clustering which is very simple, and thus, it can be used as a baseline method as in [Tuytelaars et al. 2010]. The goal of the work is to improve UVOC using BoF and the image categorisation step is as important as the other steps in the categorisation process. Thus, in this section other related categorisation methods are described and their performance in a typical UVOC task is evaluated.

The k-Means clustering algorithm has been used in many applications. One reason can be its simplicity, which is the reason why it is also used in the work as the baseline method for image categorisation. k-Means has been used earlier for UVOC by Tuytelaars et

al. [2010]. In their experiments, the k-Means clustering performed comparable to the other more complex methods. The k-Means clustering algorithm consists of two phases: a cluster assignment phase and a cluster updating phase. Initial cluster locations are usually chosen randomly. Then, each data point is assigned to its closest cluster (the cluster assignment phase). Next, cluster centroids are updated by computing the mean of data points belonging to a specific cluster (the cluster updating phase). This is repeated as long as the cluster centroids are changing or the maximum number of iterations is reached.

One possible method of categorising images is to use Self-Organizing Map (SOM) [Kohonen 1990]. In SOM, nodes on the SOM are organised so that similar nodes are closer to each other and dissimilar are further apart. It can be initialised randomly as in k-Means or by using some heuristics to obtain better initialisation, e.g., by computing the principal components and using them to give initial weights for the SOM nodes. After the initialisation, for each input sample $\mathbf{d}_1, \ldots, \mathbf{d}_{length(D)} \in \mathbf{D}$, the closest node, bm, (the Best Matching Unit (BMU)) from the codebook $\mathbf{CB}$ is searched and the weight of the BMU is changed so that it is moved towards the given data point. To maintain the topology, also BMU's neighbours (in the topology) are updated in such a way that the weights of the nodes that are closer in a topology to the BMU are changed more than the weights of the nodes that are further away. Neighbouring nodes for the BMU bm are search by finding nodes that are connected to the best match in the topology. [Kohonen 1990]

Neural Gas [Martinetz and Schulten 1991] is similar to the SOM, but in Neural Gas nodes are not organised in a topology. Instead of forcing the nodes in a predefined topology, the algorithm learns a structured manifold in the feature space which is defined based on the distances of the nodes in the input space. Thus, the learning algorithm is changed in the way that the neighbourhood of the BMU is computed using distances in the input space instead of spatial location in the topology as in SOM.

4 Spatial local feature matching

The standard BoF approach disregards all the spatial information. Therefore, the local features can be in any spatial configuration and the BoF histogram remains the same, even though the appearance of the image changes. One can think about the spatial configuration as an analogy to the order of the words in a paragraph. The order of the words is important to the interpretation of the text. Similarly the spatial configuration of the local features is important for VOC.

4.1 Proposed method

The spatial matching algorithm used in the work is based on the unsupervised landmark alignment algorithm introduced by Lankinen and Kamarainen [Lankinen and Kamarainen 2011]. However, instead of finding landmarks for a set of images from a specific category, the task here is to compute the distance between a pair of images based on the descriptors of spatially matching local features. Thus, their algorithm is used to find landmarks for a pair of images and compute the fitness of the landmarks. This information is used to define a distance between a pair of images instead of distances between two BoF codebook histograms.

The spatial matching step is computationally expensive, and thus, it is not possible to match all images against each other, especially when there are hundreds or thousands of images. One must choose candidate images carefully. Fortunately, we have the UVOC method based on BoF which can be used to find a list of candidate images for every given image. Using the BoF histograms of the images, it is trivial to find a sorted list of the most similar candidate images. A list of the N_{cand} best matching images can be given to the unsupervised landmark alignment algorithm by Lankinen and Kamarainen [2011] to spatially match the local features of the given image and candidate images, and to compute the fitness of the matching local features, which is used to define the distance between a pair of images. This approach is similar to the approach by Philbin et al. [2007] and Chum et al. [2011], but instead of recognising a specific object, the objective here is to detect the category of the object. Chum et al. [2011] did not use a hard threshold to limit the number of candidate images as is done here, instead they used an iterative method to select a cut for each query. In their method, the cut was made after 20 images in a row were predicted to be negative match. This approach was not studied in this work, but it is likely that it does not work, because it is possible that the 20 first images are from a wrong category, i.e., false matches.

The spatial matching algorithm for unsupervised landmark detection introduced by Lankinen and Kamarainen [2011] finds stable landmarks from a set of images by finding homography between a pair of images using local features and random sampling. At first, two local features are selected randomly from the seed image and then two matches are selected randomly from the top best matches in the candidate image. Then, homography is estimated using Direct Linear Transform (DLT) and spatial information of two correspondences. The homography is used to transform the local features from the candidate image to the seed image. Next, the local features that spatially match after the transformation are then matched using the descriptor part of the local features. If the local feature matches spatially and the distance between descriptors is "small", then the match is accepted. This random matching process is repeated many times in order to ensure that the correct solution (homography) is found with a reasonable confidence.

The spatial matching algorithm by Lankinen and Kamarainen [2011] returns the number of matching local features and distances between the matching local features descriptors which are used for defining the distance between the pair of images. The distance between the image pair is evaluated by computing a distance between spatially matching local features, choosing the N_{lm} best matches and computing $fScore$ the sum of the distances of the local feature descriptors of the best matches. In the case of supervised learning, $fScore$ can be used for deciding the class of the given image by choosing an image from the training set with the smallest distance to the unknown image and using its class information to predict the class of the unknown image. In unsupervised categorisation, $fScore$ can be used to find a sorted list of similar images to every given image.

In UVOC, it is not possible to match the images with candidate images with known labels. Thus, one needs to solve an image categorisation problem utilising spatial matching information $fScores$ without using labelling information. In the spatial matching, images are compared pairwise resulting in a matrix of pairwise distances. By using the pairwise distances, the images are sorted in ascending order. Next, a similarity matrix is constructed by setting the similarity value of image pair i and j as $S(i,j) = N_{cand}/rank(i,j)$, where N_{cand} is the number of images in the list of candidate images after the cut and $rank(i,j)$ is the index of image j in the list of similar images for the image i. The similarity matrix might not be symmetric because the spatial matching phase does not produce symmetric results. To fix the issue, the similarity matrix is made symmetric by refining each similarity value by $S'(i,j) = \max(S(i,j), S(j,i))$. This guarantees that the similarity matrix is symmetric. The final clustering result is computed by using the Normalised Cuts algorithm [Shi and Malik 2000].

5 Saliency detection

The way people perceive visual information has evolved during thousands of years of evolution. People can recognise thousands of objects quickly and accurately [Biederman 1987]. People perceive tremendous amounts of information through their visual system. However, only a fraction of the information is important. Thus, during the evolution, the vision system has evolved so that the focus can be changed quickly to detect important things. This "pop up" effect is called saliency [Itti and Koch 2001]. The motivation to study the saliency detection in this work is that it could be used to improve the VOC categorisation performance by detecting the foreground and using local features that are extracted from the foreground.

5.1 Improving category detection using salient regions

Saliency information can be combined with segmentation to choose an important region from an image to be used in categorisation. When the important region is detected successfully (assuming that it is the foreground), the image can be described more accurately because the codebook histogram contain only local features from the foreground (see, e.g., Fig. 4), and thus, the categorisation performance should be better. Cheng et al. [2011] have developed the Region-based Contrast (RCC) method that uses saliency information to detect important area from an image. Example results of the detected salient segments are shown in Fig. 4. The figure shows how the RCC segmentation method [Cheng et al. 2011] can detect foregrounds from the Randomised Caltech-101 images even though some of them have challenging backgrounds. The foregrounds, especially of the first three images, are detected accurately. The RCC segmentation fails on the car side image (Fig. 4j) because the saliency detector detects that the mountain is the most salient region as the mountain differs the most from its surroundings. According to the color contrast differences, it differs the most from the rest of the image, and thus it is incorrectly detected. There is a similar problem with the lotus image in Fig. 4m. The inner part of the lotus is labelled as salient, but the leaves of the flower do not differ significantly from the background, and thus, they are considered as non-salient. When this is given to the GrabCut segmentation algorithm [Rother et al. 2004], it segments the inner part of the flower and it is used as the predicted foreground. In this experiment, the data set is very challenging because it is artificial and the backgrounds can also contain salient objects. However, RCC is able to find the foreground in many cases, and thus, it can be used to detect foregrounds for VOC. An experiment using RCC predicted foregrounds is introduced in the following section.

6 Experiments

In the experiments carried out in the thesis [Kinnunen 2011] and presented here, Randomised Caltech-101 [Kinnunen et al. 2010] and Caltech-256 [Griffin et al. 2007] are used to evaluate the standard BoF UVOC method and proposed improvements. Performance of the methods is evaluated by computing a conditional entropy [Tuytelaars et al. 2010] defined as:

$$H(\mathbf{X} \mid \mathbf{Y}) = \sum_{y \in \mathbf{Y}} p(y) \sum_{x \in X} p(x \mid y) \log \frac{1}{p(x \mid y)} \quad . \qquad (1)$$

where $\mathbf{X}$ is the ground truth, $\mathbf{Y}$ is the predicted labels In addition to conditional entropy, the performance is evaluated by computing category average categorisation accuracy, $P(t, i)$, proposed by

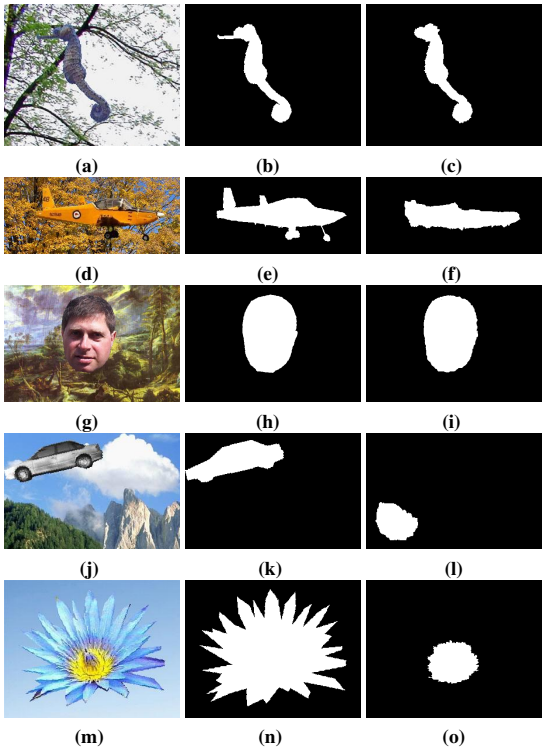


Figure 4: *Example outputs of Region-based Contrast (RCC) segmentation by Cheng et al. Images are from Randomised Caltech-101 The left column shows the original images; The middle column shows the ground truth foregrounds; The right column shows the RCC predicted foregrounds.*

Sivic et al. [2008] which is defined as follow:

$$P(t, i) = \frac{|X_i \cap Y_t|}{|X_i \cup Y_t|} \quad , \qquad (2)$$

where X_i are the ground truth images for category i and Y_t are the images assigned to node t. $|X_i \cap Y_t|$ is the number of images from category i assigned to the cluster t and $|X_i \cup Y_t|$ is the number of images belonging to the category i or being assigned to the cluster t. The average performance, $perf_{uvoc}$, is then defined as

$$perf_{uvoc} = \frac{1}{N_c} \sum_{i=1}^{N_c} \max_t P(t, i) \quad , \qquad (3)$$

where N_c is the number of categories.

With Randomised Caltech-101, the test was repeated 10 times with 5, 10, 20, 50, and 101 object categories. In each test, 30 images were chosen randomly from each category. With Caltech-256 dataset, all the images were selected from the 20 categories that were used in [Tuytelaars et al. 2010]. The Hessian-Laplace detector and SIFT descriptor were used to extract local features. A visual codebook was built using a SOM of size 200×1 as the default and codebook histograms were normalised using L2-normalisation.

6.1 Utilising spatial information

In this experiment, the performance of UVOC was measured using the BoF approach and the spatial matching approaches with

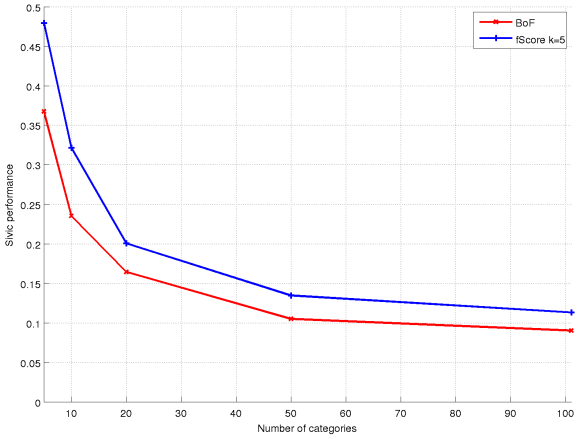


Figure 5: *Results using the standard BoF approach (red) and BoF approach with spatial matching (blue) with Randomised Caltech-101 imageset. Performance of the methods is evaluated using Sivic's method defined in Eq. (2).*

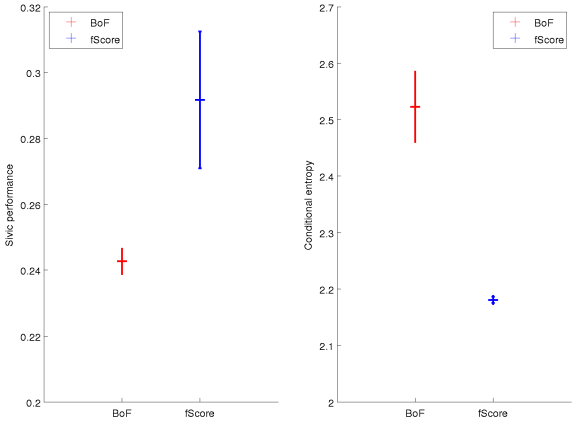


Figure 6: *Results using the standard BoF approach (red) and BoF approach with spatial matching (blue) with 20 categories from Caltech-256 image set. Performance evaluated using Sivic's method (Eq. (2)) on the left and using Conditional entropy (Eq. (1)) on the right.*

Randomised Caltech-101 dataset. The standard BoF method was used to generate a sorted list of candidate images for each given image. Then the spatial matching method was used to match all the given images with their $N_{cand} = 100$ best matching images, based on the Euclidean distances of BoF histograms. For each image pair, $fScore$ for $N_{lm} = 5$ was computed $fScore$s of pairs of image were used to build a similarity matrix that was given to the Normalised Cuts algorithm [Shi and Malik 2000] to obtain image categories. The results of the Randomised Caltech-101 UVOC experiment are shown in Fig. 5. The figure shows that the spatial matching improves the categorisation performance.

Fig. 6 shows that the performance can be improved significantly using the spatial matching. Both performance evaluation methods (i.e. Sivic performance Eq. (2) and Conditional entropy Eq. (1) show that the spatial matching improves the results significantly.

6.2 Using salient regions to improve categorisation

In the second experiment, saliency information was used to detect foreground segment from the input images which was used to select local features only from the foreground. Proposed approach was tested with Randomised Caltech-101 and Caltech-256 datasets and results are explained below.

In this experiment, salient regions were selected using the RCC detector. Codebook histograms that were generated from local features extracted from the salient region were given then to a 1-NN classifier which was used to test if salient regions could be used to improve VOC. 1-NN classifier was chosen instead of UVOC, because of its simplicity. Fig. 7 shows that the classification performance can be improved by choosing local features only from the salient segments of the images if the background does not contain relevant information about the object in the foreground as in the case of Randomised Caltech-101. However, the performance is inferior to the performance using local features only from the ground truth foregrounds. Randomised Caltech-101 image set is quite challenging especially for saliency detectors because the backgrounds can also contain salient objects.

In the experiment using the 20 selected categories from Caltech-256 image set, three different categorisation methods were tested: SOM [Kohonen 1990], Neural Gas [Martinetz and Schulten 1991], and k-Means using the full images and only the salient segments detected using RCC by Cheng et al. [2011]. The size of the SOM, number of nodes in Neural Gas and number of clusters in k-Means was set to the number of ground truth categories. Then the images were categorised using the clustering method with 20 clusters and the performance was measured by computing conditional entropy by Tuytelaars et al. [2010] as in Eq. (1). Results of the experiment comparing performance of the UVOC using full images and only the salient segments are shown in Fig. 8.

In Fig. 8, we can see that the categorisation performance was improved with RCC foreground detection when the size of the codebook is small (less than 500 words for SOM and less than 1000 words for k-Means and Neural Gas). When the size of the codebook was increased, the RCC foreground detection did not improve the categorisation performance. However, with k-Means and Neural Gas categorisation, the best performances were achieved with a small codebook and RCC foreground detection whereas with SOM categorisation and overall, the best performance was achieved with a larger codebook and without RCC foreground detection.

7 Conclusions

The number of digital images is huge and rapidly increasing both in the Internet and in the personally owned devices. The enormous number of images makes a manual image search laborious and slow. Thus, the thesis by Kinnunen [2011] studied automatic image categorisation. The most important and interesting aspects from the thesis were depicted in this work. A Bag-of-Features based framework was studied for the problem of unsupervised visual object categorisation because the Bag-of-Features approach has performed well in supervised visual object categorisation and Bag-of-Features can be scaled up to thousands of categories.

The performance of the basic unsupervised visual object categorisation using the Bag-of-Features approach suffers from false local feature matches in the feature generation step, and thus, codebook histograms can be confused between the images of different categories. In this work, the performance of the UVOC was improved by utilising spatial information to limit down the number of false

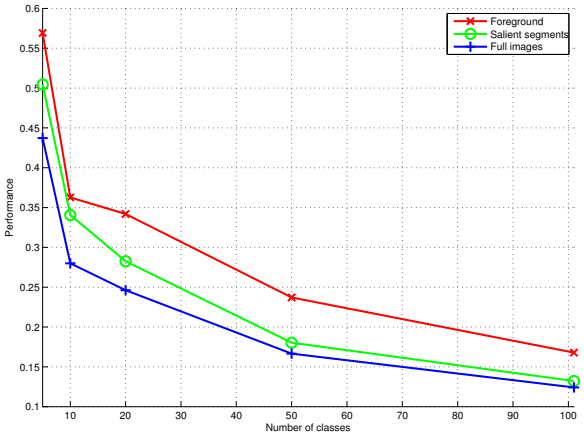


Figure 7: *Classification performance using the salient segment detector (RCC) using Randomised Caltech-101 image set: Performance using only local features from the foreground (red cross), from the salient segment (green circle), from whole image (blue plus).*

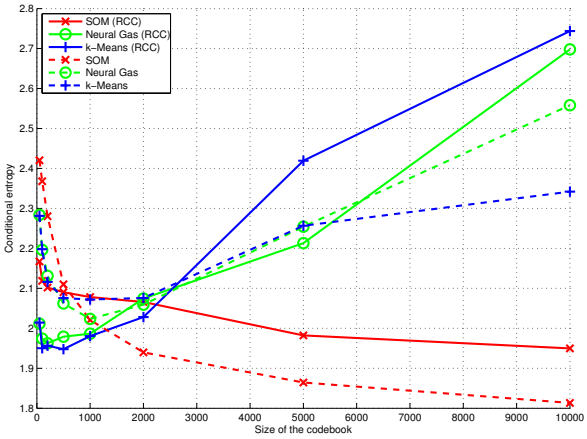


Figure 8: *Results of the UVOC experiment using RCC salient region detector using SOM (red cross), Neural Gas (green circle) and k-Means (blue plus) categorisation methods. Dashed lines denote results without RCC foreground detection and solid lines denote results with RCC foreground detection. Performance is computed using Conditional entropy Eq. (1) (lower is better).*

matches. In the spatial matching, also the spatial configuration of matching local features is verified. The spatial matching improved categorisation accuracy significantly, but it also increased computation dramatically. However, by choosing candidate images wisely using the Bag-of-Features method, the computational need can be kept reasonable.

Alternative approach, which was introduced, to limit the number of false matches was based on saliency information that was used choose only important local features. In the experiments, it was shown that salient region detection can significantly improve categorisation performance if the backgrounds do not contain important information about the foreground.

References

ALHONIEMI, E., HIMBERG, J., PARHANKANGAS, J., AND VESANTO, J., 2000. SOM Toolbox. http://www.cis.hut.fi/somtoolbox/.

BAR-HILLEL, A., AND WEINSHALL, D. 2008. Efficient learning of relational object class models. *International Journal of Computer Vision 77*, 175–198.

BART, E., PORTEOUS, I., PERONA, P., AND WELLING, M. 2008. Unsupervised learning of visual taxonomies. In *Proc. of Computer Vision and Pattern Recognition*.

BIEDERMAN, I. 1987. Recognition-by-components: A theory of human image understanding. *Psychological Review 94(2)*, 115–147.

BLEI, D. M., NG, A. Y., AND JORDAN, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, 4-5, 993–1022.

BORG, I., AND GROENEN, P. 2005. *Modern multidimensional scaling*, 2 ed. New York: Springer.

CANNY, J. 1986. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence 8*, 679–698.

CHENG, M., ZHANG, G., MITRA, N., HUANG, X., AND HU, S. 2011. Global Contrast based Salient Region Detection. In *Proc. of Computer Vision and Pattern Recognition*, 409–416.

CHUM, O., AND MATAS, J. 2010. Unsupervised discovery of co-occurrence in sparse high dimensional data. In *Proc. of Computer Vision and Pattern Recognition*, 3416–3423.

CHUM, O., MIKULIK, A., PERDOCH, M., AND MATAS, J. 2011. Total recall ii: Query expansion revisited. In *Proc. of Computer Vision and Pattern Recognition*, 889 –896.

CSURKA, G., DANCE, C., WILLAMOWSKI, J., FAN, L., AND BRAY, C. 2004. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning in Computer Vision*.

DENG, J., BERG, A., LI, K., AND FEI-FEI, L. 2010. What does classifying more than 10,000 image categories tell us? In *Proc. of European Conference on Computer Vision*, Springer, 71–84.

DUDA, R. O., AND HART, P. E. 1972. Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM 15*, 11–15.

DUECK, D., AND FREY, B. 2007. Non-metric affinity propagation for unsupervised image categorization. In *Proc. of International Conference on Computer Vision*, 1 –8.

EVERINGHAM, M., VAN GOOL, L., WILLIAMS, C. K. I., WINN, J., AND ZISSERMAN, A. 2010. The PASCAL Visual Object Classes (VOC) challenge. *International Journal of Computer Vision 88*, 2, 303–338.

FULKERSON, B., VEDALDI, A., AND S.SOATTO. 2008. Localizing objects with smart dictionaries. In *Proc. of European Conference on Computer Vision*.

GRAUMAN, K., AND DARRELL, T. 2006. Unsupervised learning of categories from sets of partially matching image features. *Proc. of Computer Vision and Pattern Recognition 1*, 19–25.

GRIFFIN, G., HOLUB, A., AND PERONA, P. 2007. Caltech-256 object category dataset. Tech. Rep. 7694, California Institute of Technology.

ITTI, L., AND KOCH, C. 2001. Computational modelling of visual attention. *Nature Reviews Neuroscience 2*, 3, 194–203.

JONES, K. S. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation 28*, 1, 11–21.

KINNUNEN, T., KAMARAINEN, J.-K., LENSU, L., AND KÄLVIÄINEN, H. 2009. Bag-of-features codebook generation by self-organization. In *Proc. of International Workshop on Self-Organizing Maps*.

KINNUNEN, T., KAMARAINEN, J.-K., LENSU, L., LANKINEN, J., AND KÄLVIÄINEN, H. 2010. Making visual object categorization more challenging: Randomized caltech 101 data set. In *Proc. of International Conference on Pattern Recognition*.

KINNUNEN, T. 2011. *Bag-of-features approach to unsupervised visual object categorisation*. PhD thesis, Lappeenranta University of Technology.

KOHONEN, T. 1990. The self-organizing map. *Proceedings of the IEEE 78*, 9, 1464–1480.

KREMERSKOTHEN, K., 2011. 6,000,000,000. http://blog.flickr.net/en/2011/08/04/6000000000/.

LANKINEN, J., AND KAMARAINEN, J.-K. 2011. Local feature based unsupervised alignment of object class images. In *Proc. of British Machine Vision Conference*.

LEIBE, B., ETTLIN, A., AND SCHIELE, B. 2008. Learning semantic object parts for object categorization. *Image and Vision Computing 26*, 15–26.

LEWIS, D. 1998. Naive (bayes) at forty: The independence assumption in information retrieval. In *Machine Learning: ECML-98*, Springer Berlin / Heidelberg, C. Nédellec and E. Rouveirol, Eds., vol. 1398 of *Lecture Notes in Computer Science*, 4–15.

LOWE, D. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision 20*, 91–110.

MARTINETZ, T., AND SCHULTEN, K. 1991. A "Neural-Gas" Network Learns Topologies. *Artificial Neural Networks I*, 397–402.

MIKOLAJCZYK, K., AND SCHMID, C. 2002. An affine invariant interest point detector. In *Proc. of European Conference on Computer Vision*, 128–142.

MIKOLAJCZYK, K., AND SCHMID, C. 2004. Scale & affine invariant interest point detectors. *International Journal of Computer Vision 60*, 63–86.

MIKOLAJCZYK, K., AND SCHMID, C. 2005. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*, 10, 1615–1630.

MIKOLAJCZYK, K., LEIBE, B., AND SCHIELE, B. 2005. Local features for object class recognition. In *Proc. of Computer Vision and Pattern Recognition*.

MIKOLAJCZYK, K., TUYTELAARS, T., SCHMID, C., ZISSERMAN, A., MATAS, J., SCHAFFALITZKY, F., KADIR, T., AND GOOL, L. V. 2005. A comparison of affine region detectors. *International Journal of Computer Vision 65*, 1/2, 43–72.

PHILBIN, J., CHUM, O., ISARD, M., SIVIC, J., AND ZISSERMAN, A. 2007. Object retrieval with large vocabularies and fast spatial matching. In *Proc. of Computer Vision and Pattern Recognition*, 1–8.

ROTHER, C., KOLMOGOROV, V., AND BLAKE, A. 2004. "grabcut": interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics 23*, 309–314.

ROWEIS, S., AND SAUL, L. 2000. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science 290*, 5500, 2323–2326.

SHI, J., AND MALIK, J. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 8, 888 –905.

SIVIC, J., AND ZISSERMAN, A. 2003. Video google: a text retrieval approach to object matching in videos. In *Proc. of Computer Vision and Pattern Recognition*, 1470 –1477.

SIVIC, J., RUSSELL, B. C., ZISSERMAN, A., FREEMAN, W. T., AND EFROS, A. A. 2008. Unsupervised discovery of visual object class hierarchies. In *Proc. of Computer Vision and Pattern Recognition*, 1–8.

SMEULDERS, A. W. M., WORRING, M., SANTINI, S., GUPTA, A., AND JAIN, R. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 1349–1380.

SONG, Z., CHEN, Q., HUANG, Z., HUA, Y., AND YAN, S. 2011. Contextualizing object detection and classification. In *CVPR*, 1585–1592.

TENENBAUM, J., DE SILVA, V., AND LANGFORD, J. 2000. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science 290*, 5500, 2319–2323.

TUYTELAARS, T., LAMPERT, C., BLASCHKO, M., AND BUNTINE, W. 2010. Unsupervised object discovery: A comparison. *International Journal of Computer Vision 88*, 2.

VAN GEMERT, J., GEUSEBROEK, J., VEENMAN, C., AND SMEULDERS, A. 2008. Kernel codebooks for scene categorization. In *Proc. of European Conference on Computer Vision*, 696–709.

ZHANG, J., MARSZALEK, M., LAZEBNIK, S., AND SCHMID, C. 2007. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision 73*, 2.

From Full-Reference to No-Reference in Quality Assessment of Printed Images

Tuomas Eerola*, Joni-Kristian Kamarainen, Lasse Lensu and Heikki Kälviäinen
Machine Vision and Pattern Recognition Laboratory (MVPR)
Department of Information Technology
Lappeenranta University of Technology (LUT)
Finland

Abstract

Measuring visual quality of printed media is important as printed products play an essential role in every day life. The image quality assessment has been an active research topic in digital image processing, but adapting the developed methods to measuring visual quality of printed media has been considered rarely and is not straightforward. In this work, different methods for the quality assessment of printed images are considered. First, the so called full-reference approach, where the original image with ideal quality is known, is presented. After that, the feasibility of using the reference-based approach for printed image quality assessments is discussed and problems related to the use of a digital reference image as the basis of the print quality analysis are shown. As a novel solution, a no-reference quality assessment approach is proposed. The solution based on a Bayesian network model of print quality is presented and its quantitative results are reported by using subjective data.

Keywords: print quality, image quality, quality assessment, full-reference, no-reference, Bayesian network, machine vision

1 Introduction

Despite the rapid development in electronic media, most people still prefer reading text printed on paper rather than reproduced on electronic displays [Imai and Omodani 2008]. Printed media can be also considered as more suitable for delivering localised news than electronic media. These, among other reasons, are why paper still has a notable role in communication, and the printed matter, such as books and newspapers, is an important part of daily life. When a customer purchases an image or printed product, one of the key factors is image quality [Engeldrum 2004]. Humans do not typically evaluate the quality of an image based on its physical parameters, but rather based on personal preferences and what they see as pleasurable [Engeldrum 2004].

The problem of how humans perceive the quality of a reproduced image is of interest to researchers of many fields related to vision science and engineering: optics and material physics, image processing (compression and transfer), printing and media technology, and psychology. The problem is especially difficult for the printed media since solving it requires understanding of paper and ink physics, visual measurements and optics and the human visual system. A measure for visual quality (print quality) cannot be defined without ambiguity because it is ultimately a subjective opinion of an "end-user" observing the result. Understandably, this evaluation has traditionally been conducted by human observers, but the recent development in computer and machine vision has made it intriguing to apply these methods to print quality evaluation. Machine vision utilises visual information reliably and may replace humans in certain laborious off-line evaluations. In addition, computational methods provide novel possibilities for on-line measurements for paper manufacturing and the printing industry.

*e-mail:tuomas.eerola@lut.fi

The image quality assessment methods are usually divided into three categories: full-reference (FR), reduced-reference (RR) and no-reference (NR) quality assessment methods. In the FR methods, the reference image with ideal quality is available, whereas in the RR methods only some information of the reference image is given as input to the algorithm. In the NR methods, the reference image is unknown. FR is the main approach for evaluating and comparing the quality of digital images, especially compressed images. The digital representations of the original image and compressed image are in correspondence, i.e., there exist no spatial transformations between the images, and the compression should retain at least photometric equivalence. Therefore, the FR metrics can be computed in a straightforward manner by computing "distance metrics". The NR image quality assessment is the most difficult task, and the majority of the proposed NR image quality assessment algorithms are designed for a single distortion type and are domain specific.

In this work, different methods for the quality assessment (QA) of printed images are presented. First, the FR assessment of printed images is considered by presenting the measurement framework proposed by the authors [Eerola et al. 2010] and by reporting the results for state-of-the-art digital image FR-QA algorithms. After that, the rationality of using the reference-based approach for printed image quality assessment and problems related to the use of a digital reference image are discussed. As a solution, a NR-QA algorithm based on the Bayesian network model and reported by the authors [Eerola et al. 2011] is presented.

