Processing of large document collections

Part 1 (Introduction, text representation, text categorization)

Helena Ahonen-Myka

Spring 2005

1. Introduction

- course organization
- introduction to the topic
  - applications
  - methods
- learning goals
- schedule

Organization of the course

- lectures (Helena Ahonen-Myka)
  - Tue 12-14, Thu 10-12 B222
  - 15.3.-28.4. (no lectures 24.3. and 29.3.)
- exercise sessions (Juha Makkonen)
  - Tue 14-16 DK118 and Fri 10-12 DK117
  - 21.3.-6.5. (no exercises 25.3. and 29.3.)
- exam: Thu 12.5. at 16-20, A111
- points: exam 50 pts, exercises 10 pts
  - required: ~30 pts (= 1-)

Course material

- slides on the course web page
- also other material available on the page
  - handouts used in the class (sample documents etc.)
  - original articles

Large document collections

- What is a document?
  - "a document records a message from people to people" (Wilkinson et al., 1998)
- each document has content, structure, and metadata (context)
  - in this course, we concentrate on content
  - particularly: textual content

Large document collections

- large?
  - some person may have written a document, but it is not possible later to process the document manually -> automatic processing is needed
  - large w.r.t to the capacity of a device (e.g. a mobile phone)
- collection?
  - documents somehow similar -> automatic processing is possible

Applications

- text categorization
- text summarization
- information extraction
- question answering
- text compression
- text indexing and retrieval
- machine translation

...

7

Text categorization

- given a predefined set of categories and a set of documents
- label each document with one or more categories

8

Text summarization

- "Process of distilling the most important information from a source to produce an abridged version for a particular user or task" (Mani & Maybury, 1999)

9

Example

A Spanish priest was charged here today with attempting to murder the Pope. Juan Fernandez Krohn, aged 32, was arrested after a man armed with a bayonet approached the Pope while he was saying prayers at Fatima on Wednesday night.

According to the police, Fernandez told the investigators today that he trained for the past six months for the assault. He was alleged to have claimed the Pope 'looked furious' on hearing the priest's criticism of his handling of the church's affairs. If found quilty, the Spaniard faces a prison sentence of 15-20 years.

10

Example

- summary could be, e.g.
  - "A Spanish priest is charged after an unsuccessful murder attempt on the Pope"
- or a set of phrases:
  - a Spanish priest was charged
  - attempting to murder the Pope
  - he trained for the assault
  - Pope furious on hearing priest´s criticisms

11

Information extraction

- "Information extraction involves the creation of a structured representation (such as a database) of selected information drawn from the text" (Grishman, 1997)

12

Example: terrorist events

19 March - A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported.

According to unofficial sources, the bomb - allegedly detonated by urban guerrilla commandos - blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).

13

Example: terrorist events

Incident type	bombing
Date	March 19
Location	El Salvador: San Salvador (city)
Perpetrator	urban guerilla commandos
Physical target	power tower
Human target	-
Effect on physical target	destroyed
Effect on human target	no injury or death
Instrument	bomb

14

Example: terrorist events

- a document collection is given
- for each document, decide if the document is about terrorist event
- for each terrorist event, determine
  - type of attack
  - date
  - location, etc.
- = fill in a template (~database record)

15

Question answering systems

- the user asks a question in a natural language
- the question answering system finds answers from a document collection, e.g. from a collection of newspaper stories

16

Example

- question:
  - When did Chuck Yeager break the sonic barrier?
- a text fragment in the collection:
  - "For many, seeing **Chuck Yeager** – who made his historic super**sonic** flight **Oct. 14, 1947** – was the highlight of this year's show, in which..."
- answer: Oct. 14, 1947

17

Methods

- typically several methods (from several research fields) are combined in each application
  - statistics (or simply counting frequencies...)
  - machine learning
  - knowledge-based methods
  - linguistic methods
  - algorithmics

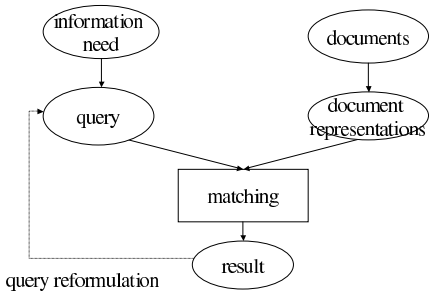
18

Learning goals

- learn to recognize components of applications/processes
- learn to recognize which (kind of) methods could be used in each component
- learn to implement some methods
- (meta)learn to control learning processes (What do I know? What should I know to solve this problem?)

19

Mapping to the information retrieval process



20

Schedule

- 15.-22.3.
  - text representation, text categorization, term selection
- 31.3.-7.4.
  - text summarization
- 12.4.-19.4.
  - information extraction
- 21.-26.4
  - question answering systems,…
- 28.4.
  - closing

21

2. Text representation

- selection of terms
- vector model
- weighting (TD*IDF)

22

Text representation

- text cannot be directly interpreted by the many document processing applications
- we need a compact representation of the content
- which are the meaningful units of text?

23

Terms

- words
  - typical choice
  - set of words, bag of words
- phrases
  - syntactical phrases (e.g. noun phrases)
  - statistical phrases (e.g. frequent pairs of words)
  - usefulness not yet known?

