Text Mining for Creative Cross-Domain Knowledge Discovery

Nada Lavrač
Jožef Stefan Institute (JSI)
Ljubljana, Slovenia

with contributors
Bojan Cestnik, Matjaž Juršič, Borut Sluban, Tanja Urbančič, et al.

(selected text mining slides by Dunja Mladenić)

Talk outline

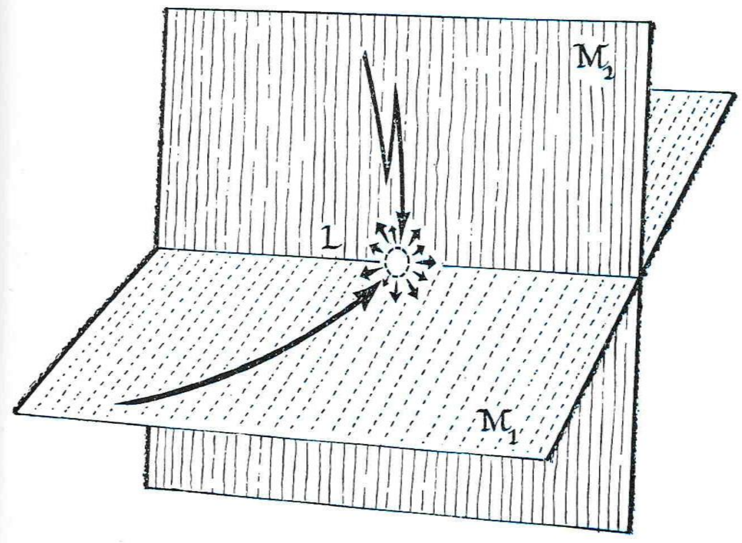
- Background and motivation
- Background technologies
  - Literature-based discovery
  - Text mining
- Cross-domain literature mining approaches
  - Outlier detection for cross-domain knowledge discovery
  - Cross-domain knowledge discovery with CrossBee
- Summary and conclusions
- CrossBee demo by Bojan Cestnik

Background

- **Boden** (The Creative Mind – Myths and Mechanisms, 2003):
  - Three types of creativity: combinatorial, exploratory, transformational
- **Koestler** (The Act od Creation, 1964):
  - “Creative act uncovers, selects, re-shuffles, combines, synthesizes already existing facts, ideas, faculties, skills. The more familiar the parts, the more striking the new whole.”
- **Berthold** (Bisociative knowledge discovery, 2012):
  - Computational tools can support humans in creative (exploratory, combinatorial) knowledge discovery

Background

- **Boden (2003):**
  - Creativity as “the ability to come up with ideas or artifacts that are new, surprising and valuable”.
- **Koestler (1964):**
  - Ideas often come from different contexts.
  - “… the perceiving of a situation or idea L, *in two self-consistent but habitually incompatible frames of reference, matrices or contexts M1 and M2*. The event L ... is not merely linked to one associative context but **bisociated** with two.”
  - Bisociation is a basis for human creativity in humor, science and art.



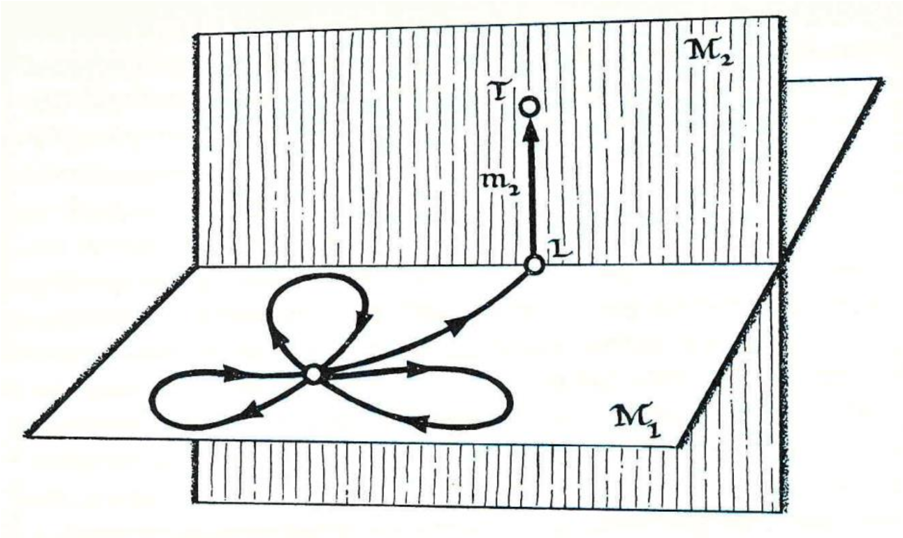
Koestler: The Archimedes example

Archimedes, a leading scientists in classical antiquity, was tasked with the problem of determining whether a crown (a present for Hiero, tyrant of Syracuse) consisted of pure gold or was adulterated with silver. To solve this problem Archimedes needed to measure the volume of the crown. At the time no method existed to determine the volume of such an irregularly shaped three-dimensional object.

Koestler: The Archimedes example

One day, while taking a bath, Archimedes noticed the rise of the water level as his body was sliding to the basin. It was at this point when he realized that the volume of water displaced was equal to the volume of the immersed parts of his own body. At this **Eureka moment** both matrices (associations of taking a bath and knowledge of geometry) were simultaneously active.