2 Full-reference quality assessment of printed images

When the quality of a compressed image is analysed by comparing it to the original (reference) image, the FR metrics can be computed in a straightforward manner by computing "distance metrics". This is possible because digital representations are in direct correspondence, i.e., there exist no rigid, partly rigid or non-rigid (elastic) spatial shifts between the images, and the compression should retain at least photometric equivalence between the images. This is not the case with printed media, however. In modern digital printing, a digital reference exists, but the image data undergoes various irreversible transformations, especially in printing and scanning, before the other digital image for the comparison is established. In the following, the system where well-known methods are combined to form an automatic framework for analysing the full reference image quality of printed product is described.

The first important consideration is related to the scanning process. Since the focus is on print quality instead of scanning quality, the scanner must be an order of magnitude better than the printing system. Fortunately, this is not difficult to achieve with the available top-quality scanners in which the sub-pixel accuracy of the original can be achieved. It is important to use sub-pixel accuracy since it prevents the scanning distortions from affecting image registration. Furthermore, to prevent photometric errors, the scanner colour mapping should be adjusted to correspond to the original colour information. This can be achieved by using the scanner pro-

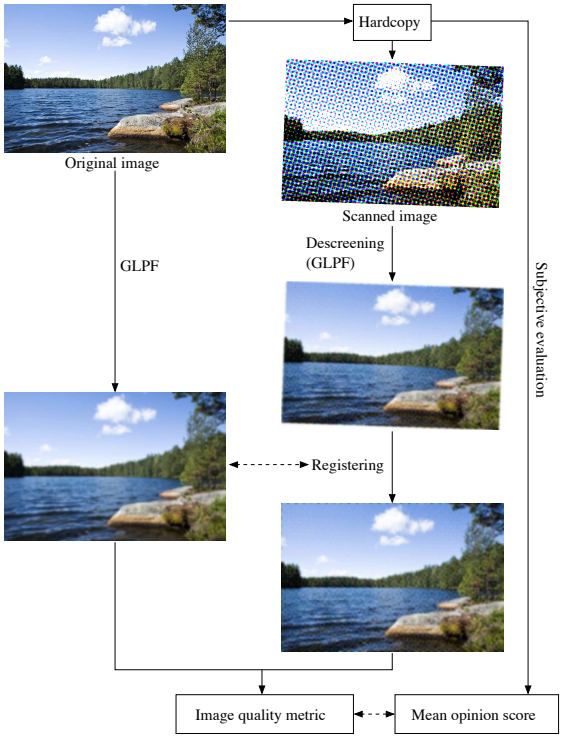


Figure 1: *The structure of the framework and data flow for computing full-reference quality assessment algorithms for printed images.*

filing software accompanying the high-quality scanners. Secondly, a printed image contains halftone patterns, and therefore, descreening is needed to remove the high halftone frequencies and form a continuous-tone image comparable to the reference image. Thirdly, the scanned image needs to be accurately registered with the original image before the FR-QA algorithm or dissimilarity between the images can be computed. The registration can be assumed to be rigid since non-rigidity is a reproduction error itself.

Based on the discussion above, it is possible to sketch the main structure of the framework [Eerola et al. 2010]. The framework structure and the data flow are illustrated in Fig. 1. Firstly, the printed halftone image is scanned using a colour-profiled scanner. Secondly, the descreening is performed using a Gaussian low-pass filter (GLPF) which produces a continuous tone image. To perform the descreening in a more perceptually plausible way, the image is converted to the CIE L*a*b* colour space in which the colour channels are filtered separately. The CIE L*a*b* spans a perceptually uniform colour space and does not suffer from problems related to, e.g., RGB where the colour differences do not correspond to the human visual system [Wyszecki and Stiles 2000]. Moreover, the filter cut-off wavelength is limited by the printing resolution and should not be higher than 0.5 mm, which is the smallest detail visually disturbing to the human eye when the unevenness of a print is evaluated from the viewing distance of 30 cm [Sadovnikov et al. 2005]. In ideal conditions, the acuity limit of the human eye can be as small as 0.017°, which corresponds to 0.1 mm [Wolfe et al. 2006]. To make the input and reference images comparable, the reference image needs to be filtered with an identical cut-off wavelength. The colour profiling of the scanner provides a "photometric registration" and the descreening a "physiological registration," and in the end, a spatial registration is needed.

2.1 Rigid image registration

Rigid image registration was considered as a difficult problem until the invention of general interest point detectors, and rotation and scale invariant descriptors. These methods provide an unparametrised approach to find accurate and robust correspondence which is essential for the registration. The most popular method which combines both interest point detection and description is David Lowe's scale-invariant feature transform (SIFT) [Lowe 2004]. The registration based on SIFT features has been utilised, for example, in mosaicking panoramic views [Brown and Lowe 2007]. The registration consists of the following stages: i) extract local features from the both images, ii) match the features (correspondence), iii) find a 2-D homography for the correspondences, and finally, iv) transform one image into the coordinate system of the other image.

The presented method performs a scale and rotation invariant extraction of local features using SIFT. The SIFT method also provides descriptors which can be used for matching. As a standard procedure, the random sample consensus (RANSAC) principle presented in [Fischler and Bolles 1981] is applied to find the best homography using exact homography estimation for the minimum number of points and linear estimation methods for all "inliers". The linear methods are robust and accurate also for the final estimation since the number of correspondences is typically quite large (several thousands of points). In the framework the implemented linear homography estimation methods are Umeyama for isometry and similarity [Umeyama 1991], and the restricted direct linear transform (DLT) for affine homography [Hartley and Zisserman 2003]. The only adjustable parameters in the method are the number of random iterations and the inlier distance threshold for RANSAC. The number of iterations can be can be seen as a trade-off between the computation time and the probability of successful registration. However, the homography estimation forms an insignificant proportion of the total computation time, and thus, a large value for the number of iterations is advisable. A tight inlier threshold was empirically shown to cause an unstable registration result and a rather high threshold value to still produce accurate registration [Eerola et al. 2009]. Based on these observations, the number of random iterations and the inlier distance threshold can be safely set respectively to 2000 and 0.7 mm (10 pixels with 360 dpi resolution). This makes the entire registration method practically free of parameters. For the image transformation, standard remapping using bicubic interpolation is utilised.

2.2 Image quality computation

In case of printed image quality assessment, the FR-QA algorithms contain special requirements. Although the above-mentioned registration works well small errors may occur. Because of this, simple pixel-wise distance formulations, such as the root mean square error (RMSE), do not work well. In other words, a good FR-QA algorithm should not be sensitive to such small registration errors. A more notable problem emerges from the subjective tests which are carried out by using printed (hardcopy) samples while the reference (original) image is in digital form. As a consequence, the reference image cannot be taken into the subjective evaluation and the evaluators do not usually see the actual reference. Therefore, those FR-QA algorithms that just compute simple similarity between the reference image and the input image do not succeed.

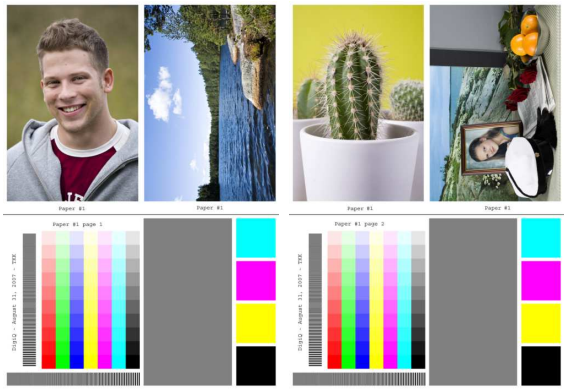


Figure 2: *The used test images and technical fields. The natural images used in subjective experiments and FR quality assessment algorithms are in the upper row. In the lower round are the technical test fields for objective (industrial) measurements.*

2.3 Experiments

2.3.1 Test sets

The objective of the study was to evaluate the effect of paper grade to the overall visual quality of printed images. Therefore, our test sets consisted of several paper grades at the cost of image contents. The set of test samples consisted of natural images printed with a production-scale electrophotographic printer on 21 different paper grades. The paper grades and the printing process were selected according to the current practices, as described in detail by the authors in previous publications [Oittinen et al. 2008; Eerola et al. 2008a; Eerola et al. 2008b]. The natural images used in the study are presented in Fig. 2. The image contents were selected based on current practices and previous experience in media technology, and they included typical content types such as objects with details (*cactus*), a human portrait (*man*) and a landscape (*landscape*). The fourth image content combined all the types (*studio*).

The printed samples were scanned using a high quality scanner with 1250 dpi resolution and 48-bit RGB colours. A colour management profile was devised for the scanner before scanning, and colour correction, descreening and other automatic settings of the scanner software were disabled. The digitised images were saved using lossless compression.

2.3.2 Subjective evaluation

The performance of the selected FR-QA algorithms was studied against the psychometrical subjective evaluations (subjective scores). The subjective evaluation procedure is described in detail by the authors in previous publication [Oittinen et al. 2008]. In brief, the sample images were attached on neutral grey frames with the size of A5 (148 x 210 mm). Evaluators were allowed to touch the frames but not the images. Samples of a specific image content were placed in a random order on a table, covered with a grey tablecloth. Labels with the numbers from 1 to 5 were also presented on the table. The evaluators were asked to select the sample image representing the lowest quality in the set and place it on the number 1. Then, the evaluator was asked to select the highest quality sample and place it on the number 5. After that, the evaluator's task was to place the remaining samples on the numbers so that the quality increased steadily from 1 to 5. The final subjective score was formed by computing the mean opinion scores (MOSs) over all evaluators (N=29). The level of illumination was 2200 lux with colour temperature 5000 K.

2.4 Processing of raw data

From the practical point of view, it is more interesting to properly order paper grades than to find the overall quality of a single printed image on some abstract quality scale. Therefore, the subjective evaluation as well as QA algorithm scores should be similar over different image contents for the same paper grade. The subjective evaluation results were always scaled to the interval of 1–5, but the image quality QA algorithm scores may differ significantly between the image contents. Therefore, either the QA algorithm scores need to be scaled to a common scale or the analysis needs to be done separately for different image contents. The first option was selected since the number of samples (21) was not enough to find statistically significant differences between the QA algorithms. Therefore, different image contents were combined to form a larger test set by scaling the QA algorithm scores. Here, the scaling was performed linearly. Let $\boldsymbol{x_n} = (x_{n,1}, ..., x_{n,M})$ represent the QA algorithm scores of one FR assessment for all samples $(1, \ldots, M)$ within a single image content n. Then, the linear model is

$$\hat{x}_{n,i} = \hat{\boldsymbol{b}}_{\boldsymbol{n}} \begin{pmatrix} 1 \\ x_{n,i} \end{pmatrix}, \tag{1}$$

where $\hat{\boldsymbol{b}}_{\boldsymbol{n}} = (b_{n,1}, b_{n,2})$ are selected by minimising the errors between the image contents as

$$\hat{\boldsymbol{b}}_{\boldsymbol{n}} = \arg\min_{\boldsymbol{b_n}} \sum_i [x_{1,i} - (b_{n,1} + b_{n,2} x_{n,i})]^2. \tag{2}$$

For the first image content, $\hat{\boldsymbol{b}}_{\boldsymbol{1}} = (0, 1)$, and for the remaining image contents, $\hat{\boldsymbol{b}}_{\boldsymbol{n}}$ are such that the QA algorithm scores are converted to values similar to the values of the first image content with the same paper grade.

2.4.1 Results

Two performance measures were used: the linear correlation coefficient between the MOS and QA algorithm score after nonlinear regression (NLCC) and the Spearman rank order correlation coefficient (SROCC). NLCC measures how well the QA algorithm scores and subjective scores correspond to each other while SROCC measures how well the QA algorithm orders the samples from the best to the worst. For NLCC, the following nonlinearity (constrained to be monotonic) was used [Sheikh et al. 2006]:

$$Q(x) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + \exp(\beta_2 x - \beta_3)} \right) + \beta_4 x + \beta_5, \tag{3}$$

where x is the modified algorithm score.

In Fig. 3 and Table 1, regression curves and performance measure values are presented for six different QA algorithms: Information fidelity criterion (IFC) [Sheikh et al. 2005], Multi-scale structural similarity metric (MS-SSIM) [Wang et al. 2003], Noise quality measure (NQM) [Damera-Venkata et al. 2000], Peak signal-to-noise ratio (PSNR), Visual information fidelity (VIF) [Sheikh and Bovik 2006], and Visual signal-to-noise ratio (VSNR) [Chandler and Hemami 2007]. As it can be seen, simple pixel-wise metrics, such as PSNR, do not perform well while the more advanced methods, such as, IFC, MS-SSIM, VIF and VSNR show high correlation coefficients (0.8-0.9) against the subjective evaluations. The SROCC values support these conclusions.

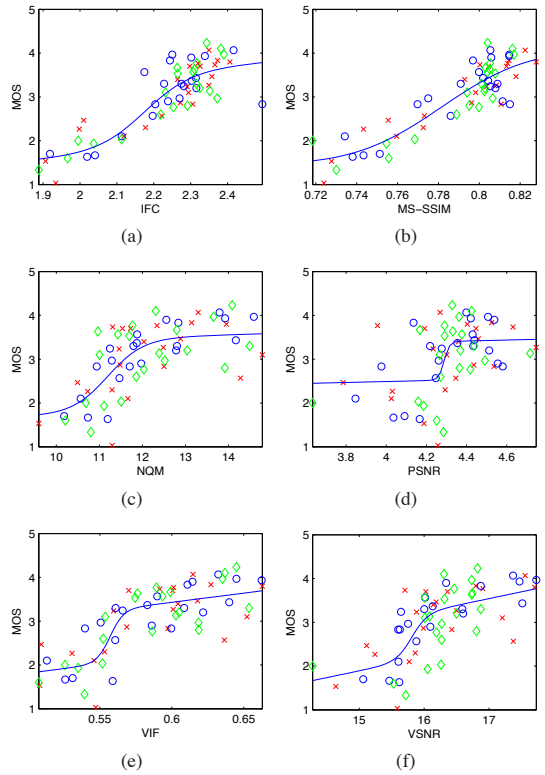


Figure 3: *Data and nonlinear regression curves (Test set B). The symbols represent the image contents: red x-mark - man, blue circle - lake, green diamond - cactus. (a) IFC; (b) MS-SSIM; (c) NQM; (d) PSNR; (e) VIF; (f) VSNR;*

2.5 Limitations of the full-reference approach

The full-reference image quality assessment has been shown to be good approach to computational evaluation of image quality, when the reference image exists. With a carefully designed measurement framework, it is possible to apply the FR approach also for printed images and relatively high correlations to subjective evaluation results can be achieved, as it was shown in the previous section. However, there exists several problems when quality of printed images is evaluated with the FR methods.

The first obvious weakness is the fact that the FR methods are suitable only when the digital reference image exists. This is not always the case with printed images.

Even more notable problems arise from the basic assumption of the FR approach: the reference image contain the ideal quality and it can be used as a basis for the quality evaluation. For the quality assessment of compressed images, this assumption is justified. A good image compression method reduces the size of the image in such manner that the visual appearance of the image changes as little as possible, i.e. the evaluated (compressed) image is visually similar to the reference (original) image. For printed images, however, it is not clear that the assumption is correct. First of all, the original image is in very different form than the printed image that is evaluated, making its use as the reference image not only difficult, but also rather questionable. It is not unambiguous how the difference between a printed photograph and a digital image should be measured. Secondly, it is not clear that the best quality is achieved when the original image is transferred unchanged to the paper. Even in a hypothetical ideal situation where the original image is of the "perfect quality" (what ever that is) on the display, the quality is not necessarily "perfect" after transferring the image visually unchanged on paper due to the different nature of the media. Thirdly, while making subjective evaluations with print samples for method development purposes, it is not often possible to show the digital reference image to the evaluators and the evaluators are forced to make the decisions without knowing what the printed image was supposed to look like.

Table 1: *Performance measures.*

QA algorithm	NLCC	SROCC
IFC	0.873	0.761
MS-SSIM	0.891	0.793
NQM	0.702	0.633
PSNR	0.547	0.428
VIF	0.818	0.738
VSNR	0.809	0.744

3 No-reference quality assessment

Because of the aforementioned reasons, it is justified to investigate NR quality assessment methods. Due to the obvious reasons, the NR quality assessment is much more difficult task than FR, and until recently there did not exist any general NR quality assessment methods. All of the proposed methods are either very application specific or measure only specific kind of distortion such as blur or noise. However, during the last few years, great developments have occurred and also more universal NR quality assessment algorithms have been established. These methods include curvelet, wavelet, and cosine transform based Hybrid No-Reference (HNR) model [Shen et al. 2011], natural scene statistics based BLind Image Integrity Notator (BLIINDS) [Saad et al. 2010], Distortion Identification-based Image Verity and INtegrity Evaluation (DIIVINE) index [Moorthy and Bovik 2011] and a (No-reference) Free-Energy-based Quality Metric (NFEQM) [Zhai et al. 2012].

None of the proposed methods is an universal NR image quality assessment method because they consider mainly the quality of digital images and the test sets for validation have been rather limited. The proposed methods produce a single scalar value that is intended to describe the visual overall quality. However, in realistic and challenging visual evaluation tasks involving aesthetic or even personal attributes, it is highly unlikely that the overall visual print quality can be measured with a single measurement and represented by a single scalar value [Keelan 2002]. Even in the restricted settings with artefactual or preferential attributes, human evaluators are likely to give different ratings for the same samples. As a possible resolution, a statistical Bayesian network model for quality assessment for printed images is presented.

4 Bayesian network model of print quality

The stochastic nature of perception and interpretation of visual information motivates to treat the overall quality and its attributes as probability distributions. For this purpose, the Bayesian theory provides a natural tool for modelling and analysis. A Bayesian network is an attractive choice since it is a probabilistic model that represents a set of random variables (instrumental measurements and subjective attributes) and their conditional independences with a directed graph. The idea of using Bayesian networks for mod-

elling visual quality is not completely new. In [de Freitas Zampolo and Seara 2004] and [Pulla et al. 2008], Bayesian networks were used to describe the overall image quality. However, these studies were not complete. In [de Freitas Zampolo and Seara 2004], a network was used to combine noise [Damera-Venkata et al. 2000] and distortion measures [de Freitas Zampolo and Seara 2003]. The work reported in [Pulla et al. 2008] was more similar to this work since the authors used the network to combine objective and subjective assessment data. The objective measurements were given as input values, and the overall image quality was viewed as a probability distribution of ratings. The preceding works did not consider the problem of how to establish the network structure automatically based on true data. Instead, they showed the potential of Bayesian networks to model image quality and similar phenomena.

The Bayesian network model presented here was originally reported by the authors in [Eerola et al. 2011]. The idea in [Pulla et al. 2008] was advanced by proposing a method which automatically optimises the structure of the Bayesian network for using it as a model of visual print quality. This was done by making elementary hypotheses about the behaviour of the overall quality with respect to the objective (instrumental) measures (prior) and by computing the model fitness through simulation. The structure optimisation method was a genetic algorithm mainly due to the complexity of the optimisation problem and to the need for simulation to evaluate the fitness of a solution. The main contribution was an evolve-estimate-simulate optimisation loop where the structure and connections are evolved using an evolutionary approach, network parameters estimated using the maximum likelihood rule, and the network performance evaluated using simulation. The final network forms statistical dependencies between the psychometric data and instrumental measurements. The network can be used as a unified model representing and explaining the phenomenon and as a more practical tool producing a single visual quality index (VQI) for any printed product.

4.1 Main structure and variables

The first step to construct a Bayesian network is to select the nodes, that is, the random variables. In this work, the basic structure depicted in Fig. 4 is used for the Bayesian network. On the left, the nodes represent the objective measures that are the essential external model inputs. In the figure, the rightmost node represents the overall quality, the network output in the form of a probability distribution within the range of possible values. The middle portion consists of the subjective attributes which represent the abstract quality concepts shared and used by the individuals to form the basis of overall quality. These attributes were identified in the human experiments using well-defined psychometric tests. The subjective attributes form an intermediate layer which transforms the objective measures into the probabilistic overall quality.

The arrows indicate the causality of the model. In the network, the subjective attributes are interpreted as "the reality" that is desired to be measured. On the one hand, the subjective attributes induce a certain measuring result (the objective measurements), and additionally, their combination forms the perceived overall quality. This is why the direction of the causality is from the subjective attributes to the objective measures as well as to the overall quality. However, constraining the direction of the causality towards the objective measures does not prevent the inference of subjective attributes and the overall quality based on the objective measures.

It is important to notice that the objective measures on the left in Fig. 4 can be accurately and repeatedly measured from printed photographs and test fields. The overall quality, or more precisely, its distributions from experiments with a jury of evaluators, can be estimated by carrying out psychometric experiments. From this viewpoint, the model produces the most likely distribution of the overall quality opinions of the evaluators, if the same material is physically presented to a number of them.

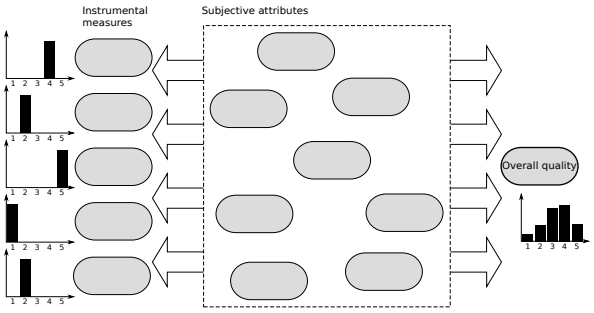


Figure 4: *The basic Bayesian network structure with example distributions of random variables representing the inputs and output of the model.*

The best objective measures for the left portion in Fig. 4 were selected according to the results of the prior works by the authors [Eerola et al. 2008a; Eerola et al. 2008b] where the most important instrumental and computational measures were surveyed and their relevance for explaining the overall quality was evaluated. The task was not possible by using the standard linear correlation; instead, non-linear relationships were evaluated and ranked using the proposed cumulative match score histogram (CMSH) [Eerola et al. 2008a]. The main idea behind the CMSH is the assumption that if two samples are visually perceived as being close to each other, they should be close to each other also based on the objective measurements. If a measure fails to meet the criterion, it was classified as irrelevant for subjective overall quality. Using the method, it was possible to rank the existing measures, and even exhaustively search for the optimal combinations of $N = 1, 2, \ldots, 6$ best measures. For digital printing (inkjet and electrophotography), the following six measures were selected: *computational mottling* [Sadovnikov et al. 2007], *colour gamut*, *mean colour density*, *print gloss*, *edge blurriness* and *edge raggedness*. This result is well in accordance with the current practises: these measures are commonly used in paper mill laboratories as well.

The selection of subjective attributes was based on systematic interviews of evaluators during the far-reaching subjective experiments. As a standard psychological interview technique [Radun et al. 2008] the evaluators were asked to describe visual factors that affected their ratings after they had given a rating for overall quality for each image. Later, a common vocabulary was established from the factors by using manual search, frequency analysis, and term mappings, and it was revised in the next independent experiments. Specifically, the most common subjective attributes were as follows: *naturalness*, *clarity*, *colourfulness*, *subjective gloss*, *graininess*, *lightness*, *contrast*, and *sharpness*. It should be noted that these subjective attributes do not necessarily correspond to their physical analogues since the semantic meaning of a term varied between the evaluators. This is typical for the natural, fuzzy concepts that naïve evaluators use in their everyday speech in contrast to the well-defined concepts used by the professionals. This difference is not necessarily caused by the incorrect use of the terms, but by the fact that visual impressions of contrast, sharpness, naturalness etc. are not unambiguously related to any physical property of an image. For example, higher colour saturation may make the image look subjectively sharper, although the use of these concepts is separated among professionals. For this reason, the graph edges cannot

be formed manually, but the relationships need to be learned.

4.2 Structure learning

Learning the optimal structure for a Bayesian network has been shown to be NP-complete [Chickering 1996]. As a consequence, full search methods are infeasible. Moreover, the laborious nature of collecting subjective data severely limits the available amount of training data. Therefore, also most heuristic methods, such as the PC algorithm [Spirtes and Glymour 1991], are not applicable. The structure optimisation is, however, essential for solving the problem and needs to be implemented into the learning process.

In the case of print quality modelling, it is possible to form a number of hypotheses on how the model should behave. For example, if the undesired solid printed area unevenness (mottling) increases while the other objective measures remain the same, the overall quality should decline. Similarly, if the colour gamut (a subset of colours a paper grade can reproduce with the available inks) expands, then the overall quality should improve. Using these heuristic and intuitively correct regulation rules, it is possible to produce a scalar value representing how logically correct a model is, that is, by randomly pruning how well model behaviour follows the hypotheses. This leads to a complex optimisation task: finding such a Bayesian network structure that the model behaves as consistently as possible after its parameters have been estimated using the training data. In this learning scheme, a network is not evaluated according to how well it represents the training data, but how well it represents the prior knowledge after the estimation with the training data. Therefore, the prior knowledge of behaviour acts as a regularisation term which enables the optimisation process with a small number of data points. The structure optimisation method is presented in more detail in [Eerola et al. 2011].

4.3 Experiments

To test the Bayesian network approach, the same test set was used as in the FR experiments (Sec. 2.3). The objective measures were computed from the technical test fields (the lower row in Fig. 2), and the subjective evaluation was carried out using the three natural image contents (human portrait, landscape and cactus). The subjective evaluations were performed using the procedure presented in Sec. 2.3.2. In addition to the overall quality, the numerical scale was revealed also for the subjective attributes by asking human evaluators to label the samples based on a single attribute, such as sharpness or graininess. The subjective evaluation was conducted separately for each image content, and the number of evaluators was 29, so the number of training samples was $21 \times 3 \times 29 = 1827$. However, it should be noted that the objective measures were constant for each paper grade, and thus, the training data is extensive only for the subjective part of the model (only 21 different combinations of objective measures).

The initial population for the genetic algorithm consisted of 20 educated guesses, 20 fully random networks, and 20 partly educated guesses (some of the edges manually selected). Due to the long computation time, the number of simulations needed to evaluate the fitness of network structure, was set to small, and thus, the error margin for the fitness function value was relatively large. Therefore, a list of the 100 best structures was maintained during the optimisation process. Moreover, due to the stochastic nature of genetic algorithm, the structure learning was repeated 10 times resulting 1000 network structures. All the 1000 networks were evaluated against the subjective MOS using leave-one-out cross-validation. The best structure according to the correlation coefficient between the model output and MOS is shown in Fig. 5. In Fig. 6, the correlation against the subjective evaluation is plotted. The expectation

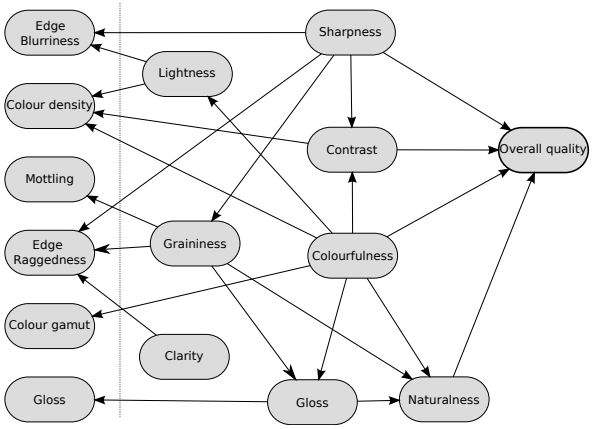


Figure 5: *The best Bayesian network structure found.*

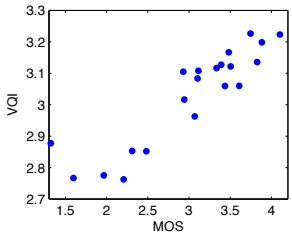


Figure 6: *Correlation between the model-produced visual quality index (distribution expectation) and subjective MOS. Correlation computed using the leave-one-out cross-validation is 0.90.*

values of the overall quality were used as the visual quality index (VQI). For a comparison, a tree-structured network learned using Chow-Liu algorithm [Pearl 1988] was also tested. With a correlation coefficient of 0.75 against MOS, the tree-structured network was outperformed by the network found using proposed structure optimisation method. A more extensive analysis of the found models can be found in [Eerola et al. 2011].

5 Conclusions and future work

In this work, different approaches to estimate the overall quality of printed images were presented. First, the full-reference approach was presented and analysed. The problems related to the full-reference approach, most notably the unclear definition of the reference (image), was discussed and investigation towards no-reference approaches justified. Different methods to no-reference quality assessment were shortly presented. A Bayesian network model of printed image quality was introduced and was shown to predict well the subjective human evaluations.

The main advantage of the Bayesian network approach is its versatility. The factors of low quality are easier to establish, since numerical values for the instrumental measurements are known. In comparison, the FR quality measures usually return only one scalar value (some measures return also a dissimilarity map) that tells whether the overall quality is high or low. The Bayesian network helps us also to understand the perceived quality as a phenomenon. The structure of the network gives information about the relations of the objective measurements to the subjective attributes, and about the relations of the subjective attributes to each other. In addition,

the Bayesian network works with incomplete measurements, i.e., only a portion of the objective measurements can be used as the evidence for predicting the overall quality. To be more specific, in the Bayesian network any nodes can be used as evidence to predict the value of any other node. This enables, for example, to fix the desired overall quality and examine the distribution of one instrumental measure with a certain combination of other instrumental measures.

The presented Bayesian network was a proof-of-concept, and more work is needed to make the approach an indispensable tool for the image quality assessment. The model needs certain technical test fields to be printed in order to define the instrumental measures used as an input. This is why the current model cannot be considered as a complete NR method. The next stage is to find alternatives for the instrumental measures, such as blur and noise measures, that can be measured directly from a printed photograph. Measuring the model input directly from natural images will lead to a more general image quality assessment method that has significant potential also for the quality assessment of digital images.

Acknowledgements

The authors would like to thank Raisa Halonen and Prof. Pirkko Oittinen from the Department of Media Technology in Helsinki University of Technology for providing the test material, Tuomas Leisti and Prof. Göte Nyman from the Department of Psychology in the University of Helsinki for providing the subjective evaluation data, and Prof. Risto Ritala from the Department of Automation Science and Engineering in Tampere University of Technology for contributions related to Bayesian networks. The authors would like to thank also the Finnish Funding Agency for Technology and Innovation (TEKES) and partners of the DigiQ project (TEKES Project No. 40176/06) for support.

References

BROWN, M., AND LOWE, D. G. 2007. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision 74*, 1, 59–73.

CHANDLER, D. M., AND HEMAMI, S. S. 2007. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Transactions on Image Processing 16*, 9 (September), 2284–2298.

CHICKERING, D. M. 1996. *Learning from Data: Artificial Intelligence and Statistics V*. Springer, ch. Learning Bayesian networks is NP-complete, 121–130.

DAMERA-VENKATA, N., KITE, T. D., GEISLER, W. S., EVANS, B. L., AND BOVIK, A. C. 2000. Image quality assessment based on a degradation model. *IEEE Transactions On Image Processing 9*, 4 (April), 636–650.

DE FREITAS ZAMPOLO, R., AND SEARA, R. 2003. A measure for perceptual image quality assessment. In *Proceedings of the International Conference on Image Processing (ICIP'03)*, 433–436.

DE FREITAS ZAMPOLO, R., AND SEARA, R. 2004. Perceptual image quality assessment based on bayesian networks. In *In proc. of the International Conference on Image Processing (ICIP'04)*, 329–332.

EEROLA, T., KAMARAINEN, J.-K., LEISTI, T., HALONEN, R., LENSU, L., KÄLVIÄINEN, H., NYMAN, G., AND OITTINEN, P. 2008. Is there hope for predicting human visual quality experience? In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*.