24

Terms

- part of the text is not considered as terms: these words can be removed
  - very common words (function words):
    - articles (a, the) , prepositions (of, in), conjunctions (and, or), adverbs (here, then)
  - numerals (30.9.2002, 2547)
- other preprocessing possible
  - stemming (recognization -> recogn), base words (skies -> sky)

25

Vector model

- a document is often represented as a vector
- the vector has as many dimensions as there are terms in the whole collection of documents

26

Vector model

- in our sample document collection, there are 118 words (terms)
- in alphabetical order, the list of terms starts with:
  - absorption
  - agriculture
  - anaemia
  - analyse
  - application
  - …

27

Vector model

- each document can be represented by a vector of 118 dimensions
- we can think a document vector as an array of 118 elements, one for each term, indexed, e.g. 0-117

28

Vector model

- let d1 be the vector for document 1
- record only which terms occur in document:
  - d1[0] = 0 -- absorption doesn't occur
  - d1[1] = 0 -- agriculture -"-
  - d1[2] = 0 -- anaemia -"-
  - d1[3] = 0 -- analyse -"-
  - d1[4] = 1 -- application occurs
  - ...
  - d1[21] = 1 -- current occurs
  - …

29

Weighting terms

- usually we want to say that some terms are more important (for some document) than the others -> weighting
- weights usually range between 0 and 1
  - 1 denotes presence, 0 absence of the term in the document

30

Weighting terms

- if a word occurs many times in a document, it may be more important
  - but what about very frequent words?
- often the **TF*IDF** function is used
  - higher weight, if the term occurs often in the document
  - lower weight, if the term occurs in many documents

31

Weighting terms: TF*IDF

- TF*IDF = term frequency * inversed document frequency
- weight of term t_k in document d_j:

$$tfidf(t_k, d_j) = \#(t_k, d_j) \cdot \log \frac{|Tr|}{\#Tr(t_k)}$$

- where
  - $\#(t_k, d_j)$: the number of times t_k occurs in d_j
  - $\#Tr(t_k)$: the number of documents in Tr in which t_k occurs
  - Tr: the documents in the collection

32

Weighting terms: TF*IDF

- in document 1:
  - term 'application' occurs once, and in the whole collection it occurs in 2 documents:
    - tfidf (application, d1) = 1 * log(10/2) = log 5 ~ 0.7
  - term 'current' occurs once, in the whole collection in 9 documents:
    - tfidf(current, d1) = 1 * log(10/9) ~ 0.05

33

Weighting terms: TF*IDF

- if there were some word that occurs 7 times in doc 1 and only in doc 1, the TF*IDF weight would be:
  - tfidf(doc1word, d1) = 7 * log(10/1) = 7

34

Weighting terms: normalization

- in order for the weights to fall in the [0,1] interval, the weights are often normalized (T is the set of terms):

$$w_{kj} = \frac{tfidf(t_k, d_j)}{\sqrt{\sum_{s=1}^{|T|} (tfidf(t_s, d_j))^2}}$$

35

3. Text categorization

- problem setting
- two examples
- two major approaches
- mapping to the information retrieval process?

36

Text categorization

- text classification, topic classification/spotting/detection
- problem setting:
  - assume: a predefined set of categories, a set of documents
  - label each document with one (or more) categories

37

Text categorization

- let
  - D: a collection of documents
  - C = $\{c_1, ..., c_{|C|}\}$: a set of predefined categories
  - T = true, F = false
- the task is to approximate the unknown target function Φ': D x C -> {T,F} by means of a function Φ : D x C -> {T,F}, such that the functions "coincide as much as possible"
- function Φ' : how documents should be classified
- function Φ : classifier (hypothesis, model...)

38

Example

- for instance
  - categorizing newspaper articles based on the topic area, e.g. into the following 17 "IPTC" categories:
    - Arts, culture and entertainment
    - Crime, law and justice
    - Disaster and accident
    - Economy, business and finance
    - Education
    - Environmental issue
    - Health
    - ...

39

Example

- categorization can be hierarchical
  - Arts, culture and entertainment
    - archaeology
    - architecture
    - bullfighting
    - festive event (including carnival)
    - cinema
    - dance
    - fashion
    - ...

40

Example

- "Bullfighting as we know it today, started in the village squares, and became formalised, with the building of the bullring in Ronda in the late 18th century. From that time,..."
- class:
  - Arts, culture and entertainment
  - Bullfighting
  - or both?

41

Example

- another example: filtering spam
- "Subject: Congratulation! You are selected!
  - It's Totally FREE! EMAIL LIST MANAGING SOFTWARE! EMAIL ADDRESSES RETRIEVER from web! GREATEST FREE STUFF!"
- two classes only: Spam and Not-spam

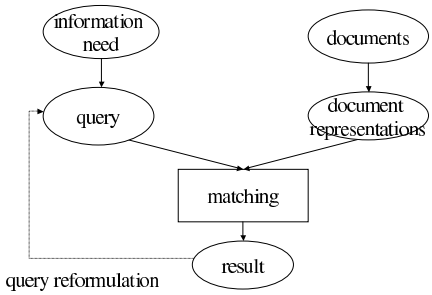
42

Text categorization

- two major approaches:
  - knowledge engineering -> end of 80's
    - manually defined set of rules encoding expert knowledge on how to classify documents under the given gategories
  - machine learning, 90's ->
    - an automatic text classifier is built by learning, from a set of preclassified documents, the characteristics of the categories

43

Mapping to the information retrieval process?



44