Koestler: The Archimedes example

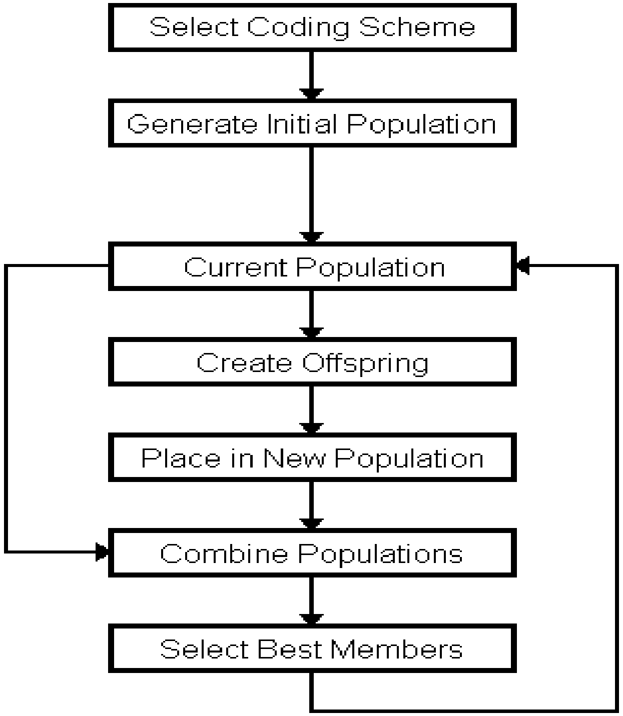


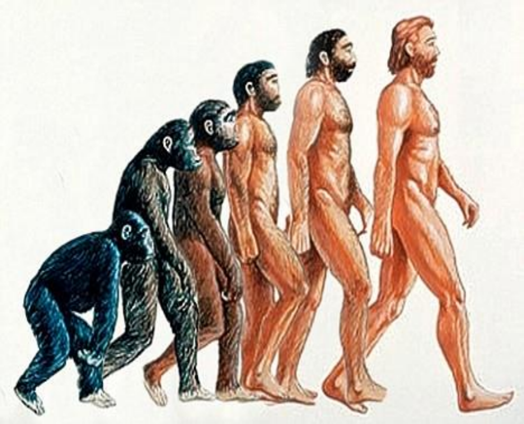
taking
a bath

computing
the volume

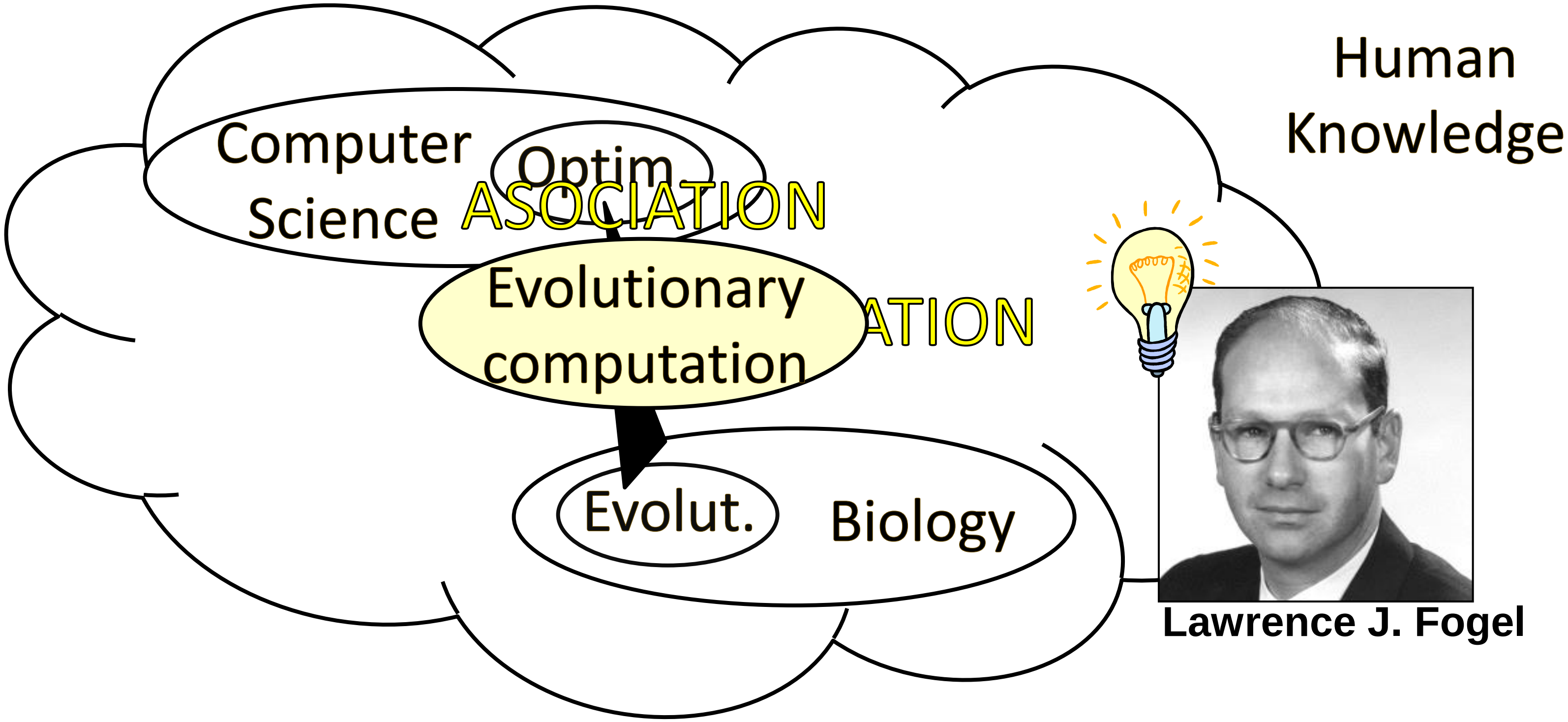
Example from the history of computer science

- **From evolution in nature to evolutionary computing (Lawrence J. Fogel, 1964)**
  - from "survival of the fittest" in nature
  - to the idea of populations of candidate solutions developing through simulated evolution





Example from the history of computer science



Human Knowledge

Computer Science

Optim.

ASOCIATION

Evolutionary computation

ATION

Evolut.

Biology

Lawrence J. Fogel

The BISON project

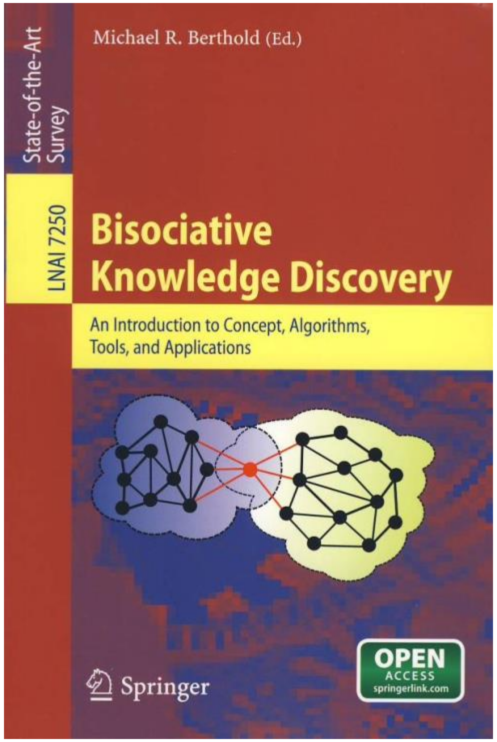
- BISON: Bisociation Networks for Creative Information Discovery, European 7FP project, www.bisonet.eu, 12 partners (2008-2011)
- Explore the idea of bisociation (Arthur Koestler, The act of creation, 1964)
- To develop computational tools which can support humans in creative (exploratory, combinatorial) knowledge discovery

The BISON project

- BISON: Bisociation Networks for Creative Information Discovery, European 7FP project, www.bisonet.eu, 12 partners (2008-2011)
- Open access book (Springer 2012):

Bisociative Knowledge Discovery

edited by M. Berthold



State-of-the-Art Survey

LNAI 7250

Michael R. Berthold (Ed.)