EEROLA, T., KAMARAINEN, J.-K., LEISTI, T., HALONEN, R., LENSU, L., KÄLVIÄINEN, H., OITTINEN, P., AND NYMAN, G. 2008. Finding best measurable quantities for predicting human visual quality experience. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*.

EEROLA, T., KAMARAINEN, J.-K., LENSU, L., AND KÄLVIÄINEN, H. 2009. Framework for applying full reference digital image quality measures to printed images. In *16th Scandinavian Conference on Image Analysis*, 99–108.

EEROLA, T., LENSU, L., KÄLVIÄINEN, H., KAMARAINEN, J.-K., LEISTI, T., NYMAN, G., HALONEN, R., AND OITTINEN, P. 2010. Full reference printed image quality: Measurement framework and statistical evaluation. *Journal of Imaging Science and Technology 54*, 1 (January/February), 1–13.

EEROLA, T., LENSU, L., KAMARAINEN, J.-K., LEISTI, T., RITALA, R., NYMAN, G., AND KÄLVIÄINEN, H. 2011. Bayesian network model of overall print quality: Construction and structural optimisation. *Pattern Recognition Letters 32*, 11, 1558–1566.

ENGELDRUM, P. G. 2004. A theory of image quality: The image quality circle. *Journal of Imaging Science and Technology 48*, 5 (September/October), 446–456.

FISCHLER, M. A., AND BOLLES, R. C. 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Graphics and Image Processing 24*, 6.

HARTLEY, R., AND ZISSERMAN, A. 2003. *Multiple View Geometry in Computer Vision*, second ed. Cambridge Universiy Press.

IMAI, J., AND OMODANI, M. 2008. Reasons why we prefer reading on paper rather than displays: Studies for seeking paper-like readability on electronic paper. *Journal of Imaging Science and Technology 52*, 5 (June), 1–5.

KEELAN, B. W. 2002. *Handbook of Image Quality: Characterization and Prediction*. Marcel Dekker Inc.

LOWE, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision 60*, 2, 91–110.

MOORTHY, A. K., AND BOVIK, A. C. 2011. Blind image quality assessment: From scene statistics to perceptual quality. *IEEE Transactions on Image Processing 20*, 12, 3350–3364.

OITTINEN, P., HALONEN, R., KOKKONEN, A., LEISTI, T., NYMAN, G., EEROLA, T., LENSU, L., KÄLVIÄINEN, H., RITALA, R., PULLA, J., AND METTÄNEN, M. 2008. Framework for modelling visual printed image quality from paper perspective. In *Image Quality and System Performance V*.

PEARL, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.

PULLA, J., METTÄNEN, M., KONKARIKOSKI, K., AND RITALA, R. 2008. Bayesian network model as an overall image quality measurement system. In *Proc. of the 12th IMEKO TC1-TC7 Joint Symposium on Man Science and Measurement*.

RADUN, J., LEISTI, T., HÄKKINEN, J., OJANEN, H., OLIVES, J.-L., VUORI, T., AND NYMAN, G. 2008. Content and quality:

Interpretation-based estimation of image quality. *ACM Transactions on Applied Perception 4*, 4.

SAAD, M. A., BOVIK, A. C., AND CHARRIER, C. 2010. A DCT statistics-based blind image quality index. *IEEE Signal Processing Letters 17*, 6, 583–586.

SADOVNIKOV, A., SALMELA, P., LENSU, L., KAMARAINEN, J., AND KÄLVIÄINEN, H. 2005. Mottling assessment of solid printed areas and its correlation to perceived uniformity. In *14th Scandinavian Conference of Image Processing*, 411–418.

SADOVNIKOV, A., LENSU, L., AND KÄLVIÄINEN, H. 2007. Automated mottling assessment of colored printed areas. In *Proceedings of the 15th Scandinavian Conference on Image Analysis*, 621–630.

SHEIKH, H. R., AND BOVIK, A. C. 2006. Image information and visual quality. *IEEE Transactions On Image Processing 15*, 2 (February), 430–444.

SHEIKH, H. R., BOVIK, A. C., AND DE VECIANA, G. 2005. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions On Image Processing 14*, 12 (December), 2117–2128.

SHEIKH, H. R., SABIR, M. F., AND BOVIK, A. C. 2006. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions On Image Processing 15*, 11 (November), 3440–3451.

SHEN, J., LI, Q., AND ERLEBACHER, G. 2011. Hybrid no-reference natural image quality assessment of noisy, blurry, jpeg2000, and jpeg images. *IEEE Transactions on Image Processing 20*, 8, 2089–2098.

SPIRTES, P., AND GLYMOUR, C. 1991. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review 9*, 1, 62–72.

UMEYAMA, S. 1991. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence 13*, 4, 376–380.

WANG, Z., SIMONCELLI, E. P., AND BOVIK, A. C. 2003. Multiscale structural similarity for image quality assesment. In *Proceedings of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, vol. 2, 1398–1402.

WOLFE, J. M., KLUENDER, K. R., AND LEVI, D. M. 2006. *Sensation & Perception*. Sinauer Associates.

WYSZECKI, G., AND STILES, W. S. 2000. *Color science : concepts and methods, quantitative data and formulae*, second ed. Wiley.

ZHAI, G., WU, X., LIN, W., AND ZHANG, W. 2012. A psychovisual quality metric in free-energy principle. *IEEE Transactions on Image Processing 21*, 1, 41–52.

MEG Mind Reading: Strategies for Feature Selection

Heikki Huttunen, Tapio Manninen and Jussi Tohka
Department of Signal Processing, Tampere University of Technology

Abstract

The regularized logistic regression classifier has shown good performance in problems where feature selection is critical, including our recent winning submissions to the ICANN2011 MEG mind reading challenge [Huttunen et al. 2011; Huttunen et al. 2012], and to the DREAM 6 AML classification challenge [Manninen et al. 2011]. The benefit of the method is that it includes an embedded feature selection step, which automatically selects a good subset of input features, thus, simplifying the classifier and improving the generalization. However, explicit wrapper feature selection methods, such as the forward and backward feature selection, are also widely used in similar problems. In this paper, we compare the efficiency of the elastic net regularized logistic regression classifier with the support vector machine classifier in combination with various sequential feature selection methods.

Keywords: Classification, feature selection, MEG, Simulated Annealing, Logistic Regression, Elastic Net

1 Introduction

The task in the supervised classification is to make predictions about the class of an unknown object (represented by a feature vector $\mathbf{x}$) given a training set of P-dimensional feature vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ with known class memberships. An important special case of supervised classification problems arises when the number of features P is larger or nearly as large than the number of training samples N. These classification problems are increasingly important in genomics and neuroimaging [Saeys et al. 2007; Pereira et al. 2009]. These setups require efficient methods for feature selection, i.e., the selection of the subset of the most relevant features among the P possible features. To give a concrete example of the nature of the problem, we outline data used in this paper. The data is drawn from a competition for classification of brain magnetoencephalography (MEG) data that was organized together with the ICANN 2011 conference [Klami et al. 2011]. The task was to train a classifier for predicting the movie being shown to the test subject. There were five classes and the data consisted of 204 channels. Each measurement was one second in length and the sampling rate was 200 Hz. From each one-second measurement, the competitors had to derive discriminative features for classification. Since there were only a few hundred measurements, the number of features (40800) easily exceed the number of measurements (i.e., the size of the training set) and, thus the key problem is to select the most relevant features efficiently.

Our method based on regularized logistic regression was the most accurate among ten submissions to the competition [Huttunen et al. 2011]. While preparing for the submission, we tested various iterative feature selection methods including the forward stepwise selection and backward stepwise selection [Pudil et al. 1994] and Simulated Annealing Feature Selection (SAFS) [Debuse and Rayward-Smith 1997], but obtained the best results using Logistic Regression with elastic net penalty leading to a joint algorithm for classifier design and feature selection. However, due to lack of time, these experiments were not systematic, but rather intuitive tests aimed at finding a direction where to proceed with the challenge submission.

In our recent paper, we discovered that the ℓ_1-regularization based feature selection outperforms both forward selection and simulated annealing feature selection for interpreting brain signals when formulated as a regression problem [Kauppi et al. 2011]. In this paper we extend the study for classification problems, and compare regularized logistic regression with sequential forward selection (SFS), sequential backward selection (SBS), and SAFS that have been combined with support vector machine (SVM) classifiers.

Before moving on, we briefly explain the central role of careful analysis of the feature selection approaches to neuroimaging studies. In functional neuroimaging studies such as described above, the point of the true importance is not the prediction of the stimulus type based on the imaging as such (we know the true type of the stimulus), but the analysis of the classifier leading to a good prediction of the stimulus type[1]. The idea is that if the classifier makes good predictions about the stimulus type then it is also informative about the distinctions of the neural representations of the two or more behavioral tasks. Especially, when the classifier is linear and sparse (i.e., uses only a small subset of all available features), the localization of the system of brain regions responding differently to the tasks becomes possible. The major benefit of this classification approach termed multivoxel pattern analysis over the more traditional method based on massively univariate statistical hypothesis testing (see, e.g., [Smith et al. 2004]) is that it (in principle) allows the identification of the set of voxels whose activity is diagnostic for engagement of a particular task while the traditional approach allows the identification of the set of voxels that are activated by a task [Poldrack et al. 2009]. The accuracy of inferences made based on the classifiers is bounded by the accuracy of the classifiers used for making the inferences. Since the feature selection is the most central component of the classifier design in these applications, it is important to understand the relative performance of the different feature selection schemes.

The reasons for using feature selection are two-fold: On one hand, using only a subset of features tends to improve the classification performance, and on the other hand, recognizing the significant features may provide insight on the mechanisms of the underlying phenomenon. For example, in our case, the MEG channels related to selected features are of interest.

In the rest of this paper, we first describe the studied feature selection and classification methods in detail in Section 2. In Section 3, we describe the data we are using. Section 4 considers the question of overlearning the training data and how to avoid the pitfalls by elaborate validation of classification performance. In Section 5, we study the efficiency of alternative classification and feature selection methods in a systematic manner, and compare them with regularized logistic regression used in our winning submission. Finally, Section 6 discusses the results and draws some conclusions on the results.

2 Feature Selection Methods

The problem of selecting the best features among all available measurements is computationally very difficult. If there are altogether

[1] There are neuroimaging applications of the supervised classification where the main interest is the prediction in itself such as brain computer interfaces [Blankertz et al. 2010] and diagnosis of neurodegenerative diseases based on imaging data [Wolz et al. 2011]

P input features, the number of subsets is 2^P, which easily becomes prohibitively large for exhaustive search. For example, in the experiments of Section 5, the number of features is $P = 408$, and there are $2^{408} \approx 7 \times 10^{122}$ subsets. If the feasibility of one subset could be tested as fast as a single clock cycle of a 3 GHz processor, exhaustive search would take 7×10^{105} years, i.e., approximately 2×10^{88} times the age of the universe. Alternatively, exploiting the earth's entire computational power and assuming 58 % annual growth [Hilbert and López 2011], the exhaustive search would take approximately 517 years.

Feature selection methods can be divided in two general categories, which are the *filter model* and the *wrapper model*. The wrapper model is the more intuitive of these: wrapper algorithms iteratively test the classifier performance with some or all of the 2^P subsets. The obvious problem is in the computational complexity especially with large data sets. Moreover, the best features for some classification algorithm might not be optimal for another one.

The objective of the filter model is to find general characteristics of the data to evaluate feature subsets without involving any classification algorithm. The statistical fitness of a feature can be evaluated with many different measures, of which popular ones include correlation-based measures and the Fisher ratio, i.e., the ratio of between-class-variance and the within-class-variance. The main drawback of the filter model is that the produced ranking of features does not take into account their joint predictive power, but instead treats them individually. In this respect, the wrapper model is typically better in finding features complementary to each other.

Because of these reasons, our study focuses on different wrapper-like algorithms. There exists vast literature on the topic, and the approaches can be roughly divided into four categories [Liu and Yu 2005].

- **Exhaustive search** is the simplest algorithm, with the obvious drawback of computational cost. However, the search space can be limited with different heuristics, such as the *branch and bound* [Narendra and Fukunaga 1977].
- **Sequential search** iteratively updates the current subset selection by adding or deleting features from the active set. The method can be thought of as an instance of greedy hill-climbing, where the optimality is not guaranteed. Typical heuristics include *forward selection* (start with empty set and add features one by one), *backward selection* (start with all the features and eliminate one at a time) and *forward-backward feature selection*, which alternates between addition and selection steps [Hastie et al. 2009].
- **Randomized search** algorithms introduces randomness into the selection through well-known randomized optimization algorithms such as *simulated annealing* [Lin et al. 2008] and *genetic algorithms* [Siedlecki and Sklansky 1989]. The randomness extends the sequential search by adding possibility of escaping from local optima. However, the difficulty in randomized algorithms is indeed their randomness: ten test runs result in ten different results, which makes drawing any conclusions difficult.
- **Embedded feature selection** is an alternative to the wrapper approach that embeds the feature selection into the performance criterion. The feature selection can be added to the performance criterion by introducing a penalty for the number of nonzero coefficients in the model. The most popular of embedded feature selection methods is the *logistic regression* model, where feature selection can be embedded by sparsity promoting prior in the Bayesian framework [Krishnapuram et al. 2005; Friedman et al. 2010].

In the subsequent sections, we will concentrate on embedded feature selection and sequential search.

2.1 Embedded Feature Selection

A recent approach to feature selection embeds the feature selection into classifier design. One of the most successful methods uses the *logistic regression* model, also known as the *logit model*. In addition to designing a classifier, the cost function includes a sparsity enforcing regularization term and, thus, works as an embedded feature selector that automatically selects the set of relevant features and channels from the pool of candidates.

More specifically, the symmetric multinomial logistic regression models the conditional probability of class $k = 1, 2, \ldots, K$ given the P-dimensional feature vector $\mathbf{x} = (x_1, x_2, \ldots, x_P)^T$ as

$$p_k(\mathbf{x}) = \frac{\exp(\beta_{k0} + \boldsymbol{\beta}_k^T \mathbf{x})}{\sum_{j=1}^{K} \exp(\beta_{j0} + \boldsymbol{\beta}_j^T \mathbf{x})}, \tag{1}$$

where β_{k0} and $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{kP})^T$ are the coefficients of the model [Hastie et al. 2009]. For this model to be valid we have to assume mixture or $\mathbf{x}$-conditional sampling [Anderson and Blair 1982] or – in a more relaxed form – that the class frequencies are (approximately) the same in the training and test data. Despite of the apparent nonlinearity of Equation (1), the resulting classifier is linear and the class k^* of a test sample $\mathbf{x}$ is selected as $k^* = \arg\max_k \left\{\beta_{k0} + \boldsymbol{\beta}_k^T \mathbf{x}\right\}$.

The training of the logistic regression model consists of maximizing the following likelihood function:

$$\sum_{i=1}^{N} \log p_{y_i}(\mathbf{x}_i), \tag{2}$$

where $y_i \in \{1, 2, \ldots, K\}$ denotes the true class of the i^{th} training sample $\mathbf{x}_i$ ($i = 1, 2, \ldots, N$).

The feature selection is introduced by adding a sparsity promoting penalty term into Equation (2). A widely used penalty term is the ℓ_1 cost, which results in the problem of maximizing the equation:

$$\sum_{i=1}^{N} \log p_{y_i}(\mathbf{x}_i) - \lambda \sum_{k=1}^{K} ||\boldsymbol{\beta}_k||_1, \tag{3}$$

where $||\boldsymbol{\beta}_k||_1$ denotes the ℓ_1 norm of the coefficient vector $\boldsymbol{\beta}_k$. The extent of regularization is controlled by the regularization parameter $\lambda \geq 0$.

The traditional regularization approach uses ℓ_2 penalty instead of ℓ_1 penalty. However, the ℓ_2 does not result in sparse solution, and its use has been limited to improved generalization due to shrinkage. Nevertheless, the ℓ_2 penalty possesses many favourable properties that ℓ_1 does not, and in many cases results in better generalization than ℓ_1. Thus, a combination of the two penalty term can result in a sparse classifier with good generalization. Such a combination is known as the *elastic net* penalty and the penalized log-likelihood function to be maximized is defined as [Zou and Hastie 2005]

$$\sum_{i=1}^{N} \log p_{y_i}(\mathbf{x}_i) - \lambda \sum_{k=1}^{K} \left(\alpha ||\boldsymbol{\beta}_k||_1 + (1-\alpha)||\boldsymbol{\beta}_k||_2^2\right), \tag{4}$$

The regularization term is a combination of the ℓ_1 and ℓ_2 norms of the coefficient vectors $\boldsymbol{\beta}_k$, and the weights for both types of norms are determined by the mixing parameter $\alpha \in [0, 1]$.

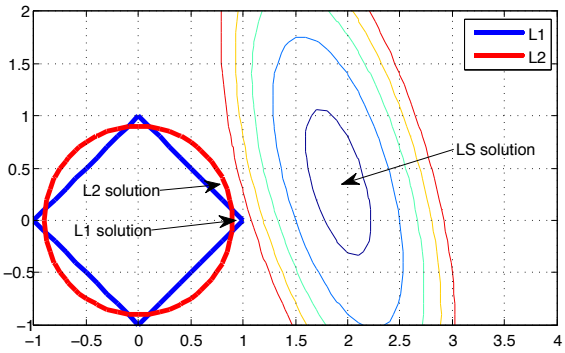


Figure 1: *The isosurfaces of the ℓ_1 and the ℓ_2 penalty. The minimum of the cost function inside the square defining the ℓ_1 penalty is along the horizontal axis thus resulting in a sparse solution.*

The role of parameter α is to determine the type of regularization. When α is zero, the ℓ_1 norm vanishes and results in a special case of *Tikhonov regularization*. The ℓ_2 penalty can be expected to work well in cases, where there are several noisy and mutually correlating features. This is because penalizing the ℓ_2 norm brings the coefficients of the correlating features closer to each other resulting in noise reduction in form of averaging. On the other hand, when α is equal to one, the ℓ_2 norm disappears, which produces the ℓ_1 regularized logistic regression of Equation (3). The ℓ_1 regularization is known for its ability to produce sparse solutions where only a few of the coefficients are non-zero, and this property carries over to the elastic net (except for the case $\alpha = 0$). Thus, both the ℓ_1 regularization and the elastic net can be efficiently used as an implicit feature selectors.

The role of the parameter λ is to control the strength of the regularization effect: the larger the value of λ, the heavier the regularization. For small values of λ, the solution is close to the maximum likelihood solution, while large values of λ allow only restricted solutions and push the coefficients towards zero. In practice, the values of both regularization parameters α and λ are determined by cross-validation, i.e., all combinations over a fixed grid are tested and the CV errors are compared.

The sparsity produced by the ℓ_1 norm is often explained graphically by a figure similar to Figure 1. The figure shows the isosurfaces of the ℓ_1 (square) and the ℓ_2 (round) penalty in terms of parameters β_1 and β_2 in a two-dimensional case when λ has been fixed. The ℓ_1 regularized solution is the point where the quadratic cost represented by the ellipses reaches its minimum inside the square. Due to the cornered shape of the constraint, this tends to appear at one of the corners of the constraint region, and the likelihood increases as the square is scaled down by increasing the value of λ. The same effect is not apparent with the round constraint area of the ℓ_2 penalty.

There are two approaches for estimating the parameters for model (1): Either through maximization of the likelihood function separately for each λ [Yamashita et al. 2008], or simultaneously for the whole regularization path [Friedman et al. 2010]. In the experiments below, we use the latter method due to higher speed.

Algorithm 1 Forward selection.

Initialize the parameter subset as $S = \emptyset$.
Initialize the classification error as $\epsilon(S) = \infty$.
while not terminated **do**
  for For each variable $x_p \notin S, p = 1, 2, \ldots, P$ **do**
    $S' = S \cup \{x_p\}$
    // M-fold CV loop:
    for $j = 1 \rightarrow M$ **do**
      Use the feature set S' and train with all training data except the j^{th} fold.
      Estimate the error ϵ_j by classifying the j^{th} fold.
    end for
    The classifier error estimate ϵ_p is the mean of ϵ_j.
  end for
  Find $\hat{p} = \arg\min_p \epsilon_p$.
  if $\epsilon_{\hat{p}} < \epsilon(S)$ **then**
    Let $S \leftarrow S \cup \{x_{\hat{p}}\}$
  end if
  If an improved subset was not found on this iteration, exit the while loop.
end while

2.2 Sequential Feature Selection Methods

2.2.1 Forward Selection and Backward Selection

Among the sequential feature selection methods, the simplest ones are forward selection and backward selection. Their difference is in the starting point of the iteration: forward selection starts with an empty feature set and iteratively adds new features, while backward selection starts with all features and deletes the most harmful features one by one. These selection algorithms are described in more detail in Algorithm 1 and Algorithm 2. The algorithms can also be combined to alternate between addition and deletion steps, and different rules of thumb can be constructed for balancing the probability of addition and deletion (see e.g., [Zhang 2011] for a recent adaptive acceptance rule). In the experiments of Section 5, we experimented also with a few forward-backward algorithms, and it turned out that the feature selection path was always one-directional, i.e., only either forward or backward steps were taken, depending on the starting point (empty feature set or full feature set).

2.2.2 Simulated Annealing Feature Selection

In order to avoid local minima, *Simulated Annealing* (SA) has also been used for feature selection [Lin et al. 2008]. SA is a randomized search heuristic with roots in condensed matter physics, where slowed-down cooling is used to reduce the defects of the material by allowing the molecule configuration to reach its global minimum state. The method has been successfully used in various optimization problems with multiple local extrema. More specifically, SA starts with the empty feature subset, and at each iteration attempts to add or remove a random feature from the set. The change in cross-validated prediction error then used to determine whether the new subset is accepted. All improved results are accepted, while worse solutions are accepted at random with probability

$$\exp\left(\frac{\epsilon(S) - \epsilon(S')}{T}\right), \tag{5}$$

where T is the simulated temperature and $\epsilon(S)$ and $\epsilon(S')$ are the error estimates for the old (better) and the new (worse) solution, respectively. The temperature T is initialized to a high value where almost all configurations are accepted, and it is decreased at each

Algorithm 2 Backward selection.

Initialize the parameter subset as $S = \{\text{all available features}\}$.
Initialize the classification error as $\epsilon(S) = \infty$.
while not terminated **do**
  for For each variable $x_p \in S$, $p = 1, 2, \ldots, P$ **do**
    $S' = S \setminus \{x_p\}$
    // M-fold CV loop:
    for $j = 1 \rightarrow M$ **do**
      Use the feature set S' and train with all training data except the j^{th} fold.
      Estimate the error ϵ_j by classifying the j^{th} fold.
    end for
    The classifier error estimate ϵ_p is the mean of ϵ_j.
  end for
  Find $\hat{p} = \arg\min_p \epsilon_p$.
  if $\epsilon_{\hat{p}} < \epsilon(S)$ **then**
    Let $S \leftarrow S \cup \{x_{\hat{p}}\}$
  end if
  If an improved subset was not found on this iteration, exit the while loop.
end while

iteration according to the rule $T \leftarrow \alpha T$ with $\alpha < 1$. The method is described in detail in Algorithm 3.

2.2.3 Wrapped Classifier: The SVM

Sequential feature selection methods always have to be coupled with a classifier, whose performance is iteratively tested with candidate feature sets. In the results below, we choose to use a support vector machine (SVM) for this task. The SVM is is a maximum margin classifier, which (in binary case) maximizes the minimum distance of the training samples from the decision boundary.

The success of the SVM is based on the observation that decision functions can be represented through inner products with the training samples as follows [Schölkopf and Smola 2001]:

$$c(\mathbf{x}) = \operatorname{sgn}\left(\sum_{n=1}^{N} \alpha_n y_n \langle \mathbf{x}, \mathbf{x}_n \rangle + \beta_0\right), \tag{6}$$

where $c(\mathbf{x}) = \{-1, +1\}$ is the predicted class label for sample $\mathbf{x}$, and $\mathbf{x}_n$ and y_n ($n = 1, 2, \ldots, N$) are the training samples and classes, respectively. Moreover, α_n, $n = 1, 2, \ldots, N$ and β_0 are the model parameters inferred from the training data. The representation (6) enables the use of the *kernel trick* [Schölkopf and Smola 2001], which implicitly maps the data into a higher dimensional space through a mapping $\phi(\mathbf{x})$. Instead of explicitly mapping the data into higher dimension, it is enough to calculate the inner products $\langle \phi(\mathbf{x}), \phi(\mathbf{x}_n) \rangle$. The kernel trick substitutes this inner product with a kernel function $\kappa(\mathbf{x}, \mathbf{x}_n) = \langle \phi(\mathbf{x}), \phi(\mathbf{x}_n) \rangle$, and it can be shown that all positive definite kernels $\kappa(\cdot, \cdot)$ correspond to a mapping $\phi(\cdot)$. Thus, the kernelized version of (6) is defined as

$$c(\mathbf{x}) = \operatorname{sgn}\left(\sum_{n=1}^{N} \alpha_n y_n \kappa(\mathbf{x}, \mathbf{x}_n) + \beta_0\right), \tag{7}$$

Despite the apparent linearity of the SVM classifier, the implicit mapping in fact allows nonlinear decision boundaries. This makes the SVM very interesting for our comparison: While preparing for the ICANN MEG challenge submission, we were concerned with how significant is the degradation in performance caused by the linearity of the decision boundaries of the logistic regression classifier.

Algorithm 3 Simulated annealing selection.

Initialize the parameter subset as $S = \emptyset$.
Initialize the classification error as $\epsilon(S) = \infty$.
Initialize the temperature T.
while not terminated **do**
  Randomly select $p \in \{1, 2, \ldots, P\}$.
  if $x_p \in S$ **then**
    $S' = S \setminus \{x_p\}$
  else
    $S' = S \cup \{x_p\}$
  end if
  // M-fold CV loop:
  for $j = 1 \rightarrow M$ **do**
    Use the feature set S' and train with all training data except the j^{th} fold.
    Estimate the error ϵ_j by classifying the j^{th} fold.
  end for
  The classifier error estimate $\epsilon(S')$ is the mean of ϵ_j.
  Let $S \leftarrow S'$ with probability $\min\{1, \exp(\epsilon(S) - \epsilon(S')/T)\}$
  $T \leftarrow \alpha T$.
  If an improved subset has not been found for N iterations, exit the while loop.
end while

In the results section we will study its efficiency with two kernels: the linear kernel and the popular radial basis function (RBF) kernel.

A multiclass classification problem can be reduced to multiple binary problems. Multiclass SVM is implemented, e.g., in the `LibSVM` package [Chang and Lin 2011], which we also use in our experiments.

3 Material

In the results section we study the efficiency of two feature selection and classification approaches using the data of the MEG Mind Reading challenge of ICANN 2011 conference[2]. The data set consists of MEG signals recorded from a test subject while watching five different video stimuli without audio:

1. **Artificial:** Animated shapes or text
2. **Nature:** Nature documentary clips
3. **Football:** Soccer match clips
4. **Bean:** Part from the comedy series "Mr. Bean"
5. **Chaplin:** Part from a Chaplin movie

The provided measurements consist of 204 gradiometer channels, and the length of each individual epoch is one second and the sampling rate is 200 Hz. Moreover, the five band-pass filtered versions of the signal are also included in the measurement data, with bands centered on the frequencies of 2 Hz, 5 Hz, 10 Hz, 20 Hz, and 35 Hz.[3]

The MEG measurements were recorded on two separate days such that the same set of video stimuli was shown to a test person on both days. Stimuli labeled as either Artificial, Nature, or Football (short clips) were presented as randomly ordered sequences of length 6 –

[2]The data can be downloaded from `http://www.cis.hut.fi/icann2011/meg/measurements.html`

[3]Note, that the challenge report [Klami et al. 2011] erroneously states the frequency features to be *the envelopes* of the frequency bands. However, the data consists of the plain frequency bands; see the erratum at `http://www.cis.hut.fi/icann2011/meg/megicann_erratum.pdf`.

26 s with a 5 s rest period between the clips, while Bean and Chaplin (movies) were presented in two consecutive clips of approximately 10 minutes. In the competition data, the measurements are cut into one-second epochs that are further divided into training and testing such that the training data with known class labels contains 677 epochs of first day data and 50 epochs of second day data while the secret test data contains 653 epochs of second day data only. Notice that the ground truth class labels for the test recordings have been released after the end of the competition.

The data is provided in a randomized order, and the complete signal cannot be reconstructed based on the individual signal epochs. During the competition, the competitors were given the information that the secret test data comes from the second day measurements only and that – similar to the training data – it is approximately class-balanced.

The division between training and test data was elaborate. In particular, 33 % of the test samples consist of recording during stimuli not seen in the training phase in order to test the ability of the classifiers to generalize to new stimuli. A more detailed description of the data can be found in the challenge report by Klami et al. [Klami et al. 2011].

For each one-second epoch of the data in 204 gradiometer channeles, we apply a feature extraction step. For the competition we experimented with numerous features fed to the classifier, and attempted to design discriminative features using various techniques [Huttunen et al. 2011]. Our understanding is that those ended up being too specific for the the first day data and eventually a simplistic solution turned out to be the best, resulting in the following features:

- The detrended mean for each channel, i.e., the parameter $\hat{b}$ of the linear model $y = ax + b$ fitted to the time series.
- The standard deviation of the residual of the fit for each channel, i.e., stddev$(\hat{y} - y)$.

Both features are calculated from the raw data; we were unable to gain any improvement from the filtered channels. Since there are 204 channel, this makes a total of 408 features from which to select in the feature selection step.

4 Performance Assessment

An important aspect for the classifier design is the error assessment. However, the designer would like to spend only a small portion of the annotated training data for testing. The typical approach is to use some kind of cross-validation technique, where a subset of training data is used for assessing the performance of the data. However, there are several pitfalls that the designer should avoid.

There are two kinds of errors that challenge the performance of a classifier: *training error* and *generalization error*. The former refers to the classification error caused by insufficient separation between the classes: Overlapping classes can not be separated no matter how good the classifier is. The training error can be efficiently estimated from the training data, and its significance can thus be easily assessed. The latter kind of error is typically more tricky: Generalization error is the error ultimately caused by training on a finite sample and testing on a finite sample. Because of this, there is always random variation present in both samples, and the classifier will learn to exploit patterns present in the training set but not in the test set. In most cases, the generalization error is the main cause of poor performance, and we believe that our assessment of generalization error was the key to our success in the two recent challenges [Huttunen et al. 2011; Manninen et al. 2011].

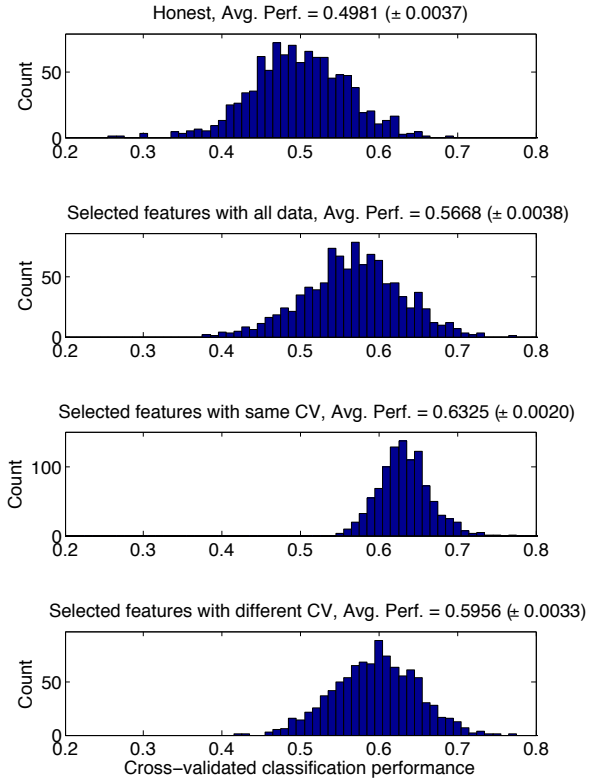


Figure 2: *Results of the wrong and right way to do cross-validation: (a) accuracy over 1000 tests without feature selection; (b) feature selection with all data; (c) feature selection using 10-fold CV twice; (d) feature selection using 10-fold CV twice with different seeds.*

In a data-rich situation the preferred solution is to split the data into three parts: *training set*, *validation set* and *test set* [Hastie et al. 2009, p. 222]. The training set is used for fitting the model, the validation set for estimating the prediction error when developing the model (e.g., when testing which features to include), and the test set for assessing the generalization error. An important point is to keep the last set "in vault" until the very last moment, and should only be introduced into the process as late as possible to avoid overlearning to it. Every time the test set is used for making decisions about the classifier, the estimated generalization error and, thus, the true performance, become more and more optimistic. The following section describes an experiment, where this principle is violated.

4.1 Wrong Ways to Use Cross Validation in Feature Selection

Hastie *et al.* describe *the wrong and right way to do cross-validation* in their book [Hastie et al. 2009], and the wrong way appears frequently in various domains [Ambroise and McLachlan 2002]. The wrong way frequently appears in scientific literature, and uses *all data* for making decisions about the model. In particular, Hastie *et al.* concentrate on feature selection: The incorrect strategy consists of three steps [Hastie et al. 2009, Sec. 7.10.2]:

1. Find a subset of good predictors that exhibit strong correlation with the class labels

2. Using only this subset, design a multivariate classifier.
3. Use cross-validation to tune the model and estimate the prediction error.

The problem of the above approach is that it uses all data in the first step and, thus, produces a too optimistic error estimate and a worse performance for truly independent test data. However, it turns out that the misuse can be a lot more subtle and difficult to recognize as we will see.

Following the example of [Hastie et al. 2009], we did the following experiment.

1. Generate $N = 100$ samples of random data of dimension $P = 30$.
2. Generate binary class labels for the samples also at random.
3. **Honest way:** Design a classifier and calculate the classification error using 10-fold cross validation. In this example we used a 3-nearest-neighbor (3NN) classifier.
4. **Cheating (version 1):** Violate the CV principles as in [Hastie et al. 2009]:
   (a) Find the single feature among $P = 30$ that gives least classifier error with 3NN.
   (b) Estimate the error as in step 3, but with only the best feature.
5. **Cheating (version 2):** Use the 10-fold CV *twice*:
   (a) Estimate classification error for each feature using the 10-fold CV.
   (b) Estimate the error using 10-fold CV as in step 3, but with only the best feature found in step 5a.
6. **Cheating (version 3):** Proceed as in step 5, but change the random seed between the two CV's, thus using different division of samples.

The results of this experiments are summarized in Figure 2. It can be seen that the honest way without any feature selection estimates the accuracy in a realistic manner; close to 0.5. However, all three ways of cheating give an optimistic estimate of the performance clearly above 0.5. One might think that using cross-validation in the feature selection step would be less harmful than using all data because the best performing feature in each CV fold is selected by testing on an independent set of data. However, it turns out that the double-CV approaches are clearly the most dangerous ways of incorrectly assessing the classification performance. There are at least two reasons for this to happen: 1) The illusion about the independence of the CV folds is broken right after the first feature selection CV, when the CV results of the best performing features are combined by averaging. 2) Both the first and the second CV use 10 folds, which allows the first feature selection CV to exploit the information about the exact sample size that is going to be used in the second CV that estimates the final classification performance.

4.2 Error Assessment in Our ICANN MEG Submission

In our submission to the ICANN MEG challenge, we used a tailored version of cross-validation, which emphasizes the performance of the second day test data. More specifically, the training and error estimation procedures consist of two nested cross-validation loops as illustrated in Algorithm 4. The outer loop is used for estimating the performance for the unlabeled test data, while the inner loop is used for selection of classifier parameters α and λ (see section 2.1). The high computational complexity of simultaneous error estimation and parameter selection can be clearly seen from the pseudo code. In order to speed up the development, our method uses parallel validation spread over numerous processors as also described in section 4. A Matlab implementation of our method can be downloaded at `http://www.cs.tut.fi/~hehu/mindreading.html`.

Algorithm 4 Error estimation and parameter selection using nested cross validation.

```
// Outer CV loop:
for n = 1 → N do
    Divide the training data into training and validation sets as
    described in section 4
    // Search over all parameter combinations:
    for α = α_min → α_max do
        for λ = λ_min → λ_max do
            // Inner M-fold CV loop:
            for j = 1 → M do
                Train with all training data except the j^th fold.
                Estimate the error e_j by classifying the j^th fold.
            end for
            The error estimate e_{α,λ} is the mean of e_j.
        end for
    end for
    Classify the test data using the classifier with smallest e_{α,λ}.
    Denote the test error by e_n.
end for
The final error estimate is the mean of all e_n.
```

A natural cross-validation (CV) error estimation technique would be the leave-one-out error estimator for the second day data. More specifically, we would train with all the first day data and 49 samples of the second day data and test with the remaining second day sample. This way there would be 50 test cases whose mean would be the leave-one-out error estimate. However, we were concerned about the small number of test cases, and decided to consider alternative divisions of the second day data to training and testing.

Instead, we randomly divided the 50 test day samples into two parts of 25 samples. The first set of 25 samples was used for training, and the other for performance assessment. Since the division can be done in $\binom{50}{25} > 10^{14}$ ways, we have more than enough test cases for estimating the error distribution. This approach gives slightly too pessimistic error estimates because only half of the second day data is used for training (as opposed to 98 % with leave-one-out), but has smaller variance due to larger number of test cases. Moreover, the pessimistic bias is not a problem, because we are primarily interested in comparing feature sets during method development rather than actually assessing the prediction error.

The imbalance in the number of samples between the first and second day data is quite significant, because with the above division the training set contains more than 25 times more first day data than second day data. Since we wanted to emphasize the role of the second day data, we increased its relative weight in training error. After experimentation, the second day weight was set to three. In training a logistic regression classifier, the weighting can be implemented in a straightforward manner by multiplying each log-odd in Equation 2 by the corresponding weight value.

The remaining problem in estimating the error distribution is the computational load. One run of training the classifier with CV of the parameters takes typically 10 – 30 minutes. If, for example, we want to test with 100 test set splits, we would be finished after a day or two. For method development and for testing different features this is certainly too slow. However, the error estimation can be easily parallelized; simply by testing each division of the test data on a different processor. For example, in our case we had access to a grid computing environment with approximately 1000

Table 1: *A comparison of the performance of different feature selection and classification methods on the ICANN2011 challenge test data. The results for the SAFS selection are averages over 20 test runs due to the stochastic nature of simulated annealing.*

Feat. sel.	Classifier	# feat.	Perf.	p-value
none	*SVM (Lin.)*	408	66.16 %	0.316
none	*SVM (RBF)*	408	67.99 %	0.765
SFS	*SVM (Lin.)*	15	55.44 %	7.01e-7
SFS	*SVM (RBF)*	18	56.66 %	6.14e-6
SBS	*SVM (Lin.)*	334	68.30 %	0.858
SBS	*SVM (RBF)*	334	67.84 %	0.721
SAFS	*SVM (Lin.)*	193.9	65.79 %	0.253
SAFS	*SVM (RBF)*	191.8	65.48 %	0.207
El. Net.	*Log. Regr.*	219	**68.76 %**	-

processors, and we were able to obtain an accurate error estimate in a matter of minutes instead of hours or days.

5 Experimental Results

In this section, we consider the multiclass classification problem of the ICANN2011 MEG mind reading challenge. The training and testing data and the feature extraction step are described in Section 3. We will compare the elastic net penalized logistic regression model (LR-ELNET) that we used in winning the challenge with different combinations of a feature selection algorithm and the SVM. For feature selection, we will consider Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), and Simulated Annealing Selection (SAFS). For classification, we will consider SVMs with linear and radial basis function (RBF) kernels. We used the `libsvm` implementation of the SVM classifier [Chang and Lin 2011], which uses a one-against-one strategy for multinomial classification. The SVM kernel width was chosen as $\gamma = \frac{1}{P}$ and the penalty parameter for the error term was chosen as $C = 1$. We do acknowledge that a cross-validated selection of these parameters would probably improve the performance, but we believe the results are close to optimal.

The test results are shown in Table 1. The first column shows the method used in feature selection and the second column shows the method used in classification. The third column shows how many features of the total 408 ended up in the selected model. The fourth column shows the percentage of correct classifications for the test data. As there are 5 different classes roughly balanced in the test set, the level of random guess is 20 %.

The method used in our original submission seems to be the best performer also in this comparison. However, there are a few alternatives that are within a small margin. The rightmost column shows the p-value of observing the respective performances under the null hypothesis that the true performance is equal to that of regularized logistic regression given the training and test sets[Dietterich 1998]. Thus, only the cases with SFS selection exhibit a poorer performance with statistical significance. However, with the competition test data the winning method remains the same: logistic regression with elastic net regularization.

In particular, both variants of the SVM seem to be successful with backward selection. The forward selection is not as successful, because it gets trapped in a local minimum and includes far too few features in the design. During the experiments, it turned out that alternating between forward and backward steps does not help with local minima: the search path is in practice always one-directional, and the optimizer takes only either addition or elimination steps.

It seems, that increasing the number of features tends to improve the performance until a saturation point somewhere near approximately 300 features. This conclusion is also supported by the fact that among the 20 test runs of the SAFS, there is a high correlation between the number of features and the prediction accuracy. An interesting coincidence is that when using SBS, both SVM variants end up with the same amount of features. The feature sets are not the same, however.

The results also reveal, that a linear classifier has enough discriminative power for the particular application: The two SVM variants are equal in performance. This is probably due to the high dimensionality of the problem, which helps in finding well separating hyperplanes even without the kernel trick.

It is slightly surprising that the simulated annealing does not perform very well. This may be because the optimization starts with an empty feature set, similarly to the forward selection. Thus, the distance to the optimum region with approx. 300 features is very long, and the process cools down too early. One solution to improve the performance could be to start with the full set of features instead of the empty set. However, this would increase the computation time of the already slow annealing even further, because the SVM design time depends on the number of features.

6 Conclusions

In this paper, we have compared the accuracy of the elastic net-regularized logistic regression classifier with the support vector machine classifier combined with various wrapper-based feature selection methods. The data for the comparison was drawn from a recent MEG mind reading competition, where regularized logistic regression outperformed other competitors. The dataset in question has nearly equal number of training samples and features when the number of features is first reduced using an ad-hoc technique described in Section 3. While the logistic regression algorithm was the most accurate in our experiments, the wrapper-SVMs combined with SBS and SAFS and SVM without any feature selection gave nearly as good results as the regularized logistic regression. In other words, the differences between the accuracies of these algorithms were non-significant. Instead, sequential forward feature selection selected always too few features and was significantly less accurate than the regularized logistic regression. We experimented with two different SVM kernels, but found virtually no difference in accuracies of classifiers produced with them. An important aspect in using wrapper methods for feature selection is the performance evaluation, usually using a cross-validation technique. As we have outlined in detail in Section 4, it deceptively easy to misuse the cross-validation and eventually produce too optimistic estimates of the actual error rates.

7 Acknowledgements

Supported by the Academy of Finland grant no. 130275.

References

AMBROISE, C., AND MCLACHLAN, G. J. 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences 99*, 10, 6562–6566.

ANDERSON, J., AND BLAIR, V. 1982. Penalized maximum likelihood estimation in logistic regression and discrimination. *Biometrika 69*, 123–136.

BLANKERTZ, B., TANGERMANN, M., VIDAURRE, C., FAZLI, S., SANNELLI, C., HAUFE, S., MAEDER, C., RAMSEY, L., STURM, I., CURIO, G., AND MLLER, K.-R. 2010. The Berlin Brain-Computer Interface: Non-Medical Uses of BCI Technology. *Front Neurosci 4*, 198.

CHANG, C.-C., AND LIN, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2*, 27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

DEBUSE, J. C., AND RAYWARD-SMITH, V. J. 1997. Feature subset selection within a simulated annealing data mining algorithm. *Journal of Intelligent Information Systems 9*, 57–81.

DIETTERICH, T. G. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput. 10*, 7 (oct), 1895–1923.

FRIEDMAN, J. H., HASTIE, T., AND TIBSHIRANI, R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*, 1, 1–22.

HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. 2009. *The elements of statistical learning: Data mining, inference, and prediction*, Second ed. Springer Series in Statistics. Springer.

HILBERT, M., AND LÓPEZ, P. 2011. The World's Technological Capacity to Store, Communicate, and Compute Information. *Science 332*, 6025 (Apr), 60–65.

HUTTUNEN, H., KAUPPI, J.-P., AND TOHKA, J. 2011. Regularized logistic regression for mind reading with parallel validation. In *ICANN2011 MEG challenge*.

HUTTUNEN, H., MANNINEN, T., KAUPPI, J.-P., AND TOHKA, J. 2012. Mind reading with regularized multinomial logistic regression. *Machine Vision and Applications* (Jan.). Submitted.

KAUPPI, J.-P., HUTTUNEN, H., KORKALA, H., JÄÄSKELÄINEN, I. P., SAMS, M., AND TOHKA, J. 2011. Face prediction from fmri data during movie stimulus: Strategies for feature selection. In *ICANN (2)*, Springer, T. Honkela, W. Duch, M. A. Girolami, and S. Kaski, Eds., vol. 6792 of *Lecture Notes in Computer Science*, 189–196.

KLAMI, A., RAMKUMAR, P., VIRTANEN, S., PARKKONEN, L., HARI, R., AND KASKI, S., 2011. ICANN/PASCAL2 Challenge: MEG Mind-Reading — Overview and Results.

KRISHNAPURAM, B., CARIN, L., FIGUEIREDO, M., AND HARTEMINK, A. 2005. Sparse multinomial logistic regression: fast algorithms and generalization bounds. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 27*, 6 (june), 957 –968.

LIN, S.-W., LEE, Z.-J., CHEN, S.-C., AND TSENG, T.-Y. 2008. Parameter determination of support vector machine and feature selection using simulated annealing approach. *Appl. Soft Comput. 8*, 1505–1512.

LIU, H., AND YU, L. 2005. Toward integrating feature selection algorithms for classification and clustering. *Knowledge and Data Engineering, IEEE Transactions on 17*, 4 (april), 491 – 502.

MANNINEN, T., HUTTUNEN, H., RUUSUVUORI, P., AND NYKTER, M. 2011. Logistic regression for AML prediction. In *Dialogue for Reverse Engineering Assessments and Methods, DREAM6*.

NARENDRA, P., AND FUKUNAGA, K. 1977. A branch and bound algorithm for feature subset selection. *Computers, IEEE Transactions on C-26*, 9 (sept.), 917 –922.

PEREIRA, F., MITCHELL, T., AND BOTVINICK, M. 2009. Machine learning classifiers and fmri: a tutorial overview. *NeuroImage 45*, Suppl 1, S199 – S209.

POLDRACK, R., HALCHENKO, Y., AND HANSON, S. 2009. Decoding the large-scale structure of brain function by classifying mental states across individuals. *Psychological Science 20*, 1364 – 1372.

PUDIL, P., NOVOVIČOVÁ, J., AND KITTLER, J. 1994. Floating search methods in feature selection. *Pattern Recogn. Lett. 15*, 11 (nov), 1119–1125.

SAEYS, Y., INZA, I., AND LARRAAGA, P. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics 23*, 2507 – 17.

SCHÖLKOPF, B., AND SMOLA, A. J. 2001. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, 1st ed. The MIT Press.

SIEDLECKI, W., AND SKLANSKY, J. 1989. A note on genetic algorithms for large-scale feature selection. *Pattern Recogn. Lett. 10*, 5 (nov), 335–347.

SMITH, S., JENKINSON, M., WOOLRICH, M., BECKMANN, C., BEHRENS, T., JOHANSEN-BERG, H., BANNISTER, P., LUCA, M. D., DROBNJAK, I., FLITNEY, D., NIAZY, R., SAUNDERS, J., VICKERS, J., ZHANG, Y., STEFANO, N. D., BRADY, J., AND MATTHEWS, P. 2004. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage 23*, S1, 208 – 219.

WOLZ, R., JULKUNEN, V., KOIKKALAINEN, J., NISKANEN, E., ZHANG, D. P., RUECKERT, D., SOININEN, H., LOTJONEN, J., AND THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE. 2011. Multi-method analysis of mri images in early diagnostics of alzheimer's disease. *PLoS ONE 6*, 10 (10), e25446.

YAMASHITA, O., SATO, M., YOSHIOKA, T., TONG, F., AND KAMITANI, Y. 2008. Sparse estimation automatically selects voxels relevant for the decoding of fmri activity patterns. *NeuroImage 42*, 4, 1414 – 1429.

ZHANG, T. 2011. Adaptive forward-backward greedy algorithm for learning sparse representations. *Information Theory, IEEE Transactions on 57*, 7 (July), 4689 –4708.

ZOU, H., AND HASTIE, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*, 2, 301–320.

Conceptual design and inductive learning of industrial processes - metallurgical processes as a case

Martti Meri *
Aalto University
Department of Materials Science,
P.O. Box 16200,
00076 Aalto University

Abstract

There are two technological breakthroughs in computer science that facilitate development of new decision support systems. First of all Internet search techniques allow fast access to data, including case data for various domains. Secondly automated reasoning systems, for both model generation and theorem proving, have become fast and scale up well to huge problems. Case based planning (CBP) generalizes from case instances accessible over the net, producing a schemata that defines actions. This schemata and local conditions together with the case data can be used for deducing, by running the reasoners, the what can be done locally. This paper outlines a prototype of a system using hydrometallurgy as a problem domain.

CR Categories: I.2.3 [Artificial Intelligence]: Deduction and theorem proving—Resolution I.2.6 [Artificial Intelligence]: Learning—Induction I.2.8 [Artificial Intelligence]: Problem solving, Control methods and Search—Plan Execution, formation and generation;

Keywords: case based reasoning, planning, industrial process design, automated reasoning

1 Introduction

How to construct the best possible process for an engineering domain, based on the knowledge of what kind of compatible processes have been used in corresponding situations elsewhere and in other times? This paper presents a sketch of a method that has a work in progress status. Conceptual design of industrial processes is a preliminary state of the design process where the aim is to keep the calculations at a symbolic level. Hydrometallurgical processes in refractory gold extraction are the case domain and the project is a joint effort of domain research and artificial intelligence research. The method comprises deductive reasoning methods and inductive reasoning methods allowing planning and learning in the same framework following a Case Based Planning (CBP) approach.

Industry, commerce and public services rely on well designed processes and work-flows. Organizations of the Internet-era are usually network organizations. They are dynamically organized around projects where firms of sufficient knowledge capital and compatible interfaces organize themselves to form an efficient, operational whole. In this paper we look at the level of co-operation where knowledge is shared over the network in a way that provides possibilities to reuse experiences of others and makes it possible to construct one's own model of behavior based on these experiences and adjusting the behavior to the local conditions.

Conceptual process design is a preliminary stage of industrial process design, where model generation is the most characteristic mode of reasoning. However the design uses case information, domain theories and requires problem solving in modes that combines several types of logical reasoning in one bundle. Domain theories and availability of case material makes this a potential application area for developing methods where generative and variant - approaches are used, meaning that models can be build from scratch or be based on stored partial solutions that are reused. The idea we present in this paper is an amalgam of reasoning modes and knowledge base structures where it is straightforward to automatically learn by inverse resolution using the case history states, and reason in forward direction using standard resolution to form a proof that acts as the frame for the conceptual process definition.

Mineralogical industry uses processes that obey laws of physical, mineralogical, and chemical domain theories. The idea is to find a sequence of unit actions that when given an ore as input produces the best possible yield of some precious metal (or something similar) and a process tail that is as harmless as possible. The fact that several kinds of theories are mixed is one of the reasons conceptual process design, a phase of the process design, is a complicated engineering problem, mainly done based on expert intuition and practical experience. Decision support systems are required in the field , and case based reasoning (CBR) -methods are one of the promising technologies to be applied [Rintala et al. 2011b; Rintala et al. 2011a]. Besides theoretical heterogeneity, the sub-domains are such that they combine sub-symbolic, symbolic and ontological levels of representation. The levels range from numeric data to highly complex description logic statements. Theoretical heterogeneity and varied levels of representation are not yet all of it, there is also an inherent need to represent change and the behavior that causes this change. Chemicals react with each other forming new chemical compounds and even some physically observable phenomena, minerals change physical attributes, making then subjective to actions that they where previously shielded from by some of their say structural characteristics. So, this is not a toy example, but rather a good workbench for testing knowledge engineering methods; trying out modular theories and testing control of mixed mode reasoning algorithms, even in parallel execution mode.

Diagnostic approach for process design for gold ores, [Torres et al. 1999] uses expert system to infer in sequence models that describe the layers:

- geological,
- mineralogical,
- behavioral,
- process options containing and item process work-flow defining

Contradictions are fed back to the previous base of inference. The decision making process is complemented with experiments taking measurements and sub-phases where the inferred knowledge and measurements are fused to form a more complete picture of the process and it's environment. The method is not case based as it

*e-mail:martti.meri@aalto.fi

assumes modeling of most common alternatives into the fixed decision tree like structure.

Conceptual process design is iterative by nature, it requires learning, finding new alternatives, composing candidate models, testing them and retaining the experiences gained by testing in case base for later use. When case data is not available, organizations have the option of conducting their own experiments, which essentially produces the missing cases. Besides public and own internal knowledge bases there are case data made available according to some restrictions from research institutes and commercial sources. The result of conceptual design is a behavioral model that is like a plan in the sense that it takes the process through a series of states to a final state that meets the goal conditions. In our case the process is that of extracting precious metals from an ore. The goal conditions can be stated by writing logical clauses saying that there is metal recovered from the ore in a certain place in sufficient quantity and form. The attributes of the original ore (initial state) and the goal conditions present a planning problem. The solution is a set of ordered operations that make the required state transformation. Literature and experience is full of cases of such problem solution pairs, often so called tacit knowledge, which we here need to make available, explicit and machine readable.

There are two alternatives in planning; re-planning (generative) and plan-repair (variant) approaches [Fox et al. 2006]. These can be used intermixed. General properties and integration of different learning techniques that suit this scenario has been studied actively in the past [Quinlan and Cameron-Jones 1993; Muggleton and de Raedt 1994; Hume and Sammut 1991]. Earlier framework for integrating case based reasoning and inductive learning is presented in [Auriol et al. 1994]. The classical phases of CBR -Cycle are *retrieve*, *reuse*, *revise* and *retain*. Since we are dealing, hopefully , with large knowledge bases that are distributed, contain partial information and present wide options that transform to huge search spaces we need to have good design of metrics for similarity, and different equality, congruence and logical entailment relations to build our knowledge retrieval and solution search mechanisms on. We need to construct the solution why retrieving new alternative model constructs on the fly. As a result a picture of the CBP-Cycle emerges where at the core operates a pair of alternating intervened logical reasoning phases that operate in resolution and inverse resolution modes. The CBP-framework directs the research towards developing specialized metrics for the states, goals and plans separately, so that we can say that we are acting in similar situation, with conforming ways to reach goals that are in some sense compatible in a logical sense.

For both the fast case retrieval and fast reasoning knowledge indexing becomes a key factor in CBP. The indexing in [Auriol et al. 1994] was based on binary trees called $k-d$ -trees. In the sequel we need to investigate how to search data elements that have subsymbolic (numeric etc.), symbolic and restricted symbolic (ontological, description logic) contents from an unified knowledge base.

There is a CBR-system named PATDEX using similarity based measure communicates at varied integration levels with a induction techniques utilizing system, named KATE-Induction, which uses information gain measure. The application domain is that of diagnosis. The indexing in PATDEX was based on binary trees called $k-d$ -trees [Wess et al. 1994]. In the sequel we need to investigate how to search data elements that have sub-symbolic (numeric etc.), symbolic and restricted symbolic (ontological, description logic) contents from an unified knowledge base.

In this paper we outline a new method based on logical theorem proving that uses case based knowledge of action-schemata that has appeared in past solutions and modifies them to constitute to the proof of the current plan to be made. This new method integrating automated reasoning with CBR manages in parallel the current proof that essentially corresponds to the plan that is the solution to the planning problem given, and the case base solutions (proofs) of case problems.

2 Knowledge - the case of hydro-metallurgical processes

The process-chain consists of three phases [Hayes and Gray 1985], where the first phase contains ore pretreatment followed by leaching and finished by recovery. Again according to this model [Hayes and Gray 1985], pretreatment falls in three categories:

- comminution and beneficiation
- chemical changes in the minerals
- structure modification.

From knowledge management perspective these processes are interesting as they change the structure and characteristics of the material in ways that are difficult to describe. It is not just that comminution breaks or grinds up the a material to form smaller particles, the new particles may have lost some of the properties of the old and gained some new.

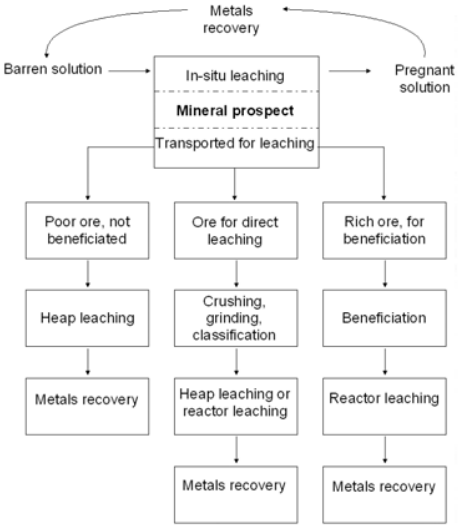


Figure 1: *Hydro-metallurgical leaching techniques*

This shows that the theoretical basis is rich and contains very different aspects. Good place for studying knowledge representation of various material substances, different collections of objects and other difficult cases presented in [Russell and Norvig 2003].

The overall process is according to [Wadsworth 1987] as shown in figure 1. The mineralogy part is beyond the scope of this paper, but hopefully this short presentation gives a flavor of the kind of things that are involved.

The type of integration requires well designed formal representation and for that purpose we use a kind of STRIPS-based action representation formalism. The formalism is such that action schema

names an operation and list the preconditions (set of positive literals) and effects (as set of positive literals and a set of negative literals), in a standard way done in AI [Russell and Norvig 2003]. For instance the fact that a reagent is in the reactor can be stated as $in(reagen_X, reactor_Y)$ .The aim at a logical theorem proving leads us to define (in SATPLAN type of thinking) logical axioms for operators implying both their effects and preconditions [Kautz and Selman 1992]. We assume that it is easier to gain access to the state fluents fluctuations along the history lines recorded by the cases. There states form a sequence and each state has some state fluents, i.e. positive or negative literals holding. Our method provides a novel view to see planning as theorem proving based on the observations of the inductive , linear nature of reasoning in environments where there are available as background information series of process case history data.

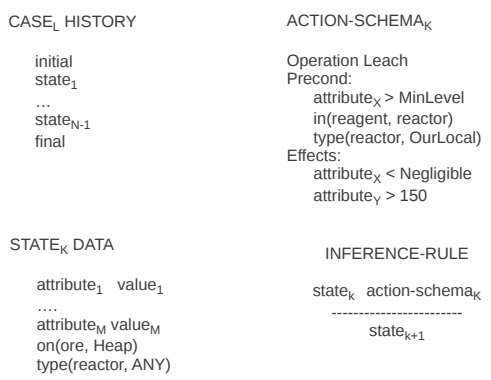


Figure 2: *Main knowledge elements*

The case histories are arrays of arrays of attribute-value pairs, that is each $state_i$ is described as an array of attribute -value pairs, and states arrays form the history, see figure 2. Figure 2 represent the kinds of knowledge that are available in an industrial process conceptual design.

As we want to describe actions in the classical planning format as operators having preconditions and effects, we map these preconditions to the case history and evaluate if the actions are executable in different cases and form chains of actions that can produce effects that are reflected in the later states on the case histories. The attribute value representation of states allows presenting the subsymbolic level of the world. We can add symbolic level by allowing, besides the object attributes also relations between objects to be represented by adding predicates to the language.

A case base of 200 cases reported in research literature will be used in the project. Each case instance records some 100 attribute value pairs describing a part of a history line for the case. The data is partial because of the set of unit processes in one case is typically small, but also because some things are unobservable and need to be completed by reasoning efforts, and some features might be even left out intentionally by the people reporting the case, for some reason. The features in cases wall in all of the categories in terms of theory type, level of symbolism and phase in process flow. Feature based similarity metrics and other approaches to case data handling are to be reported in another place at a later time. The availability of cases is such that we expect that in conceptual process design the case based approach needs back up from model based reasoning component and from human intervention. The time lines of process design projects allow such interactions as well as gradual refinement of both the models and the queries to be launched to external knowledge sources. After all there are aims for supporting creative designs, avoiding costly mistakes and minimizing unnecessary experiments by directing the experiments so that they produce new cases that support the design process.

Once the conceptual process has been formed, it is possible to complete it using model generation methods that add the resources according to the domain ontologies and check the final model against various spatial and temporal molecular sub-theories. Much it the flavor of the relationship AI planning has to scheduling. In the ontological part similarity functions can be also used [Maedche and Staab 2002]. The promising thing is to retrieve work-flows that are similar (using special similarity measures) than the candidate solution we are constructing [Bergmann and Gil 2011].

3 Reasoning

Having already presented the overall structure of the process used as a case study and after outlining the principles for representing the behavior by action schemata, we can now return to the methodological generative/variant dichotomy mentioned earlier, concerning the polarized question of merits of building process descriptions from the scratch vs. reusing proven partial solutions. In generative approach we just represent the initial state as a state where we have the ore as raw material to start with, set goal conditions as a set of constraints for the outcome of the process in terms of the yield, for example grams of gold for a ton of ore. Along the variant policy line we can construct partial plans that are promising, based on the domain theories, cumulative experience and recorded cases. In classical planning literature hierarchical task networks (HTN) is this kind of variant approach. The selection of general satisfiability based techniques diminishes this methodological difference, as it is up to the solver to select the policy to try out all action candidates whether atomic or composite, all types alike. Often the long chains and larger networks of actions can lead to faster finding of solutions, making these partial generally well working composites very valuable, as discussed in [Russell and Norvig 2003]. Selection of the logic based methods in process design has also the effect that the architecture of the decision support system is not something that is worth a diagrammatic presentation here. One just uses suitable solver for appropriate sub-theories and establish a communication link between them. The control and communication between the modules is a more complicated issue we wish to include in the further topics for later articles.

The emphasis received by knowledge representation in declarative manner, the role of inference mechanisms becomes essential. The inference rule in figure 2 (see bottom right corner of the figure) is coarse. We will refine the picture later on (see figure 5, where resolution rule is used, but merely as an example. The calculus depends of the language or languages used. What these inference rules should capture are the logical entailment relations between the behavioral (actions) and state sides of the environment the theories encode. Something like figure 3 should emerge.