Bisociative Knowledge Discovery

An Introduction to Concept, Algorithms, Tools, and Applications

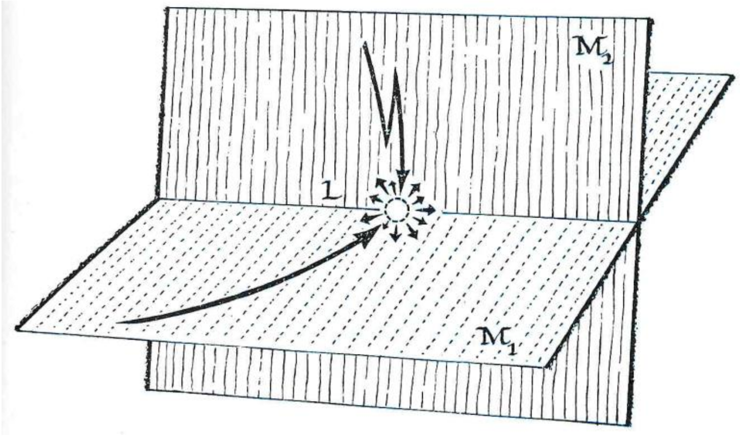
Springer

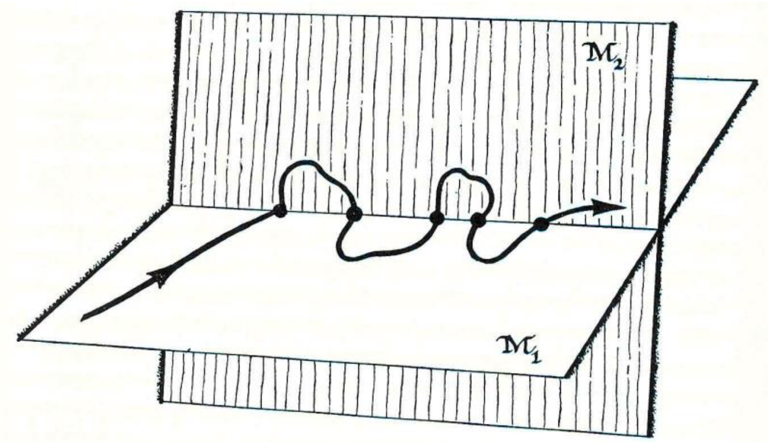
OPEN ACCESS springerlink.com

http://link.springer.com/book/10.1007%2F978-3-642-31830-6

Bisociation discovery in BISON

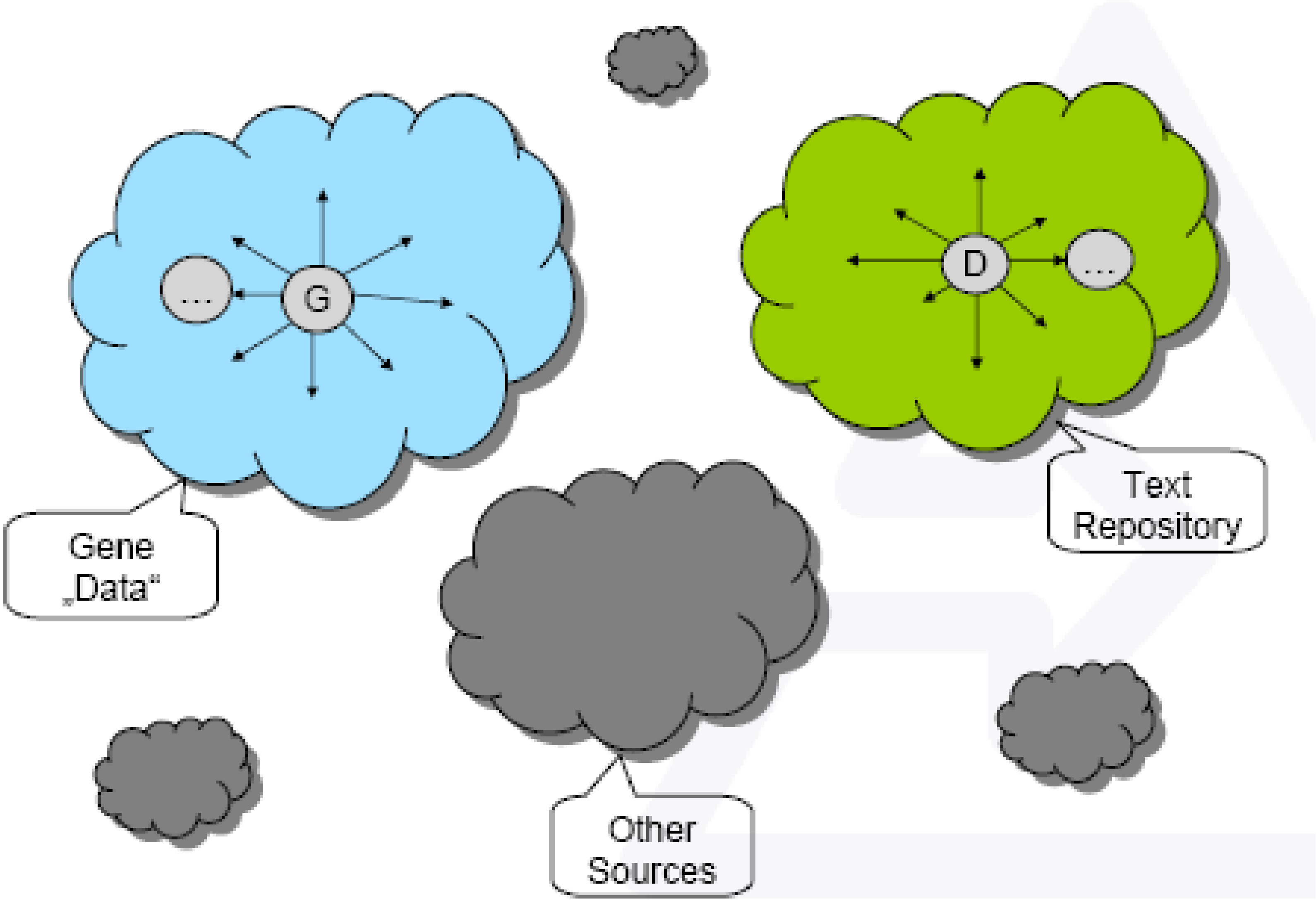
- BISON challenge:
  - Find new insights: new **bisociations,** i.e., interesting new links **accross domains**
- Two concepts are bisociated if and only if:
  - There is no direct, obvious evidence linking them
  - One has to cross contexts to find the link
  - This new link provides some novel insight