The inference rule (or syllogism) in the figure 2 can used in two ways: resolution and inverse resolution. We can deduce new states knowing current state and the action-schema, but we can also inductively form the action schema by looking at two consecutive states. When we have a set of cases that have been applied in a similar situation successfully and the action information for the first operation is similar we can collect the information about the preceding and following states and form our model of the operation based on this collected information.

Different knowledge is given different priority in figure 5, since

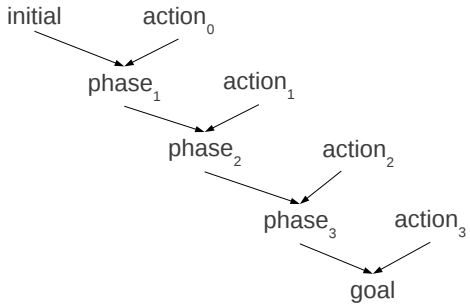


Figure 3: *Process as a proof*

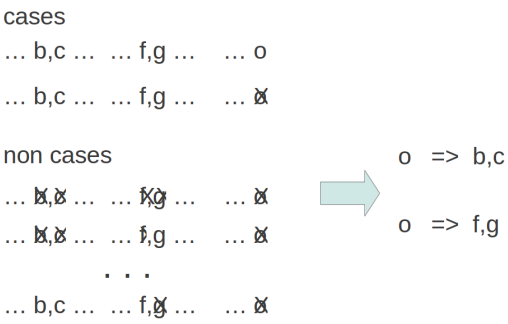


Figure 4: *Generalization of action schemata from cases*

current state is prioritized over goal state the progress is forward directed. To keep the presentation simple we use propositional logic here. The general approach is refutational, one states that there is a model where both initial state and final state conditions hold. Contrary statement to the fact that both the initial state and final state hold is to state that one condition does not hold. Special treatment is needed for the things that don't matter like the propositional variable I in figure 5, since there is no harm in them, only them existing in the clause means from the resolution point of view emergence of an empty clause and thus finishing of the proof.

Resolution methods requires use of a special factoring rule, that essentially removes the multiple entries of literals in the clausules. Were we have wanted to keep the several occurrences of O in the clausule as we have the idea to use them as an indicator that the action O is becoming a useful way to close the resolution proof. Another special feature is the step where we close the proof by combining the action (here O with two other branches of the proof tree, one for the past history and one for the future, i.e. for the branches that collect the implicating clausules corresponding to the preconditions and effects of the action O. Ultimately the idea is to translate these concepts to heuristic rules that control the resolution and allow use of main stream reasoners, resolution based or other.

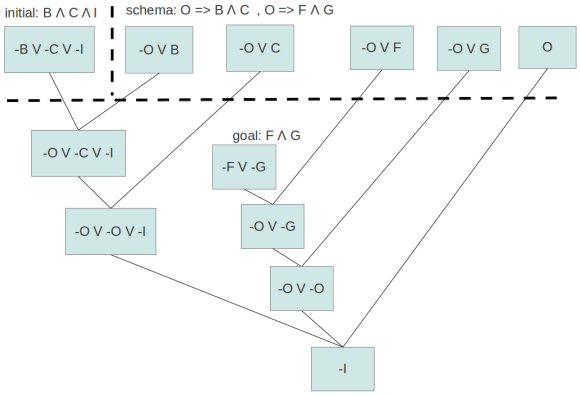


Figure 5: *Applied resolution*

Once we have inductively formed actions that work for cases, we can try to use them in forward reasoning mode for solving new problems. A sequence of actions takes the system through a sequence of states if we have a proof as presented in the figure 3. The solutions that are found are good cases as they are localized to our environment.

Rule association mining [Ting et al. 2010] can be used for handling the problem that some cases show exceptional behavior. Rules that have high support and confidence levels can be used as parts of the operation learning. In fact we see this as a possibility to statistically manage the classical ramification and qualification problems of AI-Planning.

Running into contradictions requires us to backtrack in our model generation and try other process alternatives. In case of lacking knowledge we may first of all run more specific searches to external knowledge bases and try to generalize a behavior that seems to work in similar situations so that we can use it as a schema that after localization yields processing alternatives that conform to the local conditions. The second thing that can be done is to perform local experiments for filling in the gaps of knowledge. If such cases have not been tested and reported by outside organizations then it is we might have a unique case or a case that for some reason or another has gone unreported. The general criterion for selecting between information searches and experimentation is the ration of information cost to its reliability.

4 Similarity

Fast implementation of the method will be based on efficient indexing that allows finding similar knowledge elements from the KB. Similarity, equality, conformance, matching are all relevant concept in CBP. Taxonomy for similarity concepts is provided by [Cunningham 2009]. The explanation of how there are to be used in the method is a topic of a further article.

Since, unlike in the earlier expert systems based process design methodologies, case based information is searched through process design in all the phases, similarity measures for all knowledge classes are needed. Similarity of ores, similarity of chemicals used for leaching, similarity of process parameters (including device types and capacities) and similarity of process flows are needed for accessing knowledge that might fit our purposes. This forms a different set of criteria than those used for actual model genera-

tion where sub-models are matched at boundaries so that they form larger entities that do not contradict each other.

One could try to device a semantic similarity measure that calculates similarity of classes by counting the objects in the intersection of the two classes and dividing this with the union of objects that belong to the complements in the two classes relative to the intersection.

Semantic Similarity measure [d'Amato et al. 2005]:

$$s(C,D) = \frac{|I^{\mathcal{I}}|}{|C^{\mathcal{I}}| + |D^{\mathcal{I}}| - |I^{\mathcal{I}}|} \cdot max(|I^{\mathcal{I}}|/|C^{\mathcal{I}}|, (|I^{\mathcal{I}}|/|D^{\mathcal{I}}|) \quad (1)$$

The problem in the conceptual learning phase is that when we are in the phase of learning the concepts, i.e. working with the intensions, we are hardly in the position to evaluate the sizes of the extensions. The same basic problems concerns all the possible types of measures listed in [d'Amato et al. 2005], namely filter measures, matching measures and probabilistic measures. There are several question that need to be answered. like should we try to learn the concepts of the domain before the concepts of the processing, or vise verse.

Let's consider the semantic similarity measure presented above. If a case states that a unit process uses a device that belongs to a class D of devices, how can we measure the similarity of this case with another case using a device belonging to a class C. We can easily share the descriptions and form the description of the intersection of the two class descriptions by requiring that it defines the properties of both of the original classes. Contradictory class definition means total dissimilarity. After some normalizations the count of all the properties and their restrictiveness compared to the original classes measures the similarity of the original classes. One could not effectively count the sizes of the interpretations like the semantic similarity measure requires.

5 The method

Let's have a target case $\langle \Pi, P \rangle$, where Π is the planning problem's initial and goal state description, a pair $\langle I, G \rangle$. Typically I is a complete state description and G a goal condition that has a set of complete states as refinements. Since we have generally no a priori knowledge of the plan, P is usually initially an empty set. The standard assumption is that we wish to imply partial order to the members of P. For the sake of simplicity and for the linearity of the proof we assume a total order of action to form the plan.

Now our case base is designed to consist of prior cases, source cases that have the same structure as our target case. The method here is based on finding, for the target state ${}^{i}S_T$, the n k-NN nearest neighbor states by comparing the I parts of the cases. Say these are ${}^{i}S_C^j, j \in [1, n]$.

Now that we have located n similar past states in the case base we can look at what was done in these cases. The linear structure of the plans gives us the ${}^{i}Act_C^j$, which might or might not unify in the logical sense or conform to each other. The cases also tell us the goal reached and these goals are the original goal conditions of the case planning problems, not the refined complete state descriptions. It is useful to retain the results we have gained by having followed the least commitment principle in the past planning efforts.

The decision to be taken at this point includes two dimensions; we can look at the goal part and evaluate the cases based on the distance of the case goal to our target goal definition. Is that where we are aiming at, the other dimension being the action sequence of the case, for which the question is, do we have what it takes do perform this kind of action sequence. After all, the cases can be from other sources than our own past experience.

Goal driven similarity assessment [Janetzko et al. 1992] makes similarity measures sensitive to the context of the search given by the goal. In planning context this goal is naturally the planning problem, particularly the G component in Π.

A small example of the method is in order. Let us deal with a set of unit process cases and leave more complicated plans for complete process chains to future discussions, the scope and purpose of the paper does not allow full coverage of the topic.

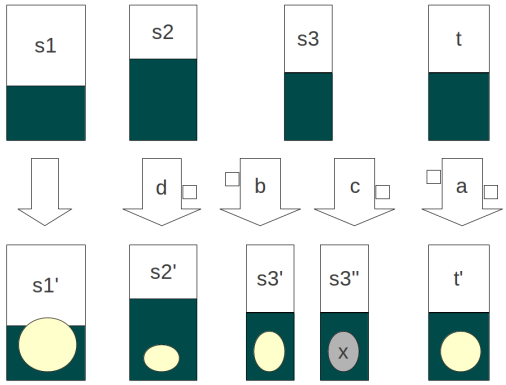


Figure 6: *Source cases and the target case*

In the figure 6 we see three source cases ($s1$,$s2$,$s3$) and the target case t. Consider here that we want to find a single step to get to a goal state where there is a quantifiable result represented by the pale circle in t'. The width, height and portions of dark and light areas of the cases representing rectangle pairs correspond to features of the states, say there are vessels where there is a liquid level that is depicted by the line separating the dark and light areas. States $s2$ and $s3$ are the 2 nearest neighbors of the target state t. The cases, there are two for $s3$,are found in the case base and we see that there are three possible actions in our current state t. Of these the one leading to state $s3''$ is not relevant from the point of view of our goal, since the resulting circle is not the color (say metal) we aim at (there is an unwanted X in) . Based on the two remaining actions and our local domain theory we can construct action a that is represented in the figure with rectangle on both side as it inherits characteristic of the two case actions that we found promising (c and b) . Now we can deduce the state t' since we know t and the action a. We then check if the circle in t' is pale enough and large enough to meet our goal conditions and if they are we store the target knowledge as promising solution candidate. In a realistic scenario the plans are longer and we would iterate in the CBP-Cycle by searching promising partial cases and constructing the target plan gradually from the partial solution candidates.

The method would use the case attribute values, attribute minimum and maximum values, the ranges, and vessel ontologies in the case search and target generation. Say we have ontological information included in the case information: in $s2$ the vessel is of type $v2$ and in $s3$ the vessel is of type $v3$. Then we can find what is the tightest category where both of the vessels belong to. In description logic this information can be deduced by answering to query $v2 \sqsubseteq v \wedge v3 \sqsubseteq v \wedge \nexists w : (v1 \sqsubseteq w \wedge v2 \sqsubseteq w \wedge w \sqsubseteq v \wedge v \neq w)$. In order for us to be able to process the state t to become the state t',

we need a vessel v_t that fulfills the general inclusion axiom $v_t \sqsubseteq v$, that is our vessel is compatible as it is of the same kind used in the relevant cases.

Since we need to deal with constraints, limits and ranges we are considering use of constraint programming and constraint logic programming. These provide tools that we can integrate with the tools based on inductive logic programming that can be used in the implementation of the learning components of the method. The next thing to consider is the ways that general automated reasoning theorem proving allows in the handling of functions by techniques that do reduction, unification and term rewriting. Again we need to restrict these topic to be outside the scope of this paper. Conceptual process design being symbolic in nature does not cover the optimization and simulation functions, but generates a process draft that can be used as a starting point for these tools doing further analysis. Still some rudimentary ability to handle similarity and produce action effects of knowledge representation containing functions is needed in the conceptual process design phase. Possibility to do reasoning in the presence of spacial function [Akbarpour and Paulson 2010] is one promising extension to the capabilities.

Mineralogical literature reports cases where ore characteristics are listed along the unit process description in terms of the contents of the leachant and the conditions under which the process is performed. Based on these reports the literature presents decision diagrams for selecting the correct process. In figure 7 some areas in the ore space is presented. For ore with grain size between 5 and 50 μm, ores containing considerable amounts of pyrites and not having a substance that interferes with the process, cyanide leaching is the recommended action. So, for cases falling in the area A (with the two parts, excluding the arc shaped area B) cyanide leaching is executable.

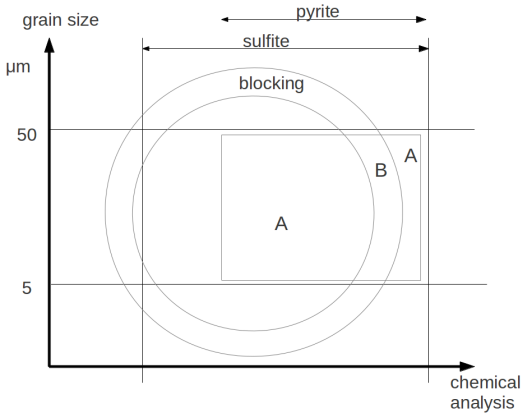


Figure 7: *Inductive hypothesis*

The classical decision tree checks variables one by one in fixed order, given by the knowledge engineer at the time the system was implemented. In a situation, like the one in CBP here, in presence of several sub-theories, several representational layers and solver modes more elaborate control mechanisms are clearly called for. The kind of fixed ordering of decision trees is too restricting. Modern theorem provers and satisfiability solvers have inbuilt control mechanisms that direct the search for solutions based on heuristics that use the logical clausal structures in deciding the search directions.

A recommended way to build such CBR-systems is to use all the available*a priori* information. So we are faced with the problem of how to integrate pure CBR with model based reasoning. If things go well the result is that we can use the experiences gained in the past. However we also aim at breaking the mold when such opportunities arise. The system should support process innovation and discovery of new ways to compose processes from elementary parts, i.e. the unit processes.

Machine learning literature [Alpaydin 2010] uses the *a priori* information in two phases. The attributes are recognized in the same way as above, we have the grain sizes and chemical compositions. Then we have a large number of cases where the attribute values and the decision made bases on the case data are given. Now what is needed is a good guess of what are the characterizing attributes and how to set the limits for values so that areas in the case base are such that they capture the cases as the case specific decisions indicate. These selections form a bias called the inductive bias.

6 Prototype

Decision trees for selecting the processes have been traditionally used in our case study field of gold extraction, like in [Marsden and House 1992]. These are reflected in the expert systems like in [Torres et al. 1999]. Prolog programming language is one of the common and flexible tools for building AI-systems [Bratko 1990]. We chose to use Prolog for building a prototype system for conceptual process design, for some of the corner stone parts of it.

The rules for deciding what unit processes to use are coded as a Prolog-program. The following rule checks if there is pyrite in the ore, the grain size of the ore is in a certain range, and there is no substance in the ore that blocks the type of leaching from succeeding. The first line of the rule, the so called head of the rule, states that situation where there is a possibility of interference with cyanide leaching occurs for an ore, here designated by the variable Ore, if the conditions represented by the rest of the rule lines, the so called rule body, hold.

```
action(cyanide_leaching, Ore) :-
         pyrite(Pyrite),
         member([Pyrite, _],Ore),
         member([grain_size, Size],Ore),
         Size >= 5,
         Size =< 50,
         member([Blocking,_],Ore),
         \+interfere(Pyrite,Blocking),
         !.
```

The rule base consisting of this kind of rules constitutes to one module of the program. The last line, with the exclamation mark here means that the proceeding *action* alternatives are not considered and the line before that means that *interfere* does not hold between two features of the ore, namely a pyrite and some blocking substance.

The domain model is represented in one knowledge-base, storing the chemical, physical and mineralogical laws of nature. Things contained are facts about what chemical compounds are pyrites and that arsenopyrites are pyrites etc. Also device ontologies are coded in Prolog.

```
pyrite(arsenopyrite).
pyrite(pyrite).

sulfide(chalcopyrite).
sulfide(pyrrhotite).
sulfide(galena).
sulfide(sphalerite).
```

```
sulfide(nickel_sulfide).
% pyrites ase also sulfides
sulfide(Pyrite) :- pyrite(Pyrite).
```

Then there is a case bases, which in the prototype is presented as a list of attribute-value pairs. In the future implementations a faster access will be implemented through a more clever indexing scheme. Once we have completed the prototype in Prolog we aim at porting it to an environment where effective data mining and automated theorem proving tools can be used. In the mean while inductive techniques and constraint logic programming techniques can experimented based on the prototype implementation.

The prototype models the executability of actions in different situations, taking into consideration various constraints that both make the actions possible, but also according to some less concrete criteria allowed. States that would be harmful in some way can be avoided in this way. Our research project is designed so that it models first the unit processes and the domain, followed by investigation of the ways to compose the overall processes. In this coming phase the search and planning heuristic techniques will be imported to the prototype.

7 Conclusions

The approach allows integration of domain theoretical elements to the framework mainly based on the state attributes and values as well as ontological information. We wish to next to include action models with resource description and temporal constraints into our reasoning method. The language is essentially first-order logic (predicate logic) without function symbols. However the possible extensions to the language in this respect are also discussed in the paper.

A systematic knowledge engineering methodology for managing the distributed, heterogeneous knowledge from varied sources and combining the reasoning systems to support the CBR-cycle at it's different phases would be nice to have and our future research directs to this direction. In this article we have presented some of the initial views of how this kind of engineering methodology could be constructed.

Naturally we can not do much about the worst case complexity that remains the same for both the generative and variant techniques, but recent advances in rule association mining [Ting et al. 2010] and kernel functions [Serina 2010] as well as ontological support bringing in representation that remain guarded from the point of view of computational complexity promise a lot. These will include some of the research directions where we aim at.

The method is generalizable to other domains and we see that in countries like Finland where process industries play a central role supported by industries producing large scale investment products to be used in the process- and manufacturing industry, this kind of enhanced conceptual process construction methodology, platform and tool set can play an important role.

Facilitated by the wider availability and faster access to case data, the ability of modern automated reasoning tools sets the stage for development of such decision support systems that have for long been on the wish list of many organizations and individuals. Were we have used industrial process design as the area of application and hydro-metallurgical processing as a case example. Case based generalization and learning evades the knowledge acquisition bottleneck experienced by the early expert system applications. Use of general description logic language makes it possible to use general purpose solvers that are shown to be efficient. The policy to use general purpose tools also promotes extensions of the description language in a way that both makes it possible to share knowledge with the outside world more flexibly and helps to avoid bringing in language structures that lead to high computational complexity.

Acknowledgments

This work has been done in the LOWGRADE project of the ELEMET research program funded by FIMECC Oy. The financial support of TEKES and Outotec Oyj is gratefully acknowledged. I would like to thank the project for providing inspiring and motivating environment for this work. Especially discussions with Lotta Rintala, directions given by Jari Aromaa and motivating comments by Olof Forsén have been invaluable. I would also like to thank the anonymous reviewers who provided constructive critique at a stage when it really was needed.

References

AKBARPOUR, B., AND PAULSON, L. 2010. Metitarski: An automatic theorem prover for real-valued special functions. *Journal of Automated Reasoning 44*, 175–205. 10.1007/s10817-009-9149-2.

ALPAYDIN, E. 2010. *Introduction to Machine Learning*, 2nd ed. The MIT Press, Cambridge.

AURIOL, E., MANAGO, M., DIETER ALTHOFF, K., WESS, S., AND DITTRICH, S. 1994. Integrating induction and case-based reasoning: Methodological approach and first evaluations. In *Proc. 17th Conference of the GfKl*, Springer Verlag, 18–32.

BERGMANN, R., AND GIL, Y. 2011. Retrieval of semantic workflows with knowledge intensive similarity measures. In *ICCBR*, Springer, A. Ram and N. Wiratunga, Eds., vol. 6880 of *Lecture Notes in Computer Science*, 17–31.

BRATKO, I. 1990. *PROLOG Programming for Artificial Intelligence, Second Edition*. Addison-Wesley.

CUNNINGHAM, P. 2009. A taxonomy of similarity mechanisms for case-based reasoning. *IEEE Trans. Knowl. Data Eng. 21*, 11, 1532–1543.

D'AMATO, C., FANIZZI, N., AND ESPOSITO, F. 2005. A semantic similarity measure for expressive description logics. In *PROCEEDINGS OF CONVEGNO ITALIANO DI LOGICA COMPUTAZIONALE, CILC05*.

FOX, M., GEREVINI, A., LONG, D., AND SERINA, I. 2006. Plan stability: Replanning versus plan repair. In *ICAPS*, AAAI, D. Long, S. F. Smith, D. Borrajo, and L. McCluskey, Eds., 212–221.

HAYES, P. C., AND GRAY, P. M. J. 1985. *Process selection in extractive metallurgy / by P.C. Hayes ; with contributions from P.M.J. Gray ... [et al.]*. Hayes Publishing, Brisbane :.

HUME, D., AND SAMMUT, C. 1991. Using inverse resolution to learn relations from experiments. In *Proceedings of the Eighth International Workshop on Machine Learning*, Morgan Kaufmann, 412–416.

JANETZKO, D., WESS, S., AND MELIS, E. 1992. Goal-driven similarity assessment. In *GWAI-92 16th German Workshop on Artificial Intelligence, volume 671 of Springer Lecture Notes on AI*, Springer Verlag, 283–298.

KAUTZ, H., AND SELMAN, B. 1992. Planning as satisfiability. In *Proceedings of the 10th European Conference of Artificial Intelligence*, John Wiley & Sons, B. Neumann, Ed., 359–363.

MAEDCHE, A., AND STAAB, S. 2002. Measuring similarity between ontologies. In *EKAW*, Springer, A. Gómez-Pérez and V. R. Benjamins, Eds., vol. 2473 of *Lecture Notes in Computer Science*, 251–263.

MARSDEN, J., AND HOUSE, I. 1992. *The chemistry of gold extraction*. Ellis Horwood series in metals and associated materials. Ellis Horwood.

MUGGLETON, S., AND DE RAEDT, L. 1994. Inductive logic programming: Theory and methods. *Journal of Logic Programming 19*, 629–679.

QUINLAN, J., AND CAMERON-JONES, R. 1993. Foil: A midterm report. In *Machine Learning: ECML-93*, P. Brazdil, Ed., vol. 667 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 1–20.

RINTALA, L., AROMAA, J., AND FORSÉN, O. 2011. The use of decision methods in the selection of leaching alternatives. In *The 6th international European Metallurgical Conference*, GDBM Informationsgesellschaft mbH, J. Harre and W. Ulrich, Eds., vol. 5, 1659 –1672.

RINTALA, L., LILLKUNG, K., AND AROMAA, J. 2011. The use of decision and optimization methods in selection of hydrometallurgical unit process alternatives. *Physicochemical Problems of Mineral Processing 46*, 229–242.

RUSSELL, S., AND NORVIG, P. 2003. *Artificial Intelligence: A Modern Approach*, 2nd edition ed. Prentice-Hall, Englewood Cliffs, NJ.

SERINA, I. 2010. Kernel functions for case-based planning. *Artificial Intelligence 174*, 1617, 1369 – 1406.

TING, S. L., WANG, W. M., KWOK, S. K., TSANG, A. H. C., AND LEE, W. B. 2010. Racer: Rule-associated case-based reasoning for supporting general practitioners in prescription making. *Expert Syst. Appl. 37*, 12 (Dec.), 8079–8089.

TORRES, V., CHAVES, A., AND MEECH, J. 1999. Process design for gold ores: A diagnostic approach. *Minerals Engineering 12*, 3, 245 – 254.

WADSWORTH, M. 1987. *Handbook Of Separation Process Technology*. Wiley & Sons, ch. Leaching - Metals Applications, 500–539.

WESS, S., DIETER ALTHOFF, K., AND DERWAND, G. 1994. Using k-d trees to improve the retrieval step in case-based reasoning. In *Topics in Case-Based Reasoning*, Springer-Verlag, S. Wess, K.-D. Althoff, and M. M. Richter, Eds.

Man-in-the-browser -hyökkäyksistä Ajax-sovelluksissa

Sampsa Rauti
Turun yliopisto
sampsa.rauti@utu.fi

Ville Leppänen
Turun yliopisto
ville.leppanen@utu.fi

1 Johdanto

Man-in-the-browser on verkkoselaimeen asentuva haittaohjelma, joka pystyy muokkaamaan verkkosivujen ulkomuotoa ja verkkoliikennettä. Haittaohjelma toimii käyttäjän ja verkkopalvelimen välissä muokaten tietoja ja esittäen kummallekin olevansa toinen osapuoli, eikä käyttäjä tai palvelin huomaa tilanteessa mitään epäilyttävää. Ohjelma voi myös taltioida tietoja, mutta tässä keskitymme datan muokkaamiseen.

Haittaohjelma voi vaikkapa odottaa, että käyttäjä kirjautuu sisään jollekin tietylle sivulle. Käyttäjän syöttäessä esimerkiksi tilisiirtoon liittyviä tietoja ohjelma muokkaa niitä käyttäjän huomaamatta ennen lähetystä. Kun palvelin lähettää takaisin varmistuksen tiedoista, voi haittaohjelma muokata verkkosivua siten, että väärät tiedot muutetaan taas oikeiksi. Käyttäjän näkökulmasta on siis tapahtunut täysin normaali toimenpide, mutta palvelin on saanut väärät tiedot.

Man-in-the-browser -hyökkäystä voidaan siis pitää perinteisen man-in-the-middle -hyökkäyksen alatyyppinä ja kehittyneempänä muotona. Koska käyttäjän koneelle on paljon helpompi hyökätä kuin palvelimille, näiden haittaohjelmien määrä on viime vuosina kasvanut nopeasti [Entrust 2010]. Hyökkäys voidaan toteuttaa monella tavalla, kuten selainlaajennuksilla tai soluttamalla haittakoodia verkkosivulle (cross site scripting -haavoittuvuus). Tämä artikkeli keskittyy Firefox-selainlaajennuksiin perustuviin man-in-the-browser -hyökkäyksiin Ajax-sovelluksissa.

2 Aiempi tutkimus

Man-in-the-browser -ongelma on saanut melko vähän huomiota. Philipp Gühring käsittelee ongelmaa ja sen mahdollisia ratkaisuja tärkeässä artikkelissaan [Gühring 2007]. Kattava kuvaus on myös lähteessä [Dougan & Curran 2012]. Lisäksi turvallisuusalan yritykset ovat laatineet aiheesta useita teknisiä raportteja.

Edistyneimpiä keinoja estää haittaohjelmia muokkaamasta lähetettäviä tietoja ovat esimerkiksi toisen tiedonvälityskanavan käyttö varmistukseen tai tietokoneeseen liitettävä tiedot digitaalisesti allekirjoittava laite. Ensimmäisessä tapauksessa esimerkiksi tilisiirto voidaan varmistaa pankin lähettämällä tekstiviestillä, johon käyttäjä vastaa. Pienellä näppäimistöllä ja näytöllä varustettu lisälaite taas voi allekirjoittaa tiedot jo ennen kuin ne siirtyvät tietokoneelle ja estää tällä tavalla niiden muokkaamisen. Näissäkin keinoissa on kuitenkin omat ongelmansa, eivätkä ne estä täysin haittaohjelmien toimintaa.

Haitallisia Firefox-laajennuksia ja toimenpiteitä niitä vastaan on tutkittu lähteessä [Ter Louw et al. 2007]. Ajaxin heikkouksista yleisesti on paljon julkaisuja, kuten [Di Paola & Fedon 2006].

3 Ajax ja selainlaajennukset

Ajax (Asynchronous JavaScript and XML) on joukko tekniikoita, joilla lisätään verkkosivujen interaktiivisuutta. Sen avulla selain voi lähettää ja vastaanottaa dataa lataamatta koko sivua uudelleen. Ajax koostuu HTML-, CSS-, JavaScript-, ja DOM-tekniikoista sekä XMLHttpRequest-olioiden käytöstä.

Selainlaajennukset ovat verkkoselaimen perustoiminnallisuutta laajentavia lisäosia. Firefoxin laajennuksilla on käytännössä samat oikeudet kuin selaimella itsellään, mikä valitettavasti tarkoittaa myös sitä, että ne voivat tehokkaasti hyödyntää Ajaxin ja selaimen heikkouksia. Laajennukset toteutetaan JavaScriptillä, ja niiden avulla hyökkääjä voi muokata minkä tahansa verkkosivun ulkoasua ja muuttaa verkkoliikennettä haluamallaan tavalla.

4 Hyökkäykset

Erilaisia hyökkäystapoja ja niiden tehokkuutta ja helppoutta tutkiaksemme toteutimme neljä erilaista haittaohjelmaa mallintavaa selainlaajennusta Firefoxille. Tässä luvussa esitellään näiden hyökkäysten pääperiaatteet.

4.1 Verkkoliikenteen muokkaaminen

XPCOM (Cross Platform Component Object Model) on Mozillan kehittämä olio- ja komponenttimalli, jonka avulla lähes kaikkia Firefox-selaimen ominaisuuksia voi käyttää laajennusten JavaScript-toteutuksissa. Tämän mallin rajapintoja hyödyntämällä Firefox-laajennus voi tarkkailla ja muokata kaikkia verkkoselaimen lähettämiä HTTP-pyyntöjä ja niiden vastauksia. Käytännössä tämä tehdään toteuttamalla XPCOM:n nsIObserver-rajapinta, jonka avulla voidaan kuunnella HTTP-pyyntöjä ja niihin saapuvia vastauksia vastaavia http-on-modify-request- ja http-on-examine-response -tapahtumia, joiden kautta päästään lukemaan ja muokkaamaan siirrettävää dataa.

Tämän hyökkäystavan toteutus on muihin menetelmiin verrattuna kohtalaisen vaativaa. Pelkän yksinkertaisen merkkijonojen korvaamisen sijaan joudutaan ensin käsittelemään ns. HTTP-kanavaa ja lukemaan siihen liittyvää tietovirtaa. Verkkosivun sisältöä onkin selvästi helpompi muokata DOM:n (Document Object Model), HTML-sivujen sisällön muokkaukseen tarkoitetun rajapinnan kautta. Toisaalta tämä menetelmä on näkymätön, sillä se ei tee muutoksia kohdesivun JavaScript-koodiin tai DOM-puuhun. Hyökkääjä voi myös tehdä dataan muutoksia joutumatta paneutumaan tarkasti kohdesivun toiminnallisuuteen.

4.2 XMLHttpRequest-prototyypin muokkaaminen

JavaScriptissä ei ole luokkia, vaan oliot luodaan ns. prototyypin pohjalta. Uudelleentoteuttamalla XMLHttpRequest-prototyyppi voidaan määritellä uudelleen kaikkien tämäntyyppisten olioiden käyttäytyminen. Ajax-sovelluksissa käytetään XMLHttpRequest-olioita lähettämään pyyntöjä palvelimelle ja vastaanottamaan vastauksia, joten hyökkääjä voi prototyyppiä muuntamalla lisätä sovellukseen haitallista toiminnallisuutta. Tätä kuvaa seuraava koodiesimerkki:

```
XMLHttpRequest.prototype.originalSend =
```

```
XMLHttpRequest.prototype.send;

var evilSend = function(data) {
  // Muokkaa dataa
  this.originalSend(data);
};

XMLHttpRequest.prototype.send = evilSend;
```

XMLHttpRequest-prototyypin send-funktio on tässä asetettu hyökkääjän kirjoittamaksi funktioksi, mikä aiheuttaa kaikkien verkkosivun JavaScript-koodissa luotujen XMLHttpRequest-olioiden muuttumisen. Man-in-the-browser -hyökkäyksessä send-funktiossa halutaan luonnollisesti muokata lähetettävää dataa, minkä jälkeen kutsutaan sen alkuperäistä toteutusta.

Palvelimelta saapuvan paluuviestin tapauksessa voidaan asettaa vastauksen käsittelijäksi hyökkääjän itse laatima funktio, joka simuloi sovelluksen normaalia toimintaa, mutta muokkaa sivulle asetettavaa dataa hyökkääjän tarkoituksiin sopivasti:

```
XMLHttpRequest.prototype.onreadystatechange
= handler;
```

4.3 DOM-puun muokkaaminen

Haittaohjelmat voivat myös muuttaa verkkosivuja muokkaamalla suoraan niiden DOM-puuta. Kun käyttäjä syöttää sovellukseen tietoja, haittaohjelma kuuntelee verkkosivun tapahtumia ja toimii seuraavaksi kuvatulla tavalla. Kun haluttu tekstikenttä muuttuu aktiiviseksi, se muutetaan CSS-tyylimuotoiluja käyttämällä näkymättömäksi ja korvataan uudella, aivan samanlaisella tekstikentällä. Alkuperäisen näkymättömän kentän arvoa muokataan hyökkääjän haluamalla tavalla. Käyttäjä kirjoittaa tietonsa uuteen kenttään luullen kaiken olevan kunnossa, mutta oikeasti palvelimelle lähtee alkuperäisessä näkymättömässä tekstikentässä oleva väärä tieto.

Ajax-sovelluksissa käyttäjän syöttämien tietojen varmistus ei tavallisesti aiheuta koko sivun lataamista uudelleen, mutta sivua pitää kuitenkin muuttaa. Näin ollen haittaohjelma voi kuunnella esimerkiksi DOMNodeInserted-tapahtumaa, joka laukaistaan, kun sivulle lisätään uusi elementti. Tämän elementin teksti muutetaan käyttäjän antamaa oikeaa arvoa vastaavaksi.

4.4 Sovelluksen toiminnallisuuden muokkaaminen

Ajax-sovelluksen toiminnallisuutta voidaan muokata suoraan muuttamalla verkkosivulla olevaa JavaScript-koodia. Tämä onnistuu luomalla uusia JavaScript-koodia sisältäviä script-tageja. Nämä voidaan liittää sivulle appendChild-metodilla, joten kyseessä on oikeastaan DOM-puun muokkauksen erikoistapaus. JavaScript-tagi voidaan myös korvata kokonaan replaceChild-metodilla, mutta uuden koodinkin lisääminen riittää usein, sillä JavaScriptissä toinen samanniminen funktio korvaa aiemman toteutuksen. Tällainen toiminnallisuuden korvaaminen edellyttää hyökkääjältä sovelluksen hyvää tuntemusta.

5 Vastatoimia

Tässä luvussa ehdotetaan muutamia JavaScriptiin itseensä perustuvia keinoja vähentää man-in-the-browser -hyökkäyksen onnistumismahdollisuuksia. Koska näissä menetelmissä muutetaan Ajax-sovelluksen lähdekoodia, jota myös hyökkääjä voi muokata, ne ovat kierrettävissä. Samalla ne kuitenkin vaikeuttavat selvästi hyökkääjän työtä.

Verkkoliikenteen muokkaamisen Firefox-laajennuksen XPCOM-rajapinnan kautta voi estää salaamalla lähetettävän tiedon. Näin hyökkääjä ei voi enää muokata viestin dataa suoraan, vaan hänen on pakko muuttaa tavalla tai toisella verkkosivua tai sillä olevaa Ajax-sovellusta. Salaaminen toimii myös muokattua XMLHttpRequest-prototyyppiä vastaan. Haittapuolena on, että viesti on salattava asiakaspuolella. Salauksen suorittavan funktion nimeä ja toteutusta voisi muutella satunnaisesti hyökkäyksen hankaloittamiseksi.

DOM-puun muokkauksen tapauksessa verkkosivun tapahtumia voi valvoa: esimerkiksi piilotetun tekstikentän arvon muuttuminen on epäilyttävää. Elementtien id-attribuutteja ja muita ominaisuuksia voi myös satunnaisesti vaihdella, ettei niihin pääse käsiksi yhtä helposti.

Vaihtelemalla funktioiden nimiä ja toteutuksia sekä sekoittamalla niiden koodia istuntokohtaisesti voidaan hämätä hyökkäjää, joka yrittää muuttaa Ajax-sovelluksen toiminnallisuutta. Funktioita voidaan jopa korvata sovelluksen ajon aikana lähettämällä palvelimelta niille uusia toteutuksia. Tämä lyhentää koodin analysointiin käytössä olevaa aikaa.

Tällaiset toimenpiteet monimutkaistaisivat sovellusta, mutta ohjelmoijan ei välttämättä tarvitse olla niistä riippuvainen. Esimerkiksi Google Web Toolkit -sovellukset voidaan kirjoittaa Javalla, josta generoidaan automaattisesti JavaScriptiä asiakaspuolelle. Tämänkaltaisiin työkaluihin voitaisiin lisätä man-in-the-browser -hyökkäyksiä ehkäisevää toiminnallisuutta, koska ohjelmoijan ei alun perinkään tarvitse välittää asiakaspuolen JavaScript-toteutuksesta.

6 Johtopäätökset

Ajax-tekniikat kasvattavat verkkosivujen hyökkäyspinta-alaa ja selainlaajennusten useissa selaimissa saamat laajat oikeudet pahentavat tilannetta entisestään. Tämän vuoksi man-in-the-browser -hyökkäykset ovat varteenotettava uhka, johon on olemassa vain osittaisia vastatoimenpiteitä.

Lähteet

Di Paola, S., Fedon, G. Subverting Ajax. Proceedings of 23rd CCC Conference, Berlin (2006).

Dougan T., Curran, K. Man in the Browser Attacks. *International Journal of Ambient Computing and Intelligence*, volume 4, issue 1, March 2012. IGI Global.

Entrust. 2010. Defeating Man-in-the-Browser. http://download.entrust.com/resources/download.cfm/24002/, viitattu maaliskuussa 2012.

Gühring, P. 2006. Concepts against Man-in-the-Browser attacks. http://www.cacert.at/svn/sourcerer/CAcert/SecureClient.pdf, viitattu maaliskuussa 2012.

Ter Louw, M., Lim J. S., Venkatakrishnan, V. N. Enhancing web browser security against malware extensions. *Journal in Computer Virology,* volume 4, number 3, August 2008. ISSN: 1772-9890. Springer Paris.

Sockets and Beyond: Assessing the Source Code of Network Applications

Miika Komu† Samu Varjonen‡ Andrei Gurtov‡
Sasu Tarkoma‡
† Aalto University, Department of Computer Science and Engineering
‡ University of Helsinki and Helsinki Institute for Information Technology
† firstname.lastname@aalto.fi ‡ firstname.lastname@hiit.fi

Abstract

The Sockets API is the low-level interface for developing network applications for TCP/IP networks. It forms also the basis for network application frameworks that abstract away the details of the Sockets API because it is burdened with legacy support and its APIs are quite intricate to program. In this work, we have analyzed open-source software to better understand how the Sockets API is used today. More specifically, we have analyzed statistically a number of C-based applications that use the Sockets API directly in Ubuntu Linux. We analyzed many aspects, including the use of different API functions and also the use OpenSSL-based security. As the most important finding, we discovered that 28.6% of the C-based network applications in Ubuntu are vulnerable to security-related attacks because they fail to initialize OpenSSL properly. We also analyzed four frameworks manually and discovered that all of them consistently failed to support UDP-based multihoming and parallel IPv4/IPv6 connection initialization for the clients. The details are published in a research report [Komu et al. 2011].

1 Introduction

The Sockets API is the basis for all internet applications. While the number of applications using it directly is large, some applications use it indirectly through intermediate libraries or frameworks to hide the intricacies of the low-level Sockets API. Nevertheless, it is then the intermediaries that still have to interface with the Sockets API. Thus, the Sockets API is important for all network applications either directly or indirectly but has been studied little. To fill in this gap, we have applied static code analysis over seven hundred C-based applications in Ubuntu Lucid Linux to understand how they utilize Sockets API related constants, structures and functions.

2 A Summary of the Sockets API Findings and Their Implications

Table 1 highlights ten of the most important findings in the Sockets APIs. Next, we go through each of them and argue their implications to the development of network applications.

Core Sockets API		
1	IPv4-IPv6 hybrids	26.9%
2	TCP-UDP hybrids	26.3%
3	Obsolete DNS resolver	43.3%
4	UDP-based apps with multihoming issue	45.7%
5	Customize networking stack	51.4%
OpenSSL-based applications		
6	Fails to initialize correctly	28.6%
7	Modifies default behavior	53.3%
8	OpenSSL-capable applications in total	10.9%
Estimations on IPv6-related extensions		
9	Potential misuse with mapped addresses	83.3%
10	Explicit IPv6 Source address selection	66.9%

Table 1: *Highlighted indicator sets and their reference ratios*

Finding 1. The number of hybrid applications supporting both IPv4 and IPv6 was fairly large. While this is a good sign for the deployment of IPv6, the dual addressing scheme doubles the complexity of address management in applications. At the client side, the application has to choose whether to handle DNS resolution over IPv4 or IPv6, and then create the actual connection with either family. As IPv6 does not even work everywhere yet, the client may initiate communications in parallel with IPv4 and IPv6 to minimize latency. Respectively, server-side applications have to listen for incoming data flows on both families.

Finding 2. The hybrid applications using both TCP and UDP amounted as much as TCP-only applications. Thus, application developers seem to write many application protocols to be run on with both transports. While it is possible to write almost identical code for the two transports, the Sockets API favors different functions for the two. This unnecessarily complicates the application code.

Finding 3. The obsolete DNS resolver was referenced twice as more than the new one. This has negative implications on the adoption of new Sockets API extensions that are dependent on the new resolver. As concrete examples, native APIs for HIP and source address selection for IPv6 may experience a slow adoption path.

*Finding 4.*We discovered a multihoming problem that we estimated to affect 45.7% of UDP-based applications: a client sends a request to an address of the server but the server responds using another, causing the client to drop the packet.

Finding 5. Roughly half of the networking software is not satisfied with the default configuration of networking stack and alters it with socket options, raw sockets or other low-level hooking. However, we did not discover any patterns (besides few popular, individually recurring socket options) to propose as new compound socket option profiles for applications.

Findings 6, 7 and 8. Roughly every tenth application was using OpenSSL but surprisingly many failed to initialize it appropriately with SSL_library_init(), thus creating potential security vulnerabilities. Half of the OpenSSL-capable applications were modifying the default configuration in some way. Many of these tweaks improved backwards compatibility at the expense of security. This opens a question why backwards compatibility is not well built into OpenSSL and why so many "knobs" are even offered to the developer.

Finding 9. IPv6-mapped IPv4 addresses should not be leaked to the wire for security reasons. As a solution, the socket option *IPV6_V6ONLY* would prevent this leakage. However, only one out of total six applications using mapped addresses were actually using the socket option. Despite the number of total applications using mapped address in general was statistically small, this is an alarming sign because the number can grow when the number of IPv6 applications increases.

Finding 10. IPv6 source address selection lets an application to choose the type of an IPv6 source address instead of explicitly choosing one particular address. The extensions are not adopted yet, but we estimated the need for them in our set of applications.

Our coarse-grained estimate is that two out of three IPv6 applications might utilize the extensions.

We have now characterized current trends with C-based applications using Sockets API directly and highlighted ten important findings. Of these, we believe findings 3, 4, 6 and 9 can be directly used to improved the existing applications in our data set. We believe that most of the remaining ones are difficult to improve without introducing changes to the Sockets API (findings 1, 2, 5) or without breaking interoperability (finding 7). Also, many of the applications appear not to need security at all (finding 8) and the adoption of extensions (finding 10) may just take some time.

As some of the findings are difficult to adapt to the applications using Sockets API directly, perhaps indirect approaches as offered by network application frameworks may offer easier migration path. For example, the first two findings are related to management of complexity in the Sockets API and frameworks can be used to hide such complexity from the applications.

3 Generic Requirements and A Summary of the Framework Results

Since frameworks can offer high-level abstractions that do not have to mimic the Sockets API layout, we organized the analysis of the frameworks in a top-down fashion. Consequently, the following list reflects the Sockets API findings as generalized categories, i.e, *R1* End-host naming, *R2* Look up of end-host names, *R3* Multiplicity of end-host names, *R4* Multiplicity of transport protocols, and *R5* Security:

R1.1 Does the API of the framework support symbolic host names in its APIs, i.e., does the framework hide the details of hostname-to-address resolution from the application? If this is true, the framework conforms to a similar API as proposed by Name Based Sockets as described in [Komu et al. 2011]. A benefit of this approach is that implementing requirements R1.2, R2.2, R3.1 and 3.3 becomes substantially easier.

R1.2 Are the details of IPv6 abstracted away from the application? In general, this requirement facilitates adoption of IPv6. It could also be used for supporting Teredo based NAT traversal transparently in the framework.

R1.3 IPv6-mapped addresses should not be present on the wire for security reasons. Thus, the framework should manually convert mapped addressed to regular IPv4 addresses before passing to any Sockets API calls. Alternatively, the frameworks can use the *AI_V4MAPPED* option as a safe guard to prevent such leakage.

R2.1 Does the framework implement DNS look ups with *getaddrinfo()*? This is important for IPv6 source address selection and native HIP API extensions because they are dependent on this particular function.

R2.2 Does the framework support parallel DNS look ups over IPv4 and IPv6 to optimize latency?

R3.1 IPv6 source address selection is not widely adopted yet but is the framework modular enough to support it especially at the client side? As a concrete example, the framework should support inclusion of new parameters to its counterpart of *connect()* call to support application preferences for source address types.

R3.2 Does the server-side multihoming for UDP work properly? As described earlier, the framework should use *SO_BINDTODEVICE* option or *sendmsg()/recvmsg()* interfaces in a proper way.

R3.3 Does the framework support parallel *connect()* over IPv4 and IPv6 to minimize the latency for connection set-up?

R4.1 Are TCP and UDP easily interchangeable? "Easy" here means that the developer merely changes one class or parameter but the APIs are the same for TCP and UDP. It should be noted that this has also implications on the adoption of SCTP and DCCP.

R5.1 Does the framework support SSL/TLS?

R5.2 Does the SSL/TLS interface provide reasonable defaults and abstraction so that the developer does not have to configure the details of the security?

R5.3 Does the framework initialize the SSL/TLS implementation automatically?

We summarize how the requirements were met by each of the four frameworks in Table 2. Some of the requirements were unmet in all of the frameworks. For example, all frameworks failed to support UDP-based multihoming (R3.2) and parallel IPv4/IPv6 connection initialization for clients (R3.3). Also, SSL/TLS initialization (R5.3) was not implemented correctly in all frameworks. In total, 56 % of our requirements were completely met in all of the frameworks.

Req.	ACE	Boost::Asio	Java.net	Twisted
R1.1	✓		✓	(✓)
R1.2	✓	✓	✓	
R1.3	✓	✓	✓	N/A
R2.1	✓	✓	✓	
R2.2				
R3.1	✓	✓	✓	✓
R3.2				
R3.3				
R4.1	✓	✓		(✓)
R5.1	✓	✓	✓	✓
R5.2	✓	✓	✓	✓
R5.3	✓	(✓)	✓	(✓)

Table 2: *Summary of how the frameworks meet the requirements*

4 Conclusions

In this article, we summarized empirical results based on a statistical analysis of open-source network software. Our aim was to understand how the Sockets APIs and its extensions are used by network applications and frameworks. We highlighted ten problems with security, IPv6 and configuration. In addition to describing the generic technical solution, we also reported the extent of the problems. As the most important finding, we discovered that 28.6% of the C-based network applications in Ubuntu are vulnerable to attacks because they fail to initialize OpenSSL properly.

We applied the findings with C-based applications to four example frameworks based on the Sockets API. Consequently, we proposed 12 networking requirements that were completely met by a little over half of the frameworks in total. Also the TLS/SSL initialization issue was present in some of the frameworks. With the suggested technical solutions for Linux, we argue that handheld devices with multiaccess capabilities have improved support for UDP, the end-user experience can be improved by reducing latency in IPv6 environments and security is improved for SSL/TLS in general.

References

KOMU, M., VARJONEN, S., TARKOMA, S., AND GURTOV, A. 2011. Sockets and Beyond: Assessing the Code of Network Applications. In *Aalto University publication series SCIENCE + TECHNOLOGY*, vol. 46. ISSN 1799-490X (pdf).

Application Awareness in Redundancy Elimination

Sumanta Saha*, Andrey Lukyanenko, Antti Ylä-Jääski
Department of Computer Science and Engineering
Aalto University

Abstract

Redundancy Elimination (RE) algorithms, which typically are used for improving network performance, work by splitting network packet payload in content dependent chunks and comparing their fingerprints for similarity. However, most of the proposed algorithms are application independent, and in spite of being a positive feature for RE algorithms, they miss many important characteristics of the traffic that could be used to improve performance. This work proposes an Application Aware Redundancy Elimination (AARE) technique, which takes into account the traffic type, and uses extrapolation for similarity estimation. Experiments show that AARE increases the line rate by 400% while keeping the compression factor comparable to that of similar algorithms.

CR Categories: C.2.2 [Computer Systems Organization]: Computer-Communication Networks—Network Protocols;

Keywords: RE, AARE, DPI, cache

1 Introduction

Redundancy elimination (RE) systems, which detect and eliminate similar regions of data from network traffic at a level lower than that of application layer objects, have been of interest in both research community [Saha et al. 2011; Muthitacharoen et al. 2001; Spring and Wetherall 2000; Anand et al. 2009], and commercial products (e.g. Cisco Systems, Inc). These systems are primarily targeted to be used in traffic aggregation points (e.g. core routers or access points) to eliminate redundant traffic across applications. The encoder in an RE system is an upstream node which strips of certain parts of the payload, knowing that it can be correctly reconstructed by the downstream nodes working as a decoder. In this way, the links between the encoder and the decoder has to carry less traffic, which, in case of core links, can be significant. However, although deployed in a limited number of commercial access routers as WAN accelerators, this technique is yet to be deployed broadly in core router environments due to expensive processing and memory accessing tasks.

A primary feature of almost all RE algorithms is application-independence, which allows the algorithms to detect similarity between two completely different application objects. However, it can be shown that limited application awareness can bring better performance out of such algorithms. This paper proposes a solution to this by integrating partial content-awareness with RE systems, which improves the performance by an average of 75% processor-loadwise and around 400%–700% line ratewise, in example cases.

The proposed algorithm, named Application Aware Redundancy Elimination (AARE) works on major content distribution protocols such as HTTP and FTP, and looks for application layer headers to acquire knowledge about the content-type. This process of peeking inside the packet payload to retrieve application signature is named as Deep Packet Inspection (DPI). AARE combines RE algorithms with DPI to make early decision about possible redundant traffic flows.

*e-mail: sumanta.saha@aalto.fi

2 Application Aware RE

From previous research [Pucha et al. 2007; Halepovic et al. 2010], it is evident that detecting similarity between two binary files is possible with a very high probability by inspecting the leading few blocks of data in them. The proposed algorithm in this paper utilizes this observation to increase the performance of existing RE algorithms, such as the ones presented by [Saha et al. 2011], or [Anand et al. 2009]. The algorithm works in two stages:

First Stage, Content Detection: The first step of AARE is based on DPI. Each packet going through the AARE engine is probed for signatures that indicate the content type of the upcoming flow. As a fast and highly effective method, we have tracked the HTTP OK response from the server and use the the HTTP fields "mime-type" and "content-length" to make decision on whether to bypass the flow and for how long.

Algorithm 1 Pseudo-code of AARE algorithm

```
while packet received do
  if HTTP GET Request then
    Set binary-cache, content-length ← 0 and aare-mode ← server-agnostic
  else if HTTP OK Response then
    if mime-type is in [image, video, audio, application] then
      Set binary-cache ← 1 for the flow
      Set content-length ← HTTP Content-Length field
      if aare-hash is present and aare-hash-feature = enable then
        Set aare-mode ← server-assisted
      end if
    end if
  else
    Find the flow for the packet
    if content-length > 0 and binary-cache = 1 then
      if (aare-mode = server-assisted and aare-hash is cache hit) or aare-mode
      = server-agnostic then
        Chunk packet with average size 512
        Check for cache hit
        if All chunks are cache hit then
          Packet Hit ← Packet Hit + 1
        else
          Store the chunks to chunk-store
        end if
        if Packet Hit = n then
          Encode the whole binary file with a shim header for downstream
          nodes to decode
        end if
        Continue to store payload bytes to local file-store
        content-length ← content-length - TCP payload length
      end if
    else
      if content-length = 0 and binary-cache = 1 then
        binary-cache ← 0
      end if
      Continue chunking and matching with traditional RE
    end if
  end if
end while
```

Second Stage, RE Bypass: Second step of AARE relies on the information from the first step to save on the processing power, maintenance, and bandwidth usage. As shown in the pseudo-code of Algorithm 1, in usual cases, all the packets are processed by RE in a similar way, i.e., chunked, fingerprinted, and stored in the cache. However, for the flows marked as *binary-cache*, only the first n packets are chunked and stored in the chunk store, while rest

are passed as is. If the first n packets are a complete match with the chunk-store, AARE decides that the binary file contained in the flow is a cache hit, and then replaces the whole file with only one shim header (a small header with necessary information to identify the actual payload it replaces) containing an identifier for the complete binary file. This replacement is done only if the RE process knows of another RE element downstream containing the same chunks in its cache, so that the downstream element can reconstruct the traffic by inspecting the shim header and replacing it again with actual payload. After several hops, the last downstream RE node having the payload corresponding to the shim-header identifier replaces the header with the actual payload. Thus, the end client never experiences the elimination process happened in the core network.

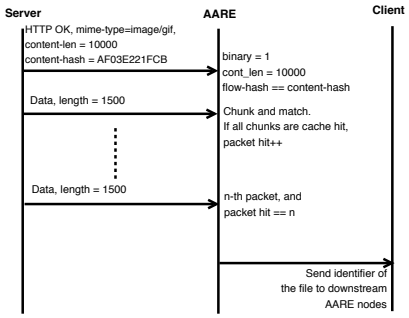


Figure 1: *Sample flowchart of AARE*

Note that, there is a small possibility that two different binary files have the same first chunks (which can happen in rare occasions for chunked binary files, such as png). For these cases, AARE accommodates a server assisted mode, where the server inserts a custom HTTP header field called "aare-hash", which is a signature of the total content, to assist the routers making bypassing decision. This two level of matching ensures a very fast processing as well as low processing requirement. A sample flowchart of AARE is depicted in Figure 1.

3 Evaluation

A proof of concept AARE implementation has been developed in pure C, which along with the two stages of AARE algorithm, contains a traditional RE engine with the CombiHeader plug-in. For details of CombiHeader implementation, which adopts an aggressive greedy nature to combine multiple shim headers into one to generate flexible length chunks for minimizing total bytes on wire, readers are referred to [Saha et al. 2011]. For implementing the first stage of AARE, as described in Section 2, we have used an open source off-the-shelf DPI tool named OpenDPI.

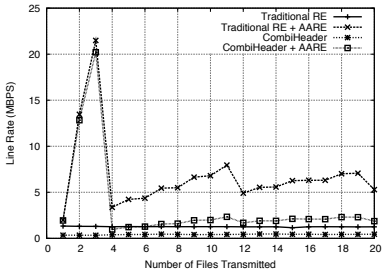


Figure 2: *Comparative line speed*

Bandwidth and Line Rate: Whenever AARE decides on a binary flow hit, rather than sending multiple packets encoded with small shim headers, AARE transmits a single packet containing a unique identifier for the whole binary content. Thus, the process actually replaces possible gigabytes of traffic with a few hundreds of bytes. For this reason, in Figure 2, we can see spikes of line rate increase for AARE induced implementations. These spikes are caused by binary file hits.

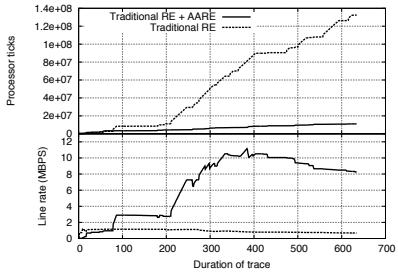


Figure 3: *Traditional RE and AARE with HTTP trace*

Real World HTTP Trace: We have also captured a short packet trace from a household access point which is used by several users. They were instructed to carry on with their regular Internet browsing, and along with that download some binary files. The result presented in Figure 3 clearly shows the improvement in both line rate as well as processor load with AARE. For a detailed analysis of processor load reduction, please refer to [Saha 2011].

4 Conclusion

Search for the optimal chunk size and needed processing power remained as the blocking factor for widespread adoption of the RE technology. Based on the widespread acceptance that the Internet is heavy tailed, and that binary files (video, audio, binary executable, compressed file) do not partially match with each other with a very high probability, we decided to address the challenge by targeting the few-in-number but large-in-volume binary files. AARE detects flows containing binary content, and treats them differently to detect a hit or miss on a very early phase. Depending on the decision, AARE then transmits just one shim header to downstream as an identifier for the whole file. This approach results in lower processing load, and higher line rate than other RE solutions.

References

ANAND, A., SEKAR, V., AND AKELLA, A. 2009. SmartRE: an architecture for coordinated network-wide redundancy elimination. In *Proc. of SIGCOMM*, ACM, 87–98.

HALEPOVIC, E., WILLIAMSON, C., AND GHADERI, M. 2010. Exploiting Non-Uniformities in Redundant Traffic Elimination. *University of Calgary*.

MUTHITACHAROEN, A., CHEN, B., AND MAZIERES, D. 2001. A low-bandwidth network file system. In *Proc. of SOSP*, ACM, 174–187.

PUCHA, H., ANDERSEN, D., AND KAMINSKY, M. 2007. Exploiting similarity for multi-source downloads using file handprints. In *Proc. of NSDI*.

SAHA, S., LUKYANENKO, A., AND YLA-JÄÄSKI, A. 2011. CombiHeader: Minimizing the Number of Shim Headers in Redundancy Elimination Systems. In *IEEE INFOCOM GI Workshop*, IEEE, 809–814.

SAHA, S. 2011. On reducing the processing load of Redundancy Elimination algorithms. In *Proc. of GLOBECOM Workshops*, IEEE, 1106–1110.

SPRING, N., AND WETHERALL, D. 2000. A protocol-independent technique for eliminating redundant network traffic. In *Proc. of SIGCOMM*, ACM, 87–95.

Enhancing Image Retrieval through Human Centered Computing Techniques

Kumaripaba M. Athukorala*
University of Helsinki

Abstract

With the explosive growth of information available in the web, locating needed and relevant information remains a difficult task. Text search engines have existed for some years now and have achieved a certain degree of success. But image retrieval systems still need to be improved to provide more accurate results. Current image retrieval techniques involve search based on user entered tags or content based image retrieval techniques. Developing accurate content based image retrieval systems is very challenging. Manual tagging is a better approach when compared to automated tagging. But explicit tagging of images had not succeeded very well. Therefore it is essential to identify explicit and implicit mechanisms for collecting metadata through human centered computing techniques. It is also import to intelligently adapt user queries to provide better results. The objective of this research is to identify mechanisms to intelligently adopt user queries and identify human computer interaction techniques for explicit/implicit metadata collection.

CR Categories: H.3.3 [Information Search and Retrieval]: Relevance feedback—Information filtering H.5.2 [User Interfaces]: Interaction styles— [I.3.6]: Methodology and Techniques—Interaction techniques;

Keywords: information retrieval, human computer interaction

1 Introduction

Retrieving images matching user queries has been studied a lot under topics such as Multimedia Information Retrieval (MIR). There had been a lot of research under content-based image retrieval. Content-based image retrieval involves retrieving images based on the image content features such as color and texture. This has been the subject of significant amount of computer vision research in the recent past [Datta et al. 2008]. Early image retrieval architectures were based on query-by-example paradigm where the best database match for user-provided query image was returned. It was then realized that the design of fully functional retrieval systems would require the support for semantic queries.

In semantic based image retrieval systems, images are annotated with semantic keywords, enabling the user to specify their query through a natural language description of the visual concepts of interest. It generated a significant interest in the problem of extracting semantic descriptions for images. Manual image labelling proves to be more accurate than automatically extracted semantic descriptions [Datta et al. 2008]. But manual image labelling is laborious and more research was conducted on finding efficient techniques. Crowd sourcing [Fei-Fei 2010], designing games with a purpose [von Ahn and Dabbish 2004] [von Ahn 2006], social tagging [Eleta and Golbeck 2012] are some findings of these research.

In recent years, tagging systems have become increasingly popular [Marlow et al. 2006]. The purpose of tagging systems is to enable users to add keywords to internet resources such as web pages, images, videos and etc. Even though user generated tags are vastly available, recent studies have shown that most users avoid annotating media such as photos [Ames and Naaman 2007]. It is found that Flickr (popular web-based photo sharing system) has succeeded in collecting a relatively high number of user generated tags by serving both personal and social purposes [Ames and Naaman 2007]. Hence it is very important to understand the incentives for annotations and methods of satisfying them.

Human Computer Interaction(HCI) involves the study, planning, and design of the interaction between people and computers. Designing of media annotation tools is an active field of research in HCI [Girgensohn et al. 2003] [Kuchinsky et al. 1999]. This research also involves identifying mechanisms to utilize both explicit and implicit user interaction strategies to collect metadata.

2 State of Art of Image Retrieval

Humans are very good in organizing things in order [Datta et al. 2008]. Over many years man learned that it is the key to progress without the loss of their current possession. Text in different languages has been set to order for efficient retrieval for over centuries.But when it comes to organizing pictures, man has traditionally outperformed machines for most tasks [Datta et al. 2008]. The main reason for this is the advanced human vision system. Building concrete descriptions of what human sees is merely an impossible task. It is even harder to teach machines to interpret things the way we see. But over the past decade impressive attempts have been made to make computers learn to interpret pictures and automatically annotate them.

Content-based image retrieval (CBIR) is the current technology which helps to organize digital picture archives by their visual content. Earliest content-based MIR systems were frequently based on computer vision algorithms which focused on feature-based similarity search [Lew et al. 2006]. The concept of similarity search was then transformed to several Internet image search engines including Webseek [Smith and Chang 1997] and Webseer [Frankel et al. 1996]. But then near the turn of the 21st century, researchers found that feature-based similarity search algorithms were not as user-friendly as expected. It was found that understanding semantics of a query is more important than understanding low-level computational features [Lew et al. 2006]. This brought in the general problem known as bridging the semantic gap. This means translating low level content-based media features to high-level terms understandable to the user. ImageScape search engine was the first content-based retrieval system to address this problem.