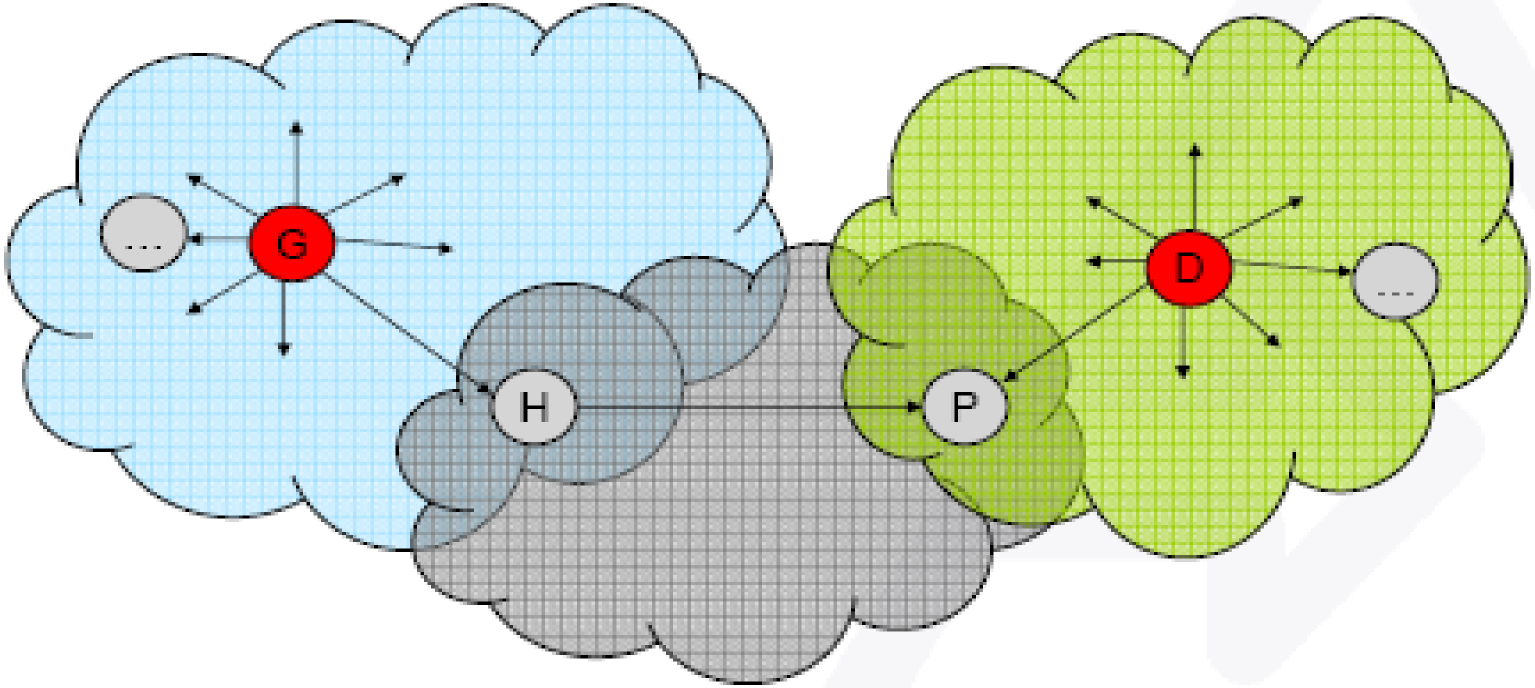


Heterogeneous data sources

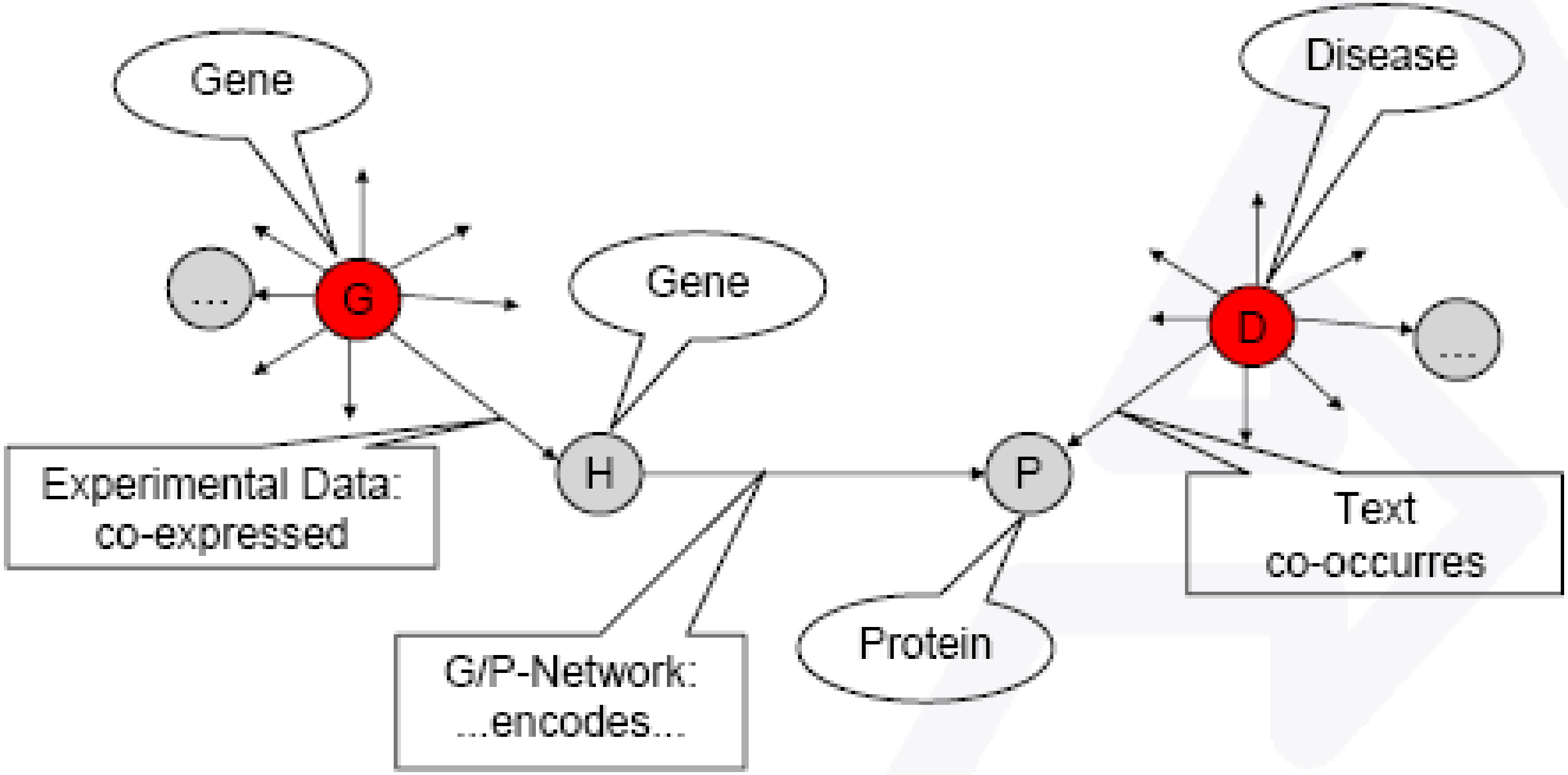
(BISON, M. Berthold, 2008)



Bridging concepts
(BISON, M. Berthold, 2008)



Chains of associations across domains (BISON, M. Berthold, 2008)

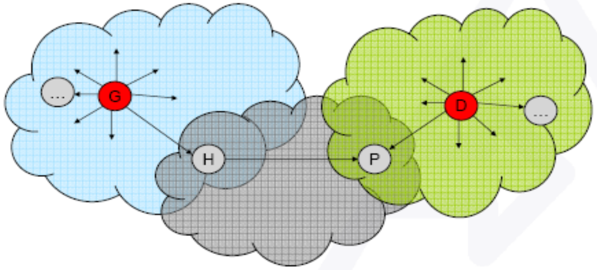


Main BISON approach

- Main approach: graph exploration
  - Find bisociations as yet unexplored links in a graph, crossing different contexts (domains)
- Open problems:
  - How to cross different types of data and knowledge sources: By fusing heterogeneous data/knowledge sources into a joint representation format - a large information network named BisoNet (consisting of nodes and relatioships between nodes)
  - How to cross different contexts (domains): By finding unexpected, previously unknown links between BisoNet nodes belonging to different contexts