The recent image retrieval systems have significant limitations, such as inability to understand a wide user vocabulary and user satisfaction level. This is mainly due to the semantic gap between computers and humans [Lew et al. 2006]. Human-centered computing is one of the prevalent research topics which have potential for improving multimedia retrieval by bridging the

*e-mail:firstname.lastname@helsinki.fi

semantic gap [Lew et al. 2006]. In these designs the user can make queries using their own terminology. User studies give an insight in to interactions between human and computer.
Even though the foundational areas for MIR were often in computer-centric fields, the primary goal of MIR is to provide the user with satisfactory result. Therefore the MIR systems need to be human-centric [Lew et al. 2006]. There have been several recent initiatives in this direction such as user understanding, experiential computing and affective computing [Rodden et al. 2001]. Rodden et al had conducted a fascinating study on how the organization of images by similarity assists browsing. This study had proved that most users prefer to use text caption similarity view over the visual content view [Rodden et al. 2001]. Frohlich et al provides a good description of user requirements for photoware [Frohlich et al. 2002]. Enser et al provides a comprehensive survey of semantic gap in visual image retrieval [Enser and Sandom 2003]. Classification of images and users were proposed by Enser et al. These studies prove that full popularity of image and user types are not broadly evaluated so far [Lew et al. 2006].
Despite the considerable progress of academic research in MIR , there had been very little impact on commercial applications [Lew et al. 2006]. The target of this research is to combine HCI techniques with MIR to enhance image retrieval. The new focus of MIR should be on providing systems for user to explore media instead of searching media.In this research we try to design and analyse new interaction techniques in information retrieval.

3 Results and Work in Progress

As the first step in this research we conducted a survey on a subset of Flickr images to evaluate the user entered tags and accessibility of these images [Hietanen H. 2011] [Hietanen H. 2012]. Two procedures were followed to evaluate the CC-By licensed image set of Flickr. In the first procedure, using the words in the wordnet ontology, queries were made to the Flickr CC-By licensed image set. The goal of this experiment was to evaluate the metadata that these images contain. In the second study the images returned through semantic based image retrieval in Flickr image subset was given to users for rating. This study proved that the number of truly useful tags in these images is very low. Only 12 percent of the CC-By licensed images were easily found. Therefore this study proved the need for strong research in this area.
The next step of this research is to experiment with implicit and explicit human computer interaction techniques to infer metadata. This is an ongoing research conducted in collaboration with three teams in multidisciplinary fields. In this research we study how to integrate human interaction techniques with a scientific article retrieval system to provide better personalized results to the user. In this research we use machine learning techniques, information retrieval techniques and human computer interaction techniques to build user profiles, annotate content and design interactive information retrieval system.
This paper is structured as follows. In Section 2 we will examine the state of the art, and then section 3 discusses the open issues that are being investigated in the current research

References

AMES, M., AND NAAMAN, M. 2007. Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, New York, NY, USA, CHI '07, 971–980.

DATTA, R., JOSHI, D., LI, J., AND WANG, J. Z. 2008. Image retrieval: Ideas, influences, and trends of the new age. ACM, New York, NY, USA, vol. 40, 5:1–5:60.

ELETA, I., AND GOLBECK, J. 2012. A study of multilingual social tagging of art images: cultural bridges and diversity. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, ACM, New York, NY, USA, CSCW '12, 695–704.

ENSER, P., AND SANDOM, C. 2003. Towards a comprehensive survey of the semantic gap in visual image retrieval. In *Proceedings of the 2nd international conference on Image and video retrieval*, Springer-Verlag, Berlin, Heidelberg, CIVR'03, 291–299.

FEI-FEI, L. 2010. Imagenet: crowdsourcing, benchmarking and other cool things. In *CMU VASC Seminar*.

FRANKEL, C., SWAIN, M. J., AND ATHITSOS, V. 1996. Webseer: An image search engine for the world wide web. Tech. rep., Chicago, IL, USA.

FROHLICH, D., KUCHINSKY, A., PERING, C., DON, A., AND ARISS, S. 2002. Requirements for photoware. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, ACM, New York, NY, USA, CSCW '02, 166–175.

GIRGENSOHN, A., ADCOCK, J., COOPER, M., FOOTE, J., AND WILCOX, L. 2003. Simplifying the management of large photo collections. In *In Proc. of INTERACT03, IOS*, Press, 196–203.

HIETANEN H., ATHUKORALA K., S. A. 2011. What's with the free images? a study of flickr's creative commons attribution images. In *To be published In Proceedings of the eleventh ACM international conference on Multimedia*, MindTrek '11.

HIETANEN H., SALOVAARA A., A. K. L. Y. 2012. ¡insert image¿: Helping in the legal use of open images. In *Full paper to be presented in CHI 2012*, CHI '12.

KUCHINSKY, A., PERING, C., CREECH, M. L., FREEZE, D., SERRA, B., AND GWIZDKA, J. 1999. Fotofile: a consumer multimedia organization and retrieval system. In *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit*, ACM, New York, NY, USA, CHI '99, 496–503.

LEW, M. S., SEBE, N., DJERABA, C., AND JAIN, R. 2006. Content-based multimedia information retrieval: State of the art and challenges. ACM, New York, NY, USA, vol. 2, 1–19.

MARLOW, C., NAAMAN, M., BOYD, D., AND DAVIS, M. 2006. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proceedings of the seventeenth conference on Hypertext and hypermedia*, ACM, New York, NY, USA, HYPERTEXT '06, 31–40.

RODDEN, K., BASALAJ, W., SINCLAIR, D., AND WOOD, K. 2001. Does organisation by similarity assist image browsing? In *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, New York, NY, USA, CHI '01, 190–197.

SMITH, J. R., AND CHANG, S.-F. 1997. Visually searching the web for content. IEEE Computer Society Press, Los Alamitos, CA, USA, vol. 4, 12–20.

VON AHN, L., AND DABBISH, L. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, New York, NY, USA, CHI '04, 319–326.

VON AHN, L. 2006. Games with a purpose. In *Computer Journal*, vol. 39, 92 –94.

Evaluation Methods for Unsupervised Natural Language Learning

Sami Virpioja*
Department of Information and Computer Science, Aalto University School of Science

Abstract

Many problems in natural language processing can be approached by unsupervised machine learning methods. Their main advantages are that there is no need for manually annotated data and that the methods are language-independent. As in all machine learning, there is a need for simple and quick evaluation methods in order to evaluate the results of a particular algorithm. A problem with the evaluation of unsupervised algorithms is that there may not be a simple way to compare the output of the algorithm to the available reference data. Our work considers evaluation methods for two popular tasks for unsupervised learning: morphological analysis and learning semantically relevant vector representations. In both cases, the output of the learning algorithm is multidimensional, and we have to find the relationships between the pairs of features in the output data and the reference data.

1 Introduction

The modern approaches for natural language processing are heavily influenced by the field of machine learning. Two fundamental tasks needed in many applications are estimating the probability distribution $p(s)$ for some pieces of written language s (e.g., sentences or documents), and encoding the semantic content of documents s in a large document collection so that the documents most similar to a given document can be found quickly and accurately. In addition, there is the third problem: selecting the basic units of representation from text strings $\boldsymbol{t} \in \Sigma^*$ (Kleene closure of the character set Σ). We will call these problems statistical language modeling, representation learning, and unit selection, respectively. Given the units of representation, sentences or documents are written as a sequence $\boldsymbol{w} = (w_1, \ldots, w_n) \in \mathcal{L}^*$, where $\mathcal{L}$ is a lexicon. Thus unit selection should determine $\mathcal{L}$ and a tokenization function $\phi : \Sigma^* \mapsto \mathcal{L}^*$. Using the sequential encoding, statistical language modeling can be formulated as a supervised problem of predicting the next unit given the observed history: $p(\boldsymbol{w}) = \prod_{i=1}^{n} p(w_i | w_1 \ldots w_{i-1})$. In representation learning, the documents are often initially represented as bag-of-features $\{w_1, \ldots, w_n\}$. The latent topics underlying the observed features are discovered by unsupervised learning, either using geometric vector space models or probabilistic topic models.

The obvious approach for unit selection is to use a lexicon of words. For many languages, this is practical because the words are easily identifiable from the text, and also works well enough. However, in some languages (e.g., Chinese), the words are not separated from each other by white space, and word tokenization is non-trivial. Moreover, in morphologically rich languages (e.g., Finnish), there are a huge number of different word forms, some of them too rare to be useful in statistical models. Even for English, it is likely that units either larger or smaller than words would be useful for some applications. In contrast to using hand-crafted, rule-based analyzers or supervised learning based on manually annotated data, unsupervised unit selection does not require human labor or existing resources for the language concerned. Thus the process of unit selection (as well as language modeling and representation learning) can be independent of the language and the style and type of the text. This is a major advantage especially for small and poorly resourced languages. Unsupervised learning has been applied successfully to many related tasks, including parts-of-speech (POS) tagging [Schütze 1995], word segmentation [de Marcken 1996], and morphological analysis [Goldsmith 2001].

Developing machine learning methods for a particular task usually requires an automatic way to measure the quality of the results. Eventually, usefulness of a method should be evaluated by how it improves the performance of the target application. However, the indirect evaluations is applications are often complicated and expensive in terms of time and manual work required. This contrasts to direct evaluations, that try to measure the quality of a certain method independently of any application. Intended to be quicker and simpler than the indirect evaluations, they are important for advancing the methodological development.

2 Evaluating Unsupervised Algorithms

In the case of supervised learning, using a direct evaluation is often easy. The annotated data is divided into training and held-out sets, and after training, the output of the method is compared to the annotations within the held-out set. For example, in supervised POS tagging, the goal is to map a sequence of words to a sequence of tags. To assess the performance of a tagger, we can simply count how many of the predicted tags x_i are the same as the reference tags y_i for a set of held-out data. However, evaluation is not as straightforward for unsupervised learning. In POS tagging, we cannot tell whether a certain predicted tag x_i (say, "tag number 23") matches the reference tag y_i (say, "adjective"). More generally, we have the predicted output $\boldsymbol{x}$ and some reference data $\boldsymbol{y}$, but there is no direct way to assess whether a single output x_i is good or bad.

Considering predictions X and references Y as random variables, a theoretically solid approach is to measure the mutual information

$$I(X;Y) = E_{X,Y}\Big[\log \frac{p(x,y)}{p(x)p(y)}\Big].$$

High mutual information means that X and Y have a strong dependency, whereas $I(X;Y) = 0$ means that they are independent. Whether the mutual information can be measured in practice or not depends on the domain and the distributions of X and Y. If they are one-dimensional Gaussian variables, Pearson's correlation coefficient is directly related to mutual information. If they are one-dimensional and discrete, mutual information can be computed over the clustering that they define over data samples [Meila 2003]. The actual problems arise with multidimensional variables. In the discrete case, X and Y have several discrete labels for each sample. One example of such a multi-label classification task is morphological analysis of words. In the continuous case, X and Y are vectors in $\mathbb{R}^n$, produced, for example, by vector space models. In either case, there is no practical way to estimate the mutual information without unrealistic assumptions.

Case I: Morphological Analysis

In morphological analysis, the task is to identify the smallest meaning-bearing units of the language, morphemes. This is especially important for synthetic languages, for which the average number of morphemes per word is high. For example, the phrase "also in a cup of coffee" can be expressed as a single word "kahvikupillisessakin" in Finnish. If each morpheme had a unique

*e-mail:sami.virpioja@aalto.fi

surface realization (morph), the task would simply be the segmentation of word forms into morphs and could be evaluated by checking whether the predicted morph boundaries match to the boundaries in the reference analysis. However, generally there is no one-to-one mapping between morphemes and morphs. Even in mainly concatenative morphology, there are different surface realizations of the same morpheme (e.g., suffixes **-d** and **-ed** for English past tense) and same surface realizations of different morphemes (e.g., suffix **-s** for both plural and 3rd person singular). Thus, each morpheme has to have a unique label separate from its surface form, and the labels produced by an unsupervised algorithm can be arbitrary.

Neglecting the order of morphemes within a word, the morphological analyses for a set of words can be represented by matrix $\mathbf{X}$, where $x_{ij} = 1$ if morpheme m_i occurs in word w_j, and zero otherwise. An equivalent representation is a bipartite graph, where $x_{ij} = 1$ indicates an edge. There are two basic approaches to compare $\mathbf{X}$ to a similar matrix $\mathbf{Y}$ constructed from reference analyses for the same set of words. Co-occurrence-based methods study whether pairs of words share the same number of morphemes in $\mathbf{X}$ and $\mathbf{Y}$ [Kurimo et al. 2010]. Mathematically, they study the differences in the word-by-word matrices $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$. Assignment-based methods try find the matching between morphemes, i.e., the rows of $\mathbf{X}$ and $\mathbf{Y}$. For example, the best one-to-one assignment can be obtained by maximizing the sum of weights $w_{ij} = \mathbf{x}_{i,:}\mathbf{y}_{j,:}^{\mathrm{T}}$ over the matched pairs $(m_i, \bar{m}_j)$ [Spiegler and Monson 2010].

In [Virpioja et al. 2011], we propose new evaluation measures using both approaches. They are compared using the database collected in Morpho Challenge competitions [Kurimo et al. 2010]. The database has almost 50 different algorithms and their results in information retrieval and machine translation tasks for several languages. We identify a few evaluation methods that can be recommended due to their robustness and high correlation to the application evaluations.

Case II: Vector Space Models

A standard vector space model [Salton et al. 1975] is constructed by first taking term-document matrix $\mathbf{D}$, in which the element d_{ij} gives the weight of term i in document j. Various machine learning techniques can be used for feature selection and extraction to construct $\mathbf{D}$. As a result, we have continuous-valued document representations $\mathbf{X}$, where the similarity $\mathrm{sim}(\mathbf{x}_i, \mathbf{x}_j)$ should be high for documents that have a similar semantic content. Direct evaluation of the vector space model is seemingly impossible, as there is no "reference vectors" to which $\mathbf{x}_i$ could be compared. However, such vectors are available by using a multilingual parallel corpus: document s and its translation t can be seen as two views for the same underlying semantics. If the evaluated model captures the language independent semantic intention that is common to the aligned documents, the generated features $\mathbf{x}_i$ and $\mathbf{y}_i$ should have a high dependence across the document pairs (s_i, t_i).

Similarly to the case with morpheme labels, we cannot directly compare one dimension of $\mathbf{X}$ to another dimension of $\mathbf{Y}$: Even if $\mathbf{X}$ and $\mathbf{Y}$ are generated by the exactly the same method, the original features (terms) are different in the two languages. In [Besançon and Rajman 2002], the samples are projected into document space similarly to the co-occurrence-based methods in morphology evaluation, but their measure is unnecessarily complicated. Our novel idea is to solve the necessary "assignment" of the features with Canonical Correlation Analysis (CCA). CCA, originally proposed by [Hotelling 1936], is a linear method that finds the maximally correlated subspaces for two sets of features. In [Virpioja et al. In press], we argue that high correlations in the CCA subspaces indicate that the representations encode information regarding the meaning of the documents, not just arbitrary features of the texts. In addition, we study the CCA-based evaluation experimentally. Its results agree well with previous findings on vector space models. Moreover, it has high correlations to the results of two indirect evaluations in document matching tasks and to a quantitative manual evaluation of the factor loadings of the features found by CCA.

3 Conclusion

While unsupervised machine learning provide important tools for natural language processing, the question of how to evaluate the output of the algorithms has gained limited attention. In this paper, we have drawn together two different tasks—morphological analysis and feature generation for vector space models—in which the desired output is multidimensional. In the multidimensional case, there is no general way to measure the mutual information between predicted output X and reference output Y. Instead, we need to find the unknown relationship between the features of X and Y. For discrete output, it is possible to optimize hard assignments between the labels. For continuous output, correlated subspaces can be found by Canonical Correlation Analysis.

References

BESANÇON, R., AND RAJMAN, M. 2002. Evaluation of a vector space similarity measure in a multilingual framework. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*. 1537–1542.

DE MARCKEN, C. G. 1996. *Unsupervised Language Acquisition*. PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

GOLDSMITH, J. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics 27*, 2, 153–189.

HOTELLING, H. 1936. Relations between two sets of variates. *Biometrika 28*, 3, 321–377.

KURIMO, M., VIRPIOJA, S., TURUNEN, V., AND LAGUS, K. 2010. Morpho challenge 2005-2010: Evaluations and results. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, 87–95.

MEILA, M. 2003. Comparing clusterings by the variation of information. In *Learning Theory and Kernel Machines*, B. Schölkopf and M. K. Warmuth, Eds., vol. 2777 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 173–187.

SALTON, G., WONG, A., AND YANG, C. 1975. A vector space model for automatic indexing. *Communications of the ACM 18*, 11, 620.

SCHÜTZE, H. 1995. Distributional part-of-speech tagging. In *Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistic (EACL)*, 141–148.

SPIEGLER, S., AND MONSON, C. 2010. EMMA: A novel evaluation metric for morphological analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 1029–1037.

VIRPIOJA, S., TURUNEN, V. T., SPIEGLER, S., KOHONEN, O., AND KURIMO, M. 2011. Empirical comparison of evaluation methods for unsupervised learning of morphology. *Traitement Automatique des Langues 52*, 2, 45–90.

VIRPIOJA, S., PAUKKERI, M., TRIPATHI, A., LINDH-KNUUTILA, T., AND LAGUS, K. In press. Evaluating vector space models with canonical correlation analysis. *Natural Language Engineering*. Available on CJO 2011 doi:10.1017/S1351324911000271.

Compression-Based Similarity Measuring in Music Information Retrieval

Teppo E. Ahonen*
Department of Computer Science
University of Helsinki

Abstract

This article presents the content of an upcoming PhD thesis [Ahonen 2012c] in music information retrieval. The topic of the thesis is measuring compositional similarity between pieces of music. The similarity is based on comparing music feature representations known as chromagrams, and for the similarity metric, a data compression based methodology is applied. The performance of the suggested approach is evaluated with information retrieval and machine learning experiments.

Keywords: music information retrieval, normalized compression distance, machine learning, data mining

1 Introduction

Music information retrieval (MIR) is a novel, interdisciplinary area of research which studies how information can be extracted and retrieved from large amounts of musical data. Considering the recent rapid change in practices of how music is distributed and consumed, a growing demand for methods that are capable to perform reliable and efficient retrieval for music has arisen. Potential MIR target groups include not only music consumers, but also musicians, musicologists, and various members of music industry.

In recent years, the method of using data compression as a tool for similarity measuring has gained significant interest from researchers in various subfields of computer science. Namely, the normalized compression distance (NCD) [Cilibrasi and Vitányi 2005] has been adapted for various different similarity measuring tasks in several domains. The advantages of the metric are mostly the independence of the domain where it is applied, and the so-called quasi-universality; it minorizes every computable distance up to a compression algorithm related error.

The thesis [Ahonen 2012c] presents the work contributed to measure similarity between pieces of music using NCD and several other compression-based measures. The similarity is based on tonal feature representation called chromagram, and in order to measure similarity between chromagrams, the continuous features need to be discretized. The thesis provides insight into how features extracted from the music should be processed, how they should be represented, what should be considered when using a data compression based similarity scheme, and how well the suggested method performs with real-world music data.

2 Background

2.1 Chromagram

The chromagram, also known as the pitch class profile, is a representation of the tonal content of the music (see, e.g. [Müller 2007] for a textbook definition). Extracted from the audio signal commonly by using a short-time Fourier transform, the chromagram maps the audio frequencies to octave pitch class (that is, C, C#, D,...) bins, and then folds all pitch classes into one octave. Thus, the audio signal is turned into a sequence of 12-dimensional vectors that represent the relative energy of all 12 pitch classes of the western tonal scale.

As the chromagram contains important tonal information, it has been successfully used for various tasks in music information retrieval. Our motivation is to measure the similarity between chromagram representations using data compression, in order to determine whether two chromagrams are different versions of a same composition. This task is known as cover song identification; although it should be mentioned that a "cover" version is a bit misleading term, since the version could be a variation, a remix, a live performance, or any other rendition of the composition. Cover song identification is, unlike some other MIR tasks, an objective way to evaluate the similarity measuring, and successful identification has various potential areas of application, such as plagiarism detection. Several methods for cover version identification have been presented (see, e.g. [Serrà et al. 2010] for a survey of most well-known methods), but applying NCD or other compression-based method for this task has been practically nonexistent, making our work highly novel.

2.2 Similarity measuring

The idea of using data compression is based on the the idea of using Kolmogorov complexity to measure similarity; a distance metric based on Kolmogorov complexity called normalized information distance (NID) can be shown to be a parameter-free, universal similarity metric [Cilibrasi and Vitányi 2005], making it a highly useful distance metric for any application. However, as Kolmogorov complexity is non-computable, the universal similarity metric is only a theoretical concept. But Kolmogorov complexity can be approximated by using a standard lossless data compression algorithm, resulting in a practical version of the normalized information distance called normalized compression distance (NCD). For a pair of strings, x and y, NCD is denoted [Cilibrasi and Vitányi 2005]

$$NCD(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}, \qquad (1)$$

where $C(x)$ is the length of the string x when compressed using data compression algorithm C, and xy is the concatenation of x and y. It can be shown that NCD is a metric, but it is not universal. However, the authors of [Cilibrasi and Vitányi 2005] denote it as quasi-universal. Also, like NID, NCD requires no background knowledge of the domain where it is applied, and it is also parameter-free.

3 Contributions

3.1 Chromagram features and representations

The continuous nature of the data of the chromagram bins is a challenge for the compression algorithms. First, continuous data do not compress very well. Second, small differences in continuous values might appear larger in practice, depending on how the continuous values are encoded. Because of these reasons, discretization of chromagram values is needed. Because NCD requires no background knowledge of the domain, the representation of the data is

*e-mail:teahonen@cs.helsinki.fi

crucial, as poorly chosen representation is likely to lead to unusable results. Also, the selected representation should preseve the essential tonal information and remain suitable for the compression algorithm at the same time.

We have experimented with several methods for chromagram data quantization. A method presented in [Ahonen and Lemström 2008; Ahonen 2009] utilizes a hidden Markov model based chord approximation method to turn the chromagram into a sequence of characters from an alphabet of 24, each representing a major or a minor triad chord. In [Ahonen et al. 2011], the 12-dimensional data was reduced to 6-dimensional tonal centroid representation and binarized, producing a more fine-grained sequence of an alphabet of 2^6 characters. Recently, in [Ahonen 2012b; Ahonen 2012d] we experimented with a 12-character quantization based on transposition indices between each chromagram frame and a global chromagram profile.

3.1.1 Feature combination

As it seems that the single quantized chromagram representation might not capture all relevant information, a combination of different representations would seem like a plausible idea. We have studied this in [Ahonen 2010a; Ahonen 2012a], combining some of the previously-mentioned features and several others, such as dynamic chromagram features. Results indicate that the feature combination does indeed provide a somewhat higher identification accuracy, but the approach comes with a trade-off: calculating several features is naturally more time-expensive, and the amount of crucial feature parameters grows vast.

3.2 Analysis on compression-based measuring

Several studies on the universal empirical performance of NCD have been presented, but in our work we are mostly interested how well compression-based measuring can be applied for the task of estimating chromagram similarity. To provide insight into this, we will evaluate several compression algorithms, and suggest various quantized representations for the data. The pros and cons of compression-based similarity measuring for the task in the hand should come out explicit.

3.3 Machine learning evaluation

In addition to the retrieval performance evaluation, we also utilize the compression-based methodology for several tasks in supervised and unsupervised machine learning. All evaluations are run on large sets of real-world audio data.

3.3.1 Classification

Using NCD as the similarity metric for classification tasks is not a completely novel method, and results in different tasks also favor using NCD as the distance metric. We examine the NCD-based k-nearest neighbor classification for chromagram data. As a more novel approach, we have experimented with extending the NCD from pairwise similarity measuring to measuring similarity in sets of objects [Ahonen 2010b].

3.3.2 Clustering

Several clustering methods that apply compression-based distance measuring have been suggested throughout the years. We bring some novelty to the approach by using a modified version of the k-medians algorithm, where we replace the commonly used Euclidean distance with compression distance [Ahonen et al. 2011; Ahonen 2012b]. As with the classification, we have also experimented with list-based compression methods for clustering tasks [Ahonen 2012d].

4 Conclusions

Measuring similarity using data compression has been applied for various tasks in music information retrieval. Here, we have presented research where the method has been utilized to measuring the similarity between the tonal contents of pieces of music. The method has been evaluated in several machine learning experiments, using large amounts of real-world music data.

Acknowledgements

The PhD studies of Ahonen have been supported by various instances. From August 2008 to the end of 2010, the work was supported by Academy of Finland, grant #129909. From 2011 onwards, the work has been supported by Helsinki Doctoral Programme in Computer Science – Advanced Computing and Intelligent Systems (Hecse).

References

AHONEN, T. E., AND LEMSTRÖM, K. 2008. Identifying cover songs using normalized compression distance. In *MML'08*.

AHONEN, T. E., LEMSTRÖM, K., AND LINKOLA, S. 2011. Compression-based similarity measures in symbolic, polyphonic music. In *ISMIR'11*.

AHONEN, T. E. 2009. Measuring harmonic similarity using PPM-based compression distance. In *WEMIS'09*.

AHONEN, T. E. 2010. Combining chroma features for cover version identification. In *ISMIR'10*.

AHONEN, T. E. 2010. Compressing lists for audio classification. In *MML'10*.

AHONEN, T. E., 2012. Chroma feature combination revisited (tentative title). Manuscript in preparation.

AHONEN, T. E. 2012. Compression-based clustering of chromgram data: New method and representations. In *CMMR'12*. To appear.

AHONEN, T. E. 2012. *Compression-based Similarity Measuring for Practical Applications in Content-based Music Information Retrieval*. PhD thesis, University of Helsinki. To appear.

AHONEN, T. E., 2012. On clustering chromagram data via list-based compression distance. Submitted.

CILIBRASI, R., AND VITÁNYI, P. M. 2005. Clustering by compression. *IEEE Trans. Inf. Theory 51*, 4 (April).

MÜLLER, M. 2007. *Information Retrieval for Music and Motion*. Springer Verlag.

SERRÀ, J., GÓMEZ, E., AND HERRARA, P. 2010. *Audio Cover Song Identification and Similarity: Background, Approaches, Evaluation, and Beyond*, vol. 274. Springer-Verlab Berlin / Heidelberg, ch. 14, 307–332.

Stacking Clouds

Toni Ruottu, Eemil Lagerspetz and Sasu Tarkoma
University of Helsinki *

Abstract

Service model stack diagrams are sometimes presented when cloud systems are discussed. These diagrams are based on service models defined by the National Institute of Standards and Technology and inspired by the desire to build cloud applications out of interchangeable parts. We consider the possibility of the three service models forming a stack. Placing existing software in the model or defining implementations for the model is hard because it is so general. We introduce terms *solution* and *component* to pinpoint partial layer implementations and pieces of software within the layers. We take into account the need to define interfaces in an iterative manner as we learn about the requirements.

Keywords: cloud, software, definition, layer model

1 Introduction

Cloud computing refers to a system distribution model where some computer resources are provided by a third party for use over the network [Armbrust et al. 2010; Vaquero et al. 2008]. The US based National Institute of Standards and Technology (NIST) defines three service models for cloud computing [Mell and Grance 2009]. The service models are used to categorize service providers based on the abstraction level of the provided service. Infrastructure as a service (IaaS) providers abstract physical hardware from their customers and provide virtual machines. Platform as a service (PaaS) providers abstract multiple computing nodes of the infrastructure into a single application development platform. Software as a service (SaaS) providers run services that are useful to the end users, mapping computations into meaningful things such as an email or a photo album.

Technology stacks are motivated by the success of the TCP/IP stack which is based on the more abstract OSI model that defines abstraction layers for data communication systems. In a stack model the user uses services provided by the top layer, and the top layer uses services provided by the lower layers. The layered stack representation implies that implementations of different layers are interchangeable. The borders between layers symbolize interfaces that implementations of the different layers use to communicate with each other. The goal of the model is to define these interfaces to allow competition on providing implementations for different layers.

We consider the possibility of the three service models forming a stack. A service of any of the three service models has its limitations. None of them can support all use cases or applications. Therefore, for example PaaS services are often not interchangeable. In practice the platform's type is defined by the kind of applications it supports. Thus the service model stack must be refined to accommodate multiple possibly overlapping elements on each layer. We introduce some new terminology to clarify the model.

2 Service Model Stack

Service model stack diagrams are sometimes presented when cloud systems are discussed [Lenk et al. 2009; Rimal et al. 2009; Bhardwaj et al. 2010]. We have included a picture of a service model stack as Figure 1. While the service model definition helps us categorize different types of cloud services, a stack model may help us visualize how some service models are more abstract than others. The stack model also implies that the upper layers are implemented on top of the lower layers. An SaaS email application may be implemented on top of a PaaS offering, while the PaaS provider may use services from an IaaS partner to acquire resources based on the timely need.

Figure 1: *The Service Model Stack.*

Placing existing software in the service model stack is hard because the model is so general. Let us consider a database system for example. It is possible to provide a database as an SaaS service by providing raw database access directly to the user. A database system can provide some support for defining application specific user interfaces to the user. This would turn it into a PaaS offering. Finally a database system often provides long term storage which implies it is part of the storage infrastructure, making it an IaaS service. A database system might be very limited either in its user friendliness, its support for custom applications or its support for complex computations. Thus a database system might not alone fulfil the requirements for a layer in the service model stack.

Defining what an implementation of a layer in the service model stack would look like seems impossible. An implementation of the SaaS layer needs to contain all possible software applications needed by humankind. The PaaS layer needs to support building all of those applications, and the IaaS layer needs to have the correct abstraction for supporting all the applications and platforms. If we can not define what the implemented layers should do we can not define how the interfaces between those layers would look like.

It is impossible to define interfaces for all future needs. Thus designing cloud interfaces needs to be an incremental process. We are seeing some competition between *Open Cloud Computing Interface* (OCCI) and *Elastic Compute Cloud* (EC2) API. In addition to the competition in implementing the interfaces it is also useful to have some competition in defining those interfaces. Some competition currently exists between *Eucalyptus*, *OpenStack* and *Amazon Web Services* (AWS) on the IaaS layer, but also between *Google App Engine*, *Heroku* and *Cloudfoundry* on the PaaS layer. To keep the competition alive we need to allow incomplete, scenario-specific interfaces to be defined and implemented.

3 Solution Stacks

We suggest that term *solution* be used when referring to competing implementations within a layer. Solutions need not be full implementations of the service model layer they target. Instead solutions may produce the cloud services needed for a specific use case.

*e-mail: {toni.ruottu, eemil.lagerspetz, sasu.tarkoma}@cs.helsinki.fi

Two SaaS solutions, say *Facebook* and *Google Docs*, may implement completely orthogonal parts of the same layer. This does not stop two IaaS solution providers behind OpenStack and Eucalyptus from working together on their shared functionalities. Depending on timely motivations the two service provides may either compete in defining a better interface or strive to standardize known good parts of the interface.

Standardizing solution interfaces allows other solutions on the layer above to be built without knowing internal structure of the services provided at the lower layer. Ideally such standards would make it possible to move from one service provider to another. While solutions may overlap in functionality and interfaces, their non-functional qualities can still be different. For example two storage services could have different reliability guarantees. Combining this with common interfaces allows competition between solutions providing similar services.

We suggest that term *component* be used for referring to a piece of software used to provide a service. Components need not be designed for one layer specifically. Thus the database system from our example in the previous chapter could be used for implementing solutions on multiple layers. In practice it may be desirable to further limit the scope when designing new software components.

When we talk about components and solutions we are talking about the technical parts that are used by a service provider to provide a service. For a piece of software to become part of a service, someone needs to run it. For example AWS is a live service and is not considered a solution in this sense while the software used to run AWS definitely counts as a solution. While the software for a solution may be publicly available, all components and solutions need not be published. Some companies may use in-house developed components in their solutions. This classification should not be affected by such details.