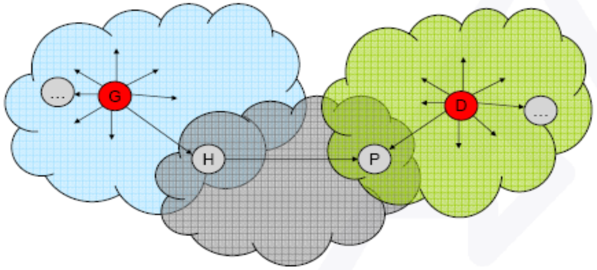
Main BISON approach

- Main approach: graph exploration
  - Find yet unexplored links in a graph, crossing different domains (contexts)

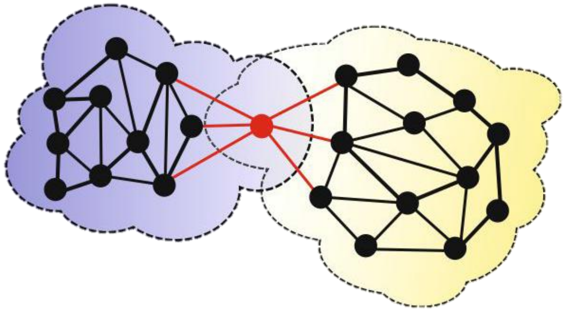


Main BISON approach

- Main approach: graph exploration
  - Find yet unexplored links in a graph, crossing different domains (contexts)



  - Simplified setting, starting from two predefined domains (i.e., the “closed discovery” setting): Find interesting bridging nodes at the intersection of the two domains



Complementary BISON approach

- Complementary approach: text mining
  - Find yet unexplored terms in the intersection of domains, crossing different contexts (domains/literatures), helping experts in cross-domain discovery for new findings

Complementary BISON approach

- Complementary approach: text mining
  - Find yet unexplored terms in the intersection of domains, crossing different contexts (domains/literatures), helping experts in cross-domain discovery for new findings
  - Addressing two settings:
    - Closed discovery setting (two predefined domains)
    - Open discovery setting (one defined domain, determining the other through exploration)
  - Closed literature-based discovery formulated in BISON:
    - Find bisociations, as bridging terms (b-terms) linking different contexts (domains)

Complementary BISON approach

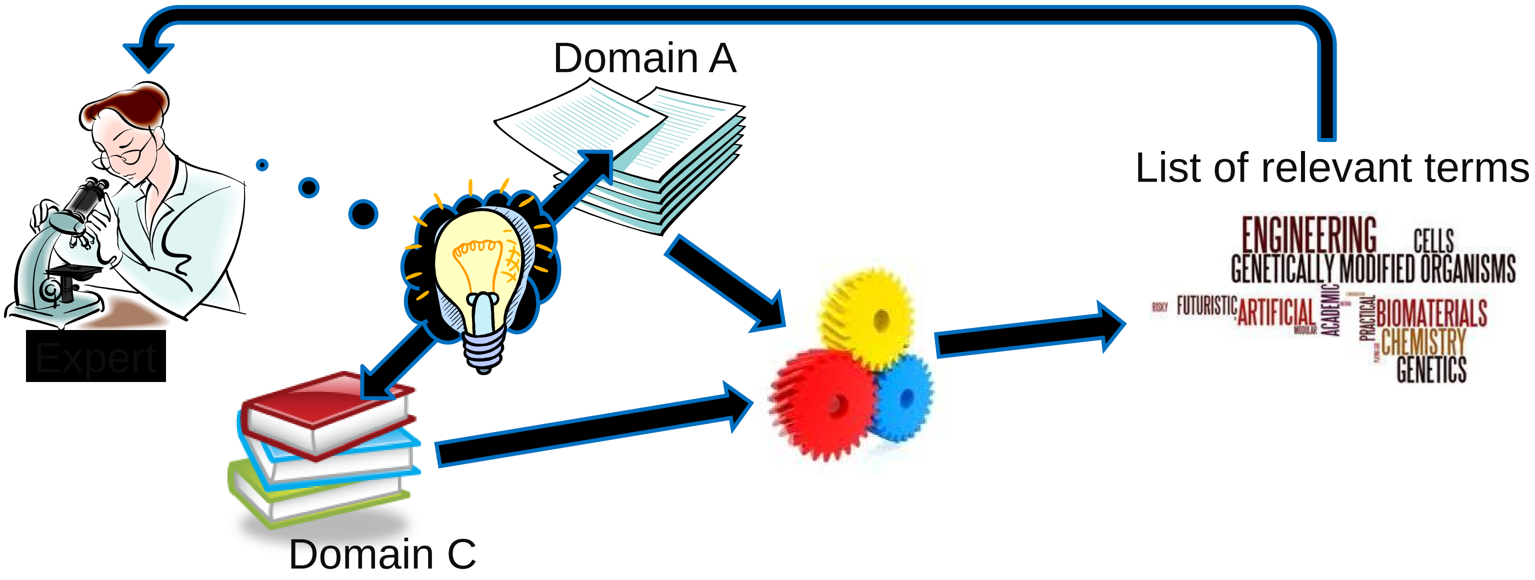
- Early related work: literature-based discovery (LBD)
  - Swanson (1988, 1990)
  - Smalheiser, Swanson (1998): ARROWSMITH
  - Weeber et al. (2001)
  - Hristovski et al. (2001): BITOLA
  - …
- Our recent work: cross-domain literature mining
  - Petrič et al. (2007, 2009): RaJoLink
  - Juršič et al. (2012): CrossBee
  - …

Talk outline

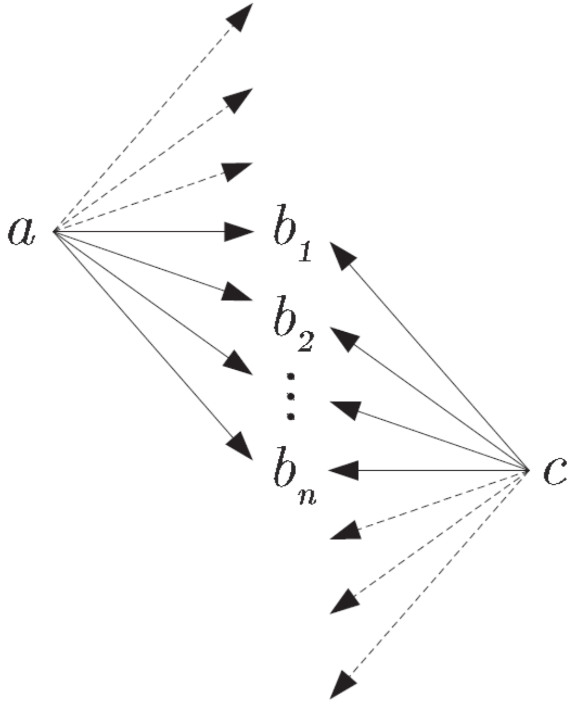
- Background and motivation
- Background technologies
  - Literature-based discovery
  - Text mining
- Cross-domain literature mining approaches
  - Outlier detection for cross-domain knowledge discovery
  - Cross-domain knowledge discovery with CrossBee
- Summary and conclusions
- CrossBee demo by Bojan Cestnik

Literature-based discovery (LBD)

- Help experts in cross-domain discovery for unknown facts/new findings
  - Early work by Swanson: Medical literature as a potential source of new knowledge, 1988, 1990
  - Closed discovery setting, bridging terms detection

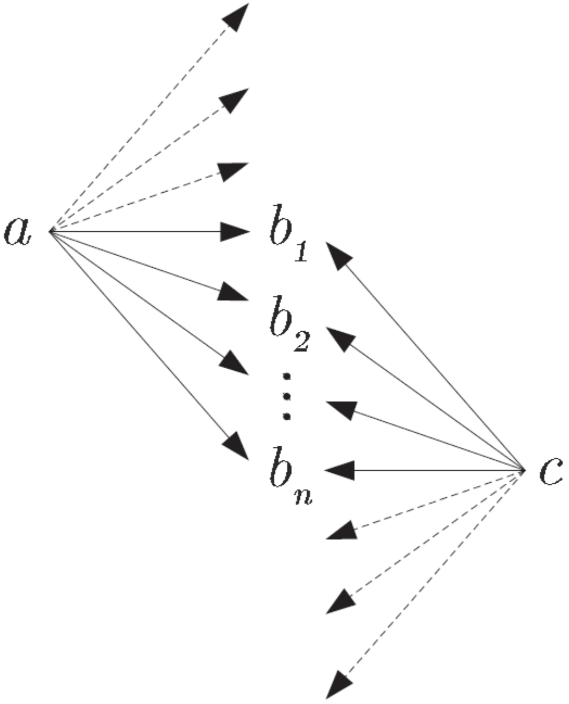


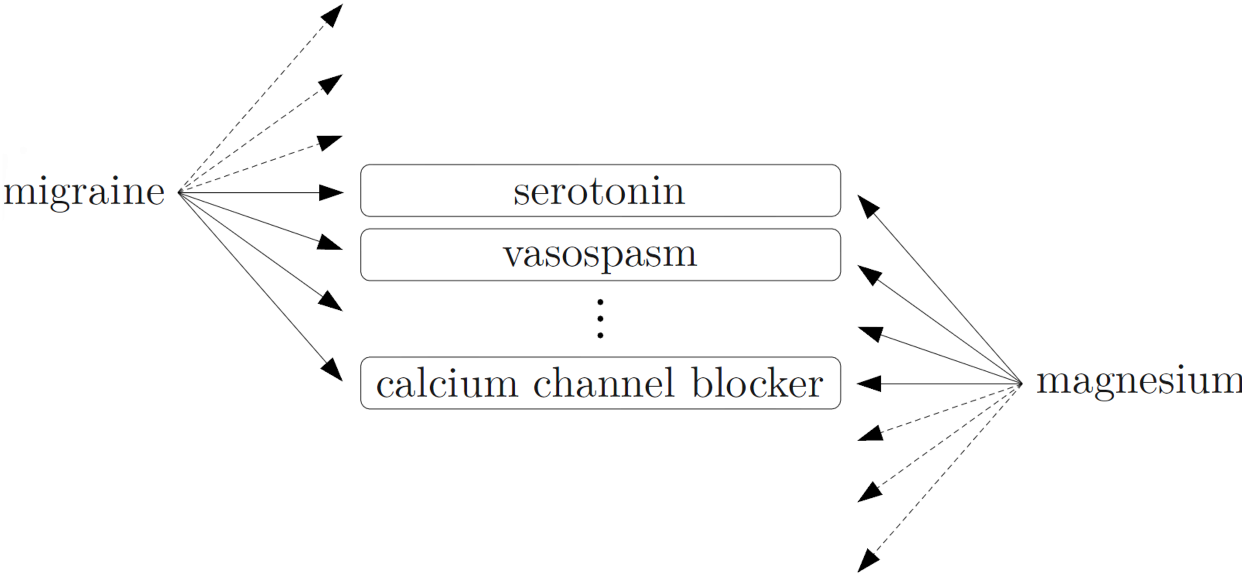
Closed discovery setting: Finding linking (bridging) terms



Swanson’s ABC model

Closed discovery setting: Finding linking (bridging) terms

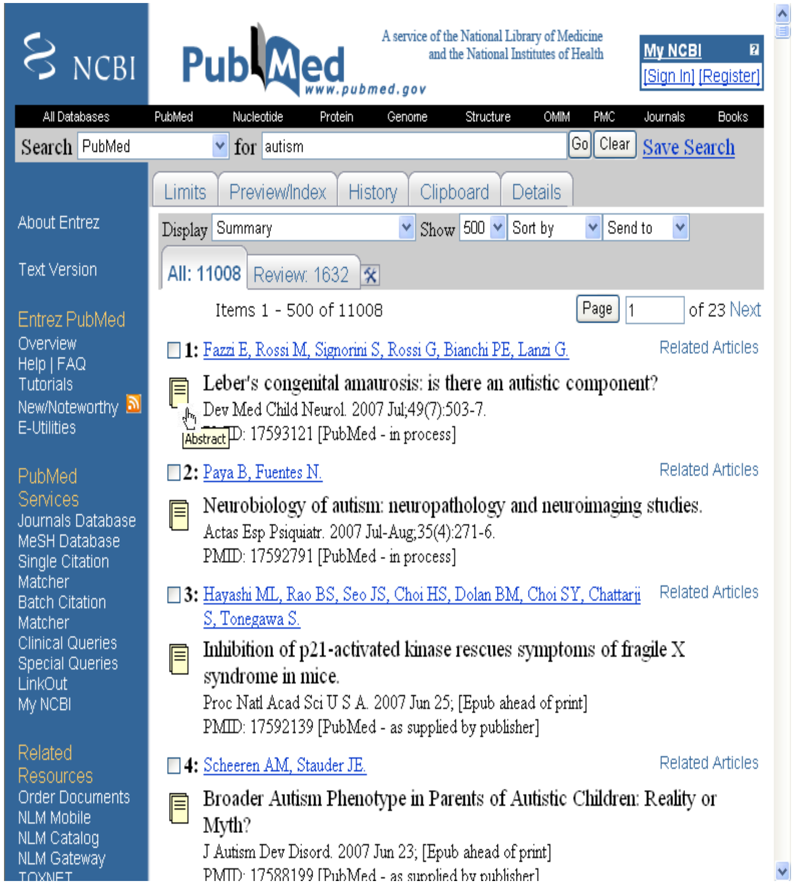




Swanson’s ABC model
B-terms: calcium channel blocker, …

Scientific literature as a source of knowledge

- Biomedical bibliographical database PubMed
- US National Library of Medicine
- More than 21M citations
- More than 5,600 journals
- 2,000 – 4,000 references added each working day!



Search PubMed for autism

Display Summary Show 500 Sort by Send to

All: 11008 Review: 1632

Items 1 - 500 of 11008 Page 1 of 23 Next

□ 1: Fazzi E, Rossi M, Signorini S, Rossi G, Bianchi PE, Lanzi G.

Leber's congenital amaurosis: is there an autistic component?
Dev Med Child Neurol. 2007 Jul;49(7):503-7.
D: 17593121 [PubMed - in process]

□ 2: Paya B, Fuentes N.

Neurobiology of autism: neuropathology and neuroimaging studies.
Actas Esp Psiquiatr. 2007 Jul-Aug;35(4):271-6.
PMID: 17592791 [PubMed - in process]

□ 3: Hayashi ML, Rao BS, Seo JS, Choi HS, Dolan BM, Choi SY, Chattarji S, Tonegawa S.

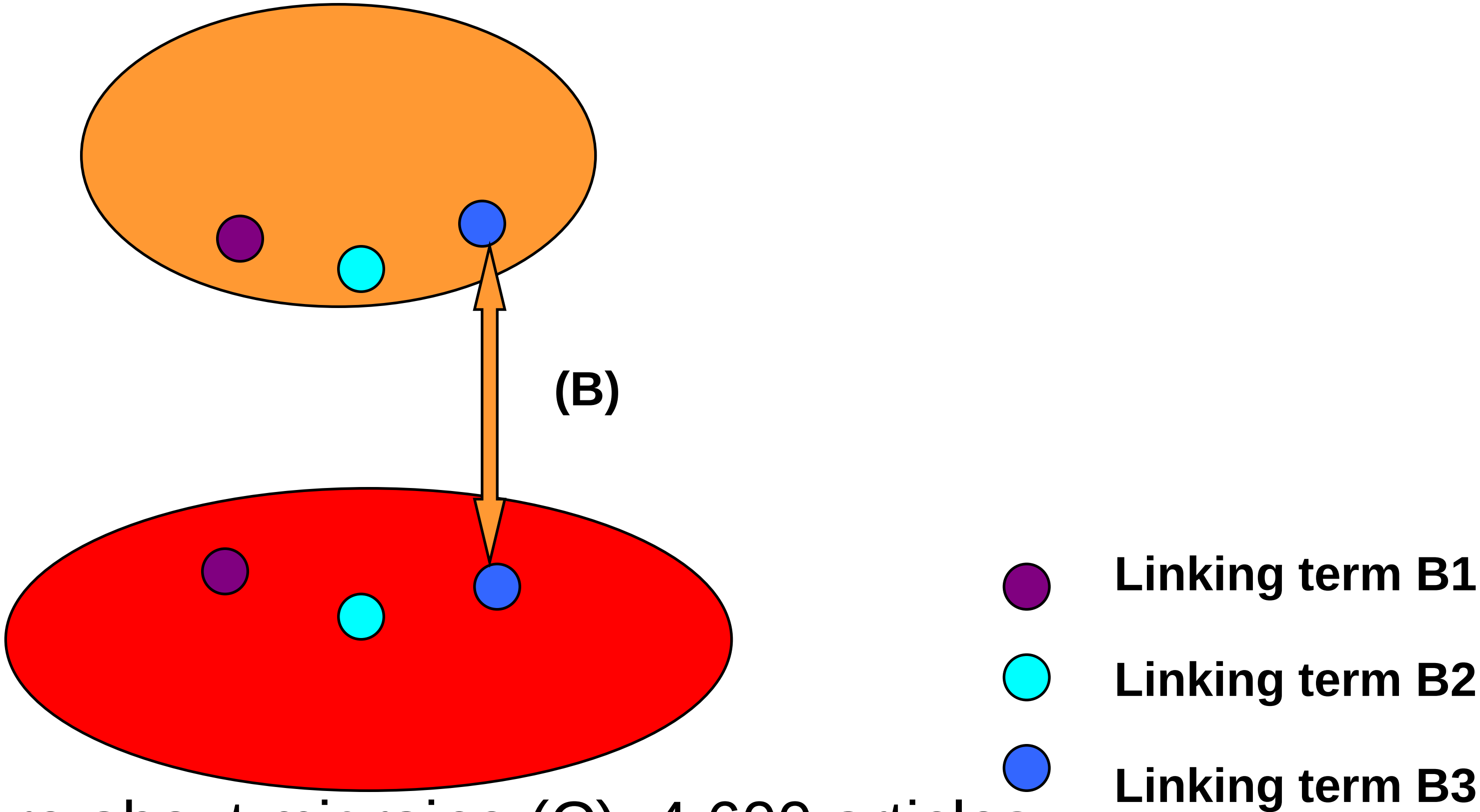
Inhibition of p21-activated kinase rescues symptoms of fragile X syndrome in mice.
Proc Natl Acad Sci U S A. 2007 Jun 25; [Epub ahead of print]
PMID: 17592139 [PubMed - as supplied by publisher]

□ 4: Scheeren AM, Stauder JE.

Broader Autism Phenotype in Parents of Autistic Children: Reality or Myth?
J Autism Dev Disord. 2007 Jun 23; [Epub ahead of print]
PMID: 17588199 [PubMed - as supplied by publisher]

Closed discovery setting: Finding linking (bridging) terms

Literature about magnesium (A): 38,000 articles



(B)

● Linking term B1

● Linking term B2

● Linking term B3

Literature about migraine (C): 4,600 articles

Closed discovery setting: Finding linking (bridging) terms

Argument 1 (magnesium literature)

- Mg is a natural calcium channel blocker.
- Stress and Type A behavior can lead to body loss of Mg.
- Magnesium has anti-inflammatory properties.
- . . .

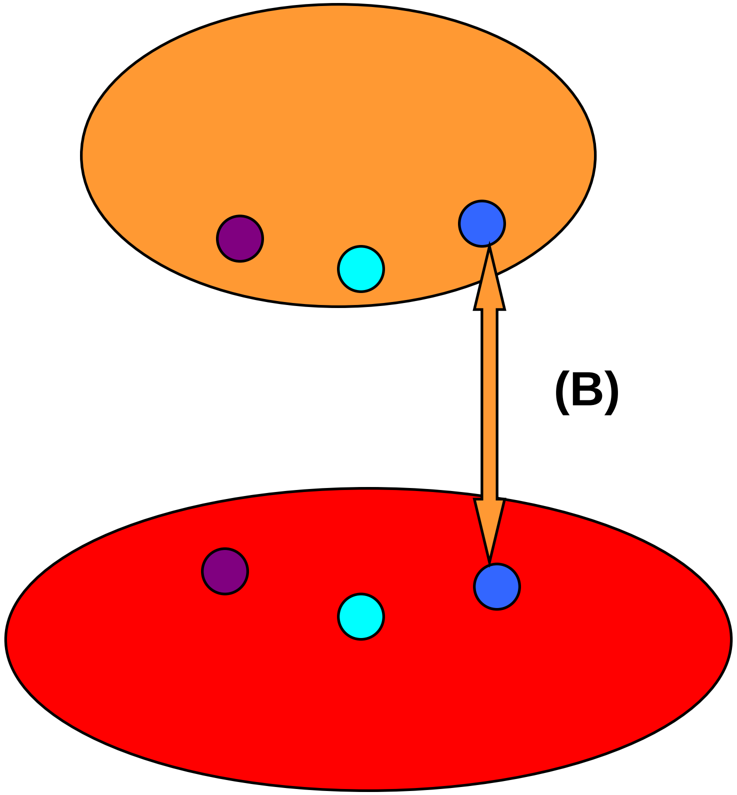
Argument 2 (migraine literature)

- Calcium channel blockers can prevent migraine attacks.
- Stress and Type A behavior are associated with migraine.
- Migraine may involve sterile inflammation of the cerebral blood vessels.
- . . .

Closed discovery setting: Finding linking (bridging) terms

Work by
Petrič et al. 2009

Literature A (calcineurin)



(B)

Literature C (autism)

● Linking term B1

● Linking term B2

● Linking term B3

Examples of b-terms

Autism literature:

- Fatemi et al. (2001) reported a reduction of ***Bcl-2*** (a regulatory protein for control of programmed brain cell death) levels in autism cerebellum.
- Huber et al. (2002) showed evidence about an important function role of fragile X protein, an identified cause of autism, in regulating activity-dependent ***synaptic plasticity*** in the brain.
- Román (2007) proposed that morphological brain changes in autism may be produced by ***maternal hypothyroxinemia*** resulting in low triiodothyronine in the fetal brain during pregnancy.

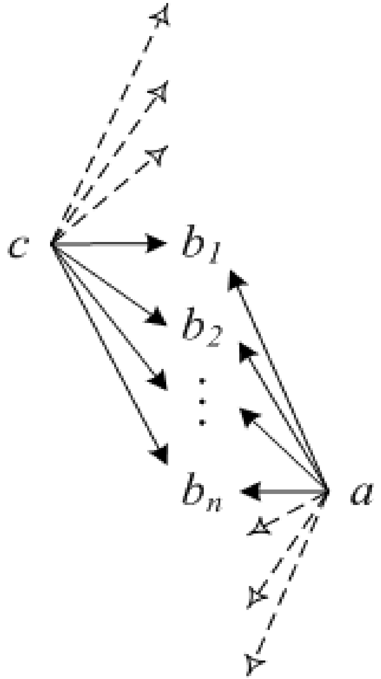
Calcineurin literature:

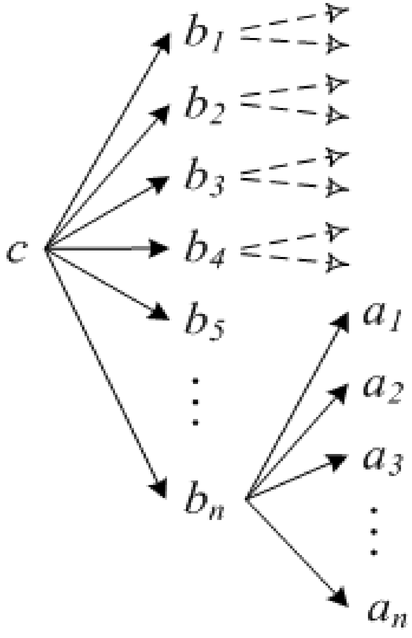
- Erin et al. (2003) observed that calcineurin occured as a coplex with ***Bcl-2*** in various regions of rat and mouse brain.
- Winder and Sweatt (2001) described the critical role of protein phosphatase 1, protein phosphatase 2A and calcineurin in the activity-dependent alterations of ***synaptic plasticity***.
- Sinha et al. (1992) found that calcineurine was compromised in young progeny when they investigated the ***maternal hypothyroxinemia*** effect during pregnancy on brain of young progeny.

From pairs of MEDLINE articles about autism and calcineurin, I. Petrič PhD Thesis

Closed vs. open discovery (Weeber et al. 2001)

- **Closed discovery:**
  - A and C are known: Given two separate literatures A and C, find bridging terms B
- **Open discovery:**
  - Only C is known: Given literature C, how do we find A?





Closed vs. open discovery (Weeber et al. 2001)

- **Closed discovery:**
  - – A and C are known: Given two separate literatures A and C, find bridging terms B
- **Open discovery:**
  - – Only C is known: Given literature C, how do we find A?
  - – Swanson: “Search proceeds via some intermediate literature (B) toward an unknown destination A. … Success depends entirely on the knowledge and ingenuity of the searcher.”
- **Text mining for cross-domain knowledge discovery:**
  - – Can we provide systematic support to the closed and open discovery process ?

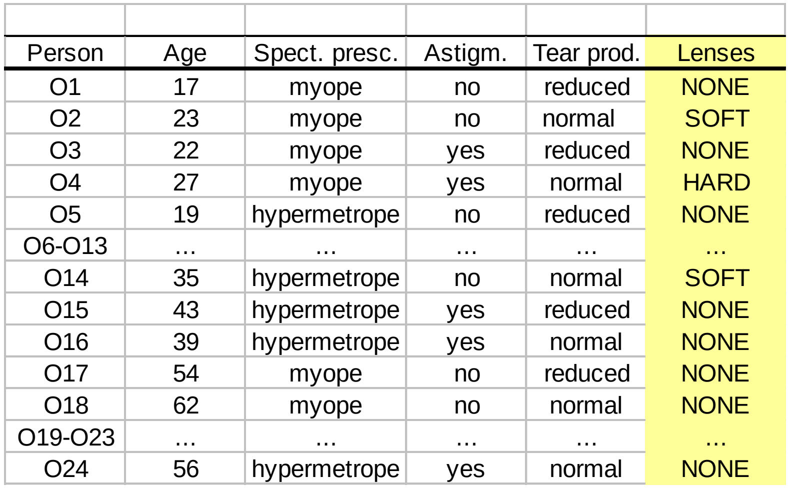
Text mining for coss-domain knowledge discovery

- **Situation:**
  - Growing speed of knowledge growth, huge ammounts of literature available on-line
  - High specialization of researchers
  - Potentially useful connections between “islands” of knowledge may remain hidden
- **Research objective:**
  - To develop methods and text mining tools to support researchers in the discovery of new knowledge from literature

Talk outline

- Background and motivation
- Background technologies
  - Literature-based discovery
  - Text mining
- Cross-domain literature mining approaches
  - Outlier detection for cross-domain knowledge discovery
  - Cross-domain knowledge discovery with CrossBee
- Summary and conclusions
- CrossBee demo by Bojan Cestnik

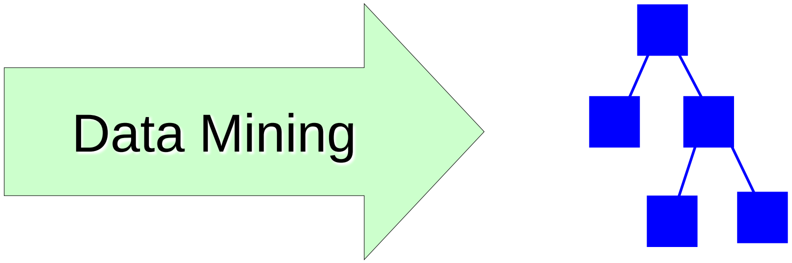
Background: Data mining



Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NONE
O2	23	myope	no	normal	SOFT
O3	22	myope	yes	reduced	NONE
O4	27	myope	yes	normal	HARD
O5	19	hypermetrope	no	reduced	NONE
O6-O13	...	...	...	...	...
O14	35	hypermetrope	no	normal	SOFT
O15	43	hypermetrope	yes	reduced	NONE
O16	39	hypermetrope	yes	normal	NONE
O17	54	myope	no	reduced	NONE
O18	62	myope	no	normal	NONE
O19-O23	...	...	...	...	...
O24	56	hypermetrope	yes	normal	NONE

data

knowledge discovery
from data



model, patterns, clusters, …