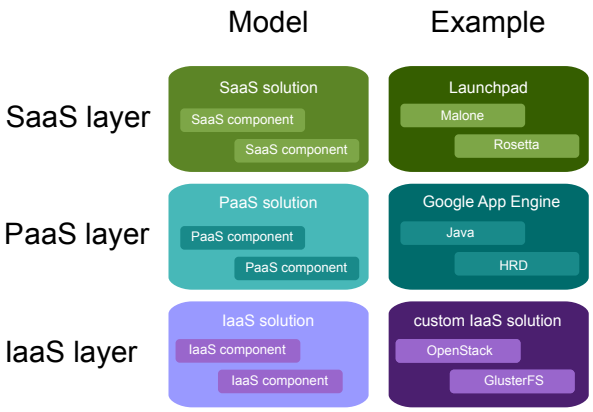


Figure 2: *Component/solution model with practical examples.*

Figure 2 visualizes how solutions and components fit into the service model stack. The figure shows the theoretical stack model next to some examples of solutions and components on different layers. On the top there is a software project management solution called *Launchpad*. It is used by software developers for bug tracking, software translation, and release planning. Malone is a bug tracking application while Rosetta is used for translating software user interfaces into foreign languages. Google App Engine is a web application platform. Its applications are written in the java programming language and it uses High Replication Datastore (HRD) for long term storage. Our example IaaS solution is a combination of OpenStack and *GlusterFS*. OpenStack is used to manage virtual machines while GlusterFS is used for longterm storage. The components within each solution in the figure are examples and do not compose the whole solution. For example Google App Engine uses Python for some applications and Launchpad requires some combination of Linux, Apache, MySQL and Perl/Php/Python (LAMP) components to run.

4 Conclusion

We studied how a cloud service stack differs from traditional networking stacks with its vague interfaces. We noticed some problems with the service model stack being too general for discussing cloud interfaces. We suggested the term *solution* to be used for the compilation of software required for providing a service and the term *component* to be used for pieces of software in such compilations.

Building cloud services from stackable parts would seem to require standardization for interfaces between the layers in the stack. At the moment it seems too early to say whether or not clouds will eventually become truly stackable in a way that would allow migration between providers of lower levels. This may not only be a question of interfaces, but also about other types of co-operation between service providers.

Acknowledgements

The authors would like to thank Ville Palkosaari, Sami Saada, Adam J. Oliner and Flutra Osmani for constructive feedback on the original draft.

References

ARMBRUST, M., FOX, A., GRIFFITH, R., JOSEPH, A. D., KATZ, R., KONWINSKI, A., LEE, G., PATTERSON, D., RABKIN, A., STOICA, I., AND ZAHARIA, M. 2010. A view of cloud computing. *Communications of the ACM 53*, 4, 50–58.

BHARDWAJ, S., JAIN, L., AND JAIN, S. 2010. Cloud computing: A study of infrastructure as a service (iaas). *International Journal of Engineering and Information Technology 2*, 1, 60–63.

LENK, A., KLEMS, M., NIMIS, J., TAI, S., AND SANDHOLM, T. 2009. What's inside the cloud? an architectural map of the cloud landscape. In *Proceedings of the 2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing*, IEEE Computer Society, Washington, DC, USA, CLOUD '09, 23–31.

MELL, P., AND GRANCE, T., 2009. The nist definition of cloud computing. Special Publication 800-145, The National Institute of Standards and Technology (NIST).

RIMAL, B., CHOI, E., AND LUMB, I. 2009. A taxonomy and survey of cloud computing systems. In *INC, IMS and IDC, 2009. NCM '09. Fifth International Joint Conference on*, 44–51.

VAQUERO, L. M., RODERO-MERINO, L., CACERES, J., AND LINDNER, M. 2008. A break in the clouds: Towards a cloud definition. *SIGCOMM Comput. Commun. Rev. 39*, 1 (December), 50–55.

Helsingin tietojenkäsittelytieteen ja -tekniikan tutkijakoulun posterit

Posters of the Helsinki Graduate School in Computer Science and Engineering

P1: Card Games for Teaching Data Structures and Algorithms

Lasse Hakulinen
Aalto University / Department of Computer Science and Engineering
lasse.hakulinen@aalto.fi

Data structures and algorithms have many elements and rules that can be used to design educational games. Many of the games used in education are played with a computer. However, cards games can also be used to teach students algorithms and to encourage them to discuss the issues faced during the games. This poster presents two card games that were used on a Data Structures and Algorithms course in Aalto University. They deal with sorting algorithms and concepts related to them. The fundamental idea of the games is to raise questions that the players need to find answers to, rather than to provide direct answers to the players.

P2: Noisy-OR Models with Latent Confounding

Antti Hyttinen
University of Helsinki / Department of Computer Science
antti.hyttinen@helsinki.fi

Generally causal models with latent variables are not identifiable from passive observational data or from experimental data in which only a few variables are subject to interventions at a time. The poster shows that if the local CPDs of the model are restricted to follow the noisy-OR parameterization, we can identify the causal model for example from passive observational data and experiments intervening only on a single variable at a time. Learning methods include a constraint-based method and maximizing the likelihood with the EM-algorithm.

P3: Backward Model Selection in Finite Mixture Models

Prem Raj Adhikari
Aalto University School of Science, Department of Information and Computer Science
prem.adhikari@aalto.fi

One of the prerequisites of using mixture models is the knowledge of number of mixture components a priori so that Expectation Maximization (EM) algorithm can be used to learn the maximum likelihood parameters of the mixture model. However, number of mixing components is often unknown and for over a century of using mixture models, determining the number of mixture components has been very central problem in mixture modelling. In this paper, we propose a search-based backward model selection method for mixture models

using progressive merging of mixture components. The paper also proposes a data driven, fast approximation of Kullback-Leibler (KL) divergence as a criterion to merge the mixture components. Furthermore, merging of mixture components is expected to eliminate the problem of local optima encountered in Expectation Maximization (EM) algorithm as we select the best model from different training runs. The proposed methodology is used in mixture modelling of two chromosomal aberration datasets showing that model selection is efficient and effective.

P4: Statistical Test for Consistent Estimation of Causal Effects in Linear non-Gaussian Models

Doris Entner
University of Helsinki, Department of Computer Science
entner@cs.helsinki.fi

In many fields of science researchers are faced with the problem of estimating causal effects from non-experimental data. A key issue is to avoid inconsistent estimators due to confounding, a problem commonly solved by 'adjusting for' a subset of the observed variables. When the data generating process can be represented by a directed acyclic graph, and this graph structure is known, there exist simple graphical procedures for determining which subset of covariates should be adjusted for to obtain consistent estimators of the causal effects. However, when the graph is not known no general and complete procedures for this task are available. In this poster we introduce such a method for linear non-Gaussian models, requiring only partial knowledge about the temporal ordering of the variables: We provide a simple statistical test for inferring whether an estimator of a causal effect is consistent when controlling for a subset of measured covariates, and we present heuristics to search for such a set.

P5: Efficient Gaussian Process Inference for Short-Scale Spatio-Temporal Modeling

Jaakko Luttinen
Aalto University / Department of Information and Computer Science
jaakko.luttinen@aalto.fi

This paper presents an efficient Gaussian process inference scheme for modeling short-scale phenomena in spatio-temporal datasets. Our model uses a sum of separable, compactly supported covariance functions, which yields a full covariance matrix represented in terms of small sparse matrices operating either on the spatial or temporal domain. The proposed inference procedure is based on Gibbs sampling, in which samples from the conditional distribution of the latent function values are obtained by applying a simple linear transformation to samples drawn from the joint distribution of the function values and the observations. We make use of the proposed model structure and the conjugate gradient method to compute the required transformation. In the experimental part, the proposed algorithm is compared to the standard approach using the sparse Cholesky decomposition and it is shown to be much faster and computationally feasible for 100-1000 times larger datasets. We demonstrate the advantages of the proposed method in the problem of reconstructing sea surface temperature, which requires processing of a real-world dataset with 106 observations.

P6: Damage Detection Methods for Structural Health Monitoring with Wireless Sensor Networks

Janne Toivola
Aalto University / Department of Information and Computer Science
janne.toivola@aalto.fi

Detecting changes in the condition of large structures, like bridges, offers a challenging data analysis task, as there are no practical sensors that would directly indicate damages and the potential future damages may have unpredictable effect on the measurements. Thus, we need to extract indirect features, insensitive to environmental variability, and novelty detection methods to detect possible unforeseen changes. This work considers the following data processing chain: feature extraction from low-power wireless accelerometer sensors, centralized and distributed feature space dimensionality reduction methods, and the final damage detection in the novelty detection framework.

P7: PicSOM Experiments in TRECVID 2011

Mats Sjöberg
Aalto University / Department of Information and Computer Science (ICS)
mats.sjoberg@aalto.fi

The poster presents an overview of our semantic indexing and automatic known-item search task experiments in the TRECVID 2011 video retrieval evaluation organized by NIST.

P8: Random Projection Method for Scalable Malware Classification

Jozsef Hegedus
Aalto University, Information and Computer Science Department
jhegedus42@gmail.com

In this poster, a two-stage methodology to analyze and detect behavioral-based malware is presented. In the first stage, a random projection (RP) is decreasing the variable dimensionality of the problem and is simultaneously reducing the computational time of the classification task by two orders of magnitude. In the second stage, a modified K-Nearest Neighbors classifier is used. This methodology is applied to a large number of file samples provided by F-Secure Corporation, for which a dynamic feature has been extracted during DeepGuard sandbox execution. We compare the speed and accuracy of two modified k-NN classifiers; 1) using approximated or 2) exact cosine similarity measures.

P9: Random Graph Ensemble in Multi-Task Classification

Hongyu Su
ICS, Aalto University
hongyu.su@aalto.fi

We present an ensemble of multi-task classifiers for multilabel classification. As the base classifiers of ensemble, we use Maximum Margin Conditional Random Field (MMCRF) Model. Source diversity of base classifiers arises from the different random output structures, a different approach from boosting or bagging. Experimental result shows that ensembles of random networks outperform other approaches.

P10: Analyzing Parliamentary Elections Based on Voting Advice Application Data

Jaakko Talonen
Aalto / Department of Information and Computer Science
jaakko.talonen@aalto.fi

The values of Finnish citizens and the members of the parliament are modelled. To achieve this goal, two databases are combined: voting advice application data and the results of the parliamentary elections in 2011. First, the data is converted to a high-dimension space. Then, it is projected to two principal components. The projection allows us to visualize the main differences between the parties. The value grids are produced with a kernel density estimation method without explicitly using the questions of the voting advice application.

P11: Conflict-Driven XOR-Clause Learning

Tero Laitinen
Aalto University / Department of Information and Computer Science
tero.laitinen@aalto.fi

Modern conflict-driven clause learning (CDCL) SAT solvers are very good in solving conjunctive normal form (CNF) formulas. However, some application problems involve lots of parity (xor) constraints which are not necessarily efficiently handled if translated into CNF. This paper studies solving CNF formulas augmented with xor-clauses in the DPLL(XOR) framework where a CDCL SAT solver is coupled with a separate xor-reasoning module. New techniques for analyzing xor-reasoning derivations are developed, allowing one to obtain smaller CNF clausal explanations for xor-implied literals and also to derive and learn new xor-clauses. It is proven that these new techniques allow very short unsatisfiability proofs for some formulas whose CNF translations do not have polynomial size resolution proofs, even when a very simple xor-reasoning module capable only of unit propagation is applied. The efficiency of the proposed techniques is evaluated on a set of challenging logical cryptanalysis instances.

P12: Emergence of Representation from Natural Data

Jaakko Väyrynen
Aalto University, Department of Information and Computer Science
jaakko.j.vayrynen@aalto.fi

This PhD work studies how unsupervised methods represent natural visual and textual data. The focus is on how statistical concepts beyond correlation learn structure that is cognitively meaningful and can be interpreted. Digital images can be directly encoded as vectors of pixels values. Learning based on sparseness, temporal coherence and topographic organization from small image patches and their sequences results in spatial and spatio-temporal receptive fields similar to simple-cells in the visual cortex. The learned receptive fields show characteristics of edge and line detectors with different orientations and scale. Textual elements, such as words and documents, can be statistically represented as frequencies of contextual units with the bag-of-units approach from a text corpus, after which distances in the created vector space measure semantic relatedness between the elements. Independent component analysis finds a sparse representation for a word vector space, in which the emergent features separate part-of-speech categories and correlate with semantic word category norms.

P13: SMT-based Induction Methods for Timed Systems

Roland Kindermann
Aalto University, Dept. of Information and Computer Science
roland.kindermann@aalto.fi

Verification techniques determine whether a given system (hardware, software, etc.) satisfies a given specification. Unlike testing, verification techniques are guaranteed to find bugs should they exist. Induction methods are a family of verification methods that verifies systems by generating an inductive proof for their correctness. In my recent research, I have been extending induction methods from finite state systems to timed systems, i.e., systems that can measure time using real-valued clock variables and thus have an infinite number of states. Key to this extension is the use of Satisfiability Modulo Theories (SMT) solvers.

P14: Probabilistic Proactive Timeline Browser

Antti Ajanki
Aalto University School of Science, Department of Information and Computer Science
antti.ajanki@aalto.fi

We have developed a browser suitable for finding events from timelines, in particular from life logs and other timelines containing a familiar narrative. The system infers the relevance of events based on the user's browsing behavior and increases the visual saliency of relevant items along the timeline. As recognized images are strong memory cues, the user can quickly determine if the salient images are relevant and, if they are, it is quick and easy to select them by clicking since they are salient. Even if the inferred relevance was not correct, the timeline will help: The user may remember if the sought event was before or after a saliently shown event which limits the search space. A user study shows that the browser helps in locating

relevant images quicker, and augmenting explicit click feedback with implicit mouse movement patterns further improves the performance.

P15: Automated Testing of Multithreaded Programs with Unfoldings

Kari Kähkönen
Aalto University / Department of Information and Computer Science
kari.kahkonen@aalto.fi

Both input data from environment and the nondeterministic interleavings of concurrent events can affect the behavior of multithreaded programs. One approach to systematically explore the nondeterminism caused by input data is dynamic symbolic execution. For testing multithreaded programs we present a new approach that combines dynamic symbolic execution with unfoldings, a method originally developed for Petri nets but also applied to many other models of concurrency. The new algorithm can explore the reachable control states of each thread with a significantly smaller number of test runs when compared to existing algorithms combining dynamic symbolic execution and partial-order reductions. In some cases the reduction to the number of test runs can be even exponential.

P16: Mask Estimation and Sparse Imputation for Missing Data Speech Recognition in Multisource Reverberant Environments

Heikki Kallasjoki
Aalto University School of Science / Department of Information and Computer Science
heikki.kallasjoki@aalto.fi

This work presents an automatic speech recognition system, which uses a missing data approach to compensate for environmental noise. The missing, noise-corrupted components are identified using binaural features or a support vector machine (SVM) classifier. To perform speech recognition using the partially observed data, the missing components are substituted with clean speech estimates calculated using sparse imputation. Evaluated on the CHiME reverberant multisource environment corpus, the missing data approach significantly improves the keyword recognition accuracy in moderate and poor SNR conditions.

P17: Communication Applications of Mobile Audio-Augmented Reality

Robert Albrecht
Aalto University, Department of Media Technology
robert.albrecht@aalto.fi

Audio-augmented reality could provide the means for applications where communication channels are integrated into the acoustic environment of the user. For example, the voices of remote participants in a conference could be heard as if they sat among the local participants. Smart and seamless augmented reality communications require, among other things, techniques for identifying if two persons are in the same acoustic space, and for integrating virtual sound sources into the local acoustic environment.

P18: Partial Order MCMC for Structure Discovery in Bayesian Networks

Teppo Niinimäki
University of Helsinki / Department of Computer Science
teppo.niinimaki@helsinki.fi

We present a new Markov chain Monte Carlo method for estimating posterior probabilities of structural features in Bayesian networks. The method samples partial orders on the nodes; for each sample, the conditional probabilities of interest are computed exactly. Compared to previous methods our algorithm obtains a significant reduction in the size of sample space with negligible increase in computation time.

P19: Environmental Proxy Selection Problems in Temperature Reconstruction

Mikko Korpela
Aalto University / Department of Information and Computer Science
mikko.korpela@aalto.fi

Direct temperature measurements are only available from the past few hundred years. Therefore, proxy measurements must be used. We study the use of different environmental proxy variables for temperature reconstruction. Differences in both the time coverage of the proxies and the temperature signal present in them pose a challenge to the recovery of reliable temperature records.

P20: Model Checking Asynchronous Control Systems

Tuomas Launiainen
Aalto School of Science, Department of information and computer science
tuomas.launiainen@aalto.fi

The current trend in industrial automation, e.g. nuclear safety systems, is to shift responsibilities from human operators to computerised control systems. For this, much effort is required to verify that the automation system does not malfunction. These systems often exhibit asynchronous communication between components, making their analysis difficult. This research is focused on methods for analysing them with model checkers, i.e. automated verification tools that exhaustively explore the behaviours of systems.

Modelling time as a real value is very useful in the analysis of asynchronous systems. Some model checkers support this, but the ones that do represent the state of the model explicitly, i.e. each distinct state is stored in a distinct location in memory. This limits the size of the models that can be analysed: it is impractical to verify models that do not fit into memory. Symbolic model checking is a method that combats the problem by storing compact representations of state sets instead of single states. None of the current model checkers with real valued clocks support symbolic model checking, however. Fitting these two approaches together is also our goal.

P21: Identifying Regulatory Modules in Genome

Jarkko Toivonen
Department of Computer Science, University of Helsinki
jarkko.toivonen@cs.helsinki.fi

Transcription factors control which genes are expressed, that is, which genes are used to produce proteins. Transcription factors play a major part in cell differentation and in controlling cell's functioning. When a transcription factor binds to DNA close to some gene it can control how the gene is transcribed and later translated to proteins. To understand this control mechanism we need a way to describe and predict the sites in DNA where the factors prefer to bind. In addition to locating individual binding sites of transcription factors, we are also interested in understanding the way the factors cooperate together, and how to measure the effect this cooperativity has on the strength of the binding. Knowing the transcription factor mechanism is also important in understanding the causes for cancer. We have concentrated on analysing the data from the high-throughput SELEX method.

P22: A Symbolic Model Checking Approach to Verifying Satellite Onboard Software

Xiang Gan
Department of Information and Computer Science, Aalto University
xiang.gan@aalto.fi

The use of symbolic model checking technology to verify the design of an embedded satellite software control system called attitude and orbit control system (AOCS) is discussed. An executable AOCS implementation by Space Systems Finland has been provided to us in Ada source code form. In order to use symbolic model checking methods, the Ada implementation of the system was modeled at a quite detailed implementation level using the input language of the symbolic model checker NuSMV 2. We describe the modeling techniques and abstractions used to alleviate the state explosion problem due to handling of timers and the large number of system components controlled by AOCS. The specification of the required system behaviour was also provided to us in a form of extended state machine diagrams with prioritized transitions. These diagrams have been translated to a set of temporal logic properties, allowing the piecewise checking of the system behaviour one extended state machine transition at a time. The generated properties are in general liveness properties. We describe how the liveness checking is translated to safety checking, since safety-checking algorithms are simpler and more efficient than liveness checking. In this way, existing safety checking techniques can also be used for liveness checking.

P23: Enabling Continuous Real-time Bodily Interaction with Virtual Characters

Klaus Förger
Aalto University, Department of Media Technology
klaus.forger@aalto.fi

Interaction with virtual characters is common in modern games and other interactive animations. However, much of the interaction is task-oriented and ignores the messages that humans continuously send with the style of their motions and bodily postures. The continuous bodily interaction is important when creating virtual characters are intended to

appear natural and emotional. This poster presents three main challenges in enabling such interaction in real-time. The first challenge is related to detecting meaningful features from human skeletal motion. The next problem is defining how a virtual character should react to observed human motions. Lastly, the desired motion style of the virtual character should be taken into account in motion synthesis. If these challenges are solved, it is possible to create a continuous interaction loop using bodily motions between a human and a virtual character.

P24: Material Appearance Capture Using Frequency Measurements

Miika Aittala
Aalto University / Department of Media Technology
miika.aittala@aalto.fi

Realistic appearance of surface materials is a key aspect in photorealistic rendering in computer graphics. The ability to easily capture reflectance properties from real world surfaces would enable rapid content creation for applications such as games, film and virtual and augmented reality. We propose a capture setup that allows one to infer such properties by photographing the response of a surface to sinusoidal illumination patterns of various frequencies.

P25: Data Aggregation: Balancing Delay and Communication Costs

Lauri Ahlroth
Aalto-yliopisto, ICS
lauri.ahlroth@aalto.fi

We present the problem of online data aggregation, which models buffered packet sending over a network with minimum sum of transmission and delay costs. We give approximation algorithms that extend best-known results from trees to more general network classes.

P26: DataFinland - A Semantic Portal for Open and Linked Datasets

Matias Frosterus
Aalto University School of Science / Department of Media Technology
matias.frosterus@aalto.fi

The number of open datasets available on the web is increasing rapidly with the rise of the Linked Open Data (LOD) cloud and various governmental efforts for releasing public data in different formats, not only in RDF. The aim in releasing open datasets is for developers to use them in innovative applications, but the datasets need to be found first and metadata available is often minimal, heterogeneous, and distributed making the search for the right dataset often problematic. To address the problem, we present DataFinland, a semantic portal featuring a distributed content creation model and tools for annotating and publishing metadata about LOD and non-RDF datasets on the web. The metadata schema for DataFinland is based on a modified version of the voiD vocabulary for describing linked RDF datasets, and annotations are done using an online metadata editor SAHA connected to ONKI ontology services providing a controlled set of annotation concepts. The content is published instantly on an integrated faceted search and browsing engine HAKO for human users, and as a SPARQL

endpoint and a source file for machines. As a proof of concept, the system has been applied to LOD and Finnish governmental datasets.

P27: Ensemble Computation with OR- and SUM-circuits

Janne H. Korhonen
University of Helsinki / Department of Computer Science + HIIT
janne.h.korhonen@helsinki.fi

Given a Boolean function as input, a fundamental problem is to find a Boolean circuit with the least number of elementary gates (AND, OR, NOT) that computes the function. The problem generalises naturally to the setting of multiple Boolean functions: find the smallest Boolean circuit that computes all the functions simultaneously. We study an NP-complete variant of this problem titled Ensemble Computation under two monotone circuit classes: OR-circuits and SUM-circuits. In particular, we are interested in understanding the separation between these classes. The main motivation for this work is the relationship between the problem of rewriting in subquadratic time a given OR-circuit to a SUM-circuit and the existence of non-trivial algorithms for NP-hard problems, e.g. CNF-SAT. We also present computational results on the sizes of small OR- and SUM-circuits.

P28: Distinguishing Between Major and Minor Chords in Automatic Chord Transcription

Antti Laaksonen
University of Helsinki, Department of Computer Science
ahslaaks@cs.helsinki.fi

Automatic chord transcription is a problem of extracting the harmonic content from a music signal and representing it through chord symbols. We focus on distinguishing between major and minor chords in automatic chord transcription. We are especially interested in the role of the musical context in this process. We conduct an experiment where human listeners are asked to classify chords, which a computer transcriber has failed to recognize when evaluated using a collection of Beatles songs. Based on this experiment and our analysis, we conclude that the musical context is often needed in distinguishing between major and minor chords. Furthermore, sometimes the quality of a chord cannot be unambiguously determined even if the full musical context is available.

P29: Do Biological Names Mean anything? Managing the Underlying Meanings on the Semantic Web

Jouni Tuominen
Aalto University School of Science, Department of Media Technology
jouni.tuominen@aalto.fi

Periodic changes characterize the scientific naming system. As a result, the biggest challenge lies in ascertaining the actual meaning of names, when multiple taxonomic concepts are associated with them. This makes it hard to integrate biological information from different sources, such as publications, online databases, and museum collections, and search for it. On the Semantic Web, the problem can be approached by representing taxa, checklists, and their

relations as ontologies that are decipherable for machines. Our goal is to establish a centralized ontology repository of biological names and classifications in Finland.

P30: Multi-Pattern Matching with Bidirectional Indexes

Kalle Karhu
Aalto University, Department of Computer Science and Engineering
kalle.karhu@aalto.fi

We study multi-pattern matching in a scenario where the pattern set is to be matched to several texts and hence indexing the pattern set is affordable. This kind of scenarios arise e.g. in metagenomics, where pattern set represents DNA of several species and the goal is to find out which species are represented in the sample and in which quantity. We develop a generic search method that exploits bidirectional indexes both for the pattern set and texts, and analyze the best and worst case running time of the method on worst-case text. We show that finding the instance of the search method with minimum best case running time on worst case text is NP-hard. The positive result is that an instance with logarithm-factor approximation to minimum best case running time can be found in polynomial time using a bidirectional index called affix tree. We further show that affix trees can be simulated space-efficiently using bidirectional variant of compressed suffix trees.

P31: Measuring Adjective Spaces (presented in ICANN2010)

Tiina Lindh-Knuutila
Aalto University, School of Science, Department of Information and Computer Science
tiina.lindh-knuutila@aalto.fi

In this article, a set of adjectives is modeled using a vector space representation built using Wikipedia corpus for English. We then use three different dimension reduction methods, the Principal Component Analysis (PCA), the Self-Organizing Map (SOM), and the Neighbor Retrieval Visualizer (NeRV) in the projection and visualization task and evaluate the results with the antonym test. The results of the analysis between the three methods are comparable: all of the methods are able to preserve meaningful information for further analysis, but the NeRV perfoms the best of the three.

P32: Periodic Finite State Controllers for Efficient POMDP and DEC-POMDP Planning

Joni Pajarinen
Aalto University / Department of Information and Computer Science
Joni.Pajarinen@aalto.fi

Agents, such as robots and wireless devices, must plan their actions in an uncertain environment with noisy, partial observations. When the goal is to maximize the accumulated reward in a Markovian world, for a single agent a partially observable Markov decision process (POMDP), and for multiple distributed agents a decentralized POMDP (DEC-POMDP) is used to find the optimal policy. One common representation for the policy of agents are finite state controllers (FSCs). We introduce a novel class of periodic FSCs, composed of layers connected only to the previous and next layer. Our periodic FSC method finds a deterministic finite-horizon policy and converts it to an initial periodic infinite-horizon policy, which is

optimized by a new infinite-horizon algorithm to yield deterministic periodic policies. A new expectation maximization algorithm optimizes stochastic periodic policies. Our approach yields better results than state-of-the-art comparison methods and can compute larger solutions than what is possible with regular FSCs.

P33: The Use of Weighted Metabolic Graphs to Investigate the Large-scale Evolution of Metabolism

Fang Zhou
University of Helsinki, Department of Computer Science
fang.zhou@cs.helsinki.fi

We are interested in better understanding the evolution of metabolic biodiversity in bacteria - Archaea and the Eubacteria. To investigate this question, we introduce the use of weighted graphs to integrate large amounts of genomic data. We propose two ways of measuring the importance of enzymes, and apply the graph compression method to compare the importance of pathways in the different kingdoms.

P34: Automatic and Convenient Sleep Measurement in a Normal Bed

Joonas Paalasmaa
University of Helsinki, Department of Computer Science
joonas.paalasmaa@helsinki.fi

The quality and quantity of sleep is measured with a flexible piezo-electric sensor place under the mattress. Sleep analysis is performed by first detecting the heart rate, respiration rate and movements from the signal, and using those information to analyze sleep. The sleep information is presented to the user in a web service.

P35: Climate Induced Changes in Benthic Macrofauna – A non-Linear Model Approach

Dusan Sovilj
Aalto; School of Science / ICS department
dusan.sovilj@aalto.fi

The non-linear methods “optimally pruned extreme learning machine” (OPELM) and “optimally pruned k–nearest neighbours” (OPKNN) are applied to relate various climate indices to time series of biomass, abundance and species number of benthic macrofauna communities in the southern North Sea for the period 1978-2005. The results of these methods show that the performance in forecasting macrofauna communities is as poor as linear statistical downscaling if only one climate index is used as a predictor. If a multivariate predictor is used, OPKNN shows a good forecast for biomass and species number, but not for abundance. The improvement of the forecast is of major relevance especially in the presence of biological and climate regime shifts which occurred in the considered period.

P36: Analysis Pipeline to Detect RNA-binding Protein Binding Sites from Deep Sequencing Data

Kari Nousiainen
Aalto University School of Science, Department of Information and Computer Science
Kari.Nousiainen@Aalto.FI

The RNA-binding proteins (RBP) are known to regulate the function of the cell by binging the RNA transcripts. A new experimental procedure called PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation) enables transcriptome-level studies to detect the binding sites and the binding motifs of RBPs. The information elucidates the impact of the RBP on the function of the cell. There is no consensus method of analyzing PAR-CLIP data since the published studies are controversial. In this work, we constructed a pipeline to analyze PAR-CLIP data. The pipeline is used to find the binding sites and to predict the binding motifs of a novel RBP.

P37: Distributed String Mining Algorithm for High-Throughput Sequencing Data

Niko Välimäki
University of Helsinki / Department of Computer Science
nvalimak@cs.helsinki.fi

The goal of frequency constrained string mining is to extract substrings that discriminate two (or more) datasets. The existing algorithms are practical only up to a few gigabytes of input. We introduce a distributed algorithm and apply it to a large-scale metagenomics study.

P38: Redescription Mining Outside the Boolean World

Esther Galbrun
Department of Computer Science, HIIT, University of Helsinki.
esther.galbrun@cs.helsinki.fi

Redescription mining is a powerful data analysis tool that is used to find multiple descriptions of the same entities. Consider geographical regions as an example. They can be characterized by the fauna that inhabits them on one hand and by their meteorological conditions on the other hand. Finding such redescriptors, a task known as niche-finding, is of much importance in biology. We will present our contribution on extending redescription mining to non-Boolean data.

P39: Predicting Students' Performance from Their Activity Patterns in Online Learning Environments

Tapio Auvinen
Aalto University, Department of Computer Science and Engineering
tapio.auvinen@aalto.fi

In higher education, courses are often large and course staff do not have time to monitor the performance of individual students. Thus, students who are struggling with a course may easily drop out unless they ask for support by themselves. Online learning environments collect detailed data about students' activities but it is laborious for course staff to analyze. In this poster, we compare different statistical analysis and machine learning methods to see if the failure to pass a course could be automatically predicted from a student's activity pattern in a learning environment, so that support could be offered to those belonging to risk groups. As a benchmark, we use data from two online computer science learning environments that automatically assess programming exercises.

P40: Sparsity Penalization for LDA in High Dimensional Data

Nima Reyhani
Aalto University/ School of Science
nima.reyhani@aalto.fi

Linear discriminant analysis (LDA) relies on an estimation of the covariance matrix, which is a major issue in applying many statistical methods for high dimensional data. In this setting, due to inconsistency of the sample estimation, the LDA classification is no better than a fair coin. Here, we empirically study the effect of sparse penalization in improving the performance of LDA.

P41: Learning Creativity Using Remote Associates Tests

Oskar Gross
University of Helsinki / Department of Computer Science and HIIT
ogross@cs.helsinki.fi

We design minimally supervised methods that can solve (with reasonable accuracy) remote association test (RAT) questions, which are a well-known psychometric measure of creativity. Our main goal is to provide some simple principles, which we establish while analysing and solving RAT questions. These principles are used to develop methods for more general word associations model. We will show, that these methods could be used in lexical creativity support systems and also could be a small step towards possible lexical creativity in computers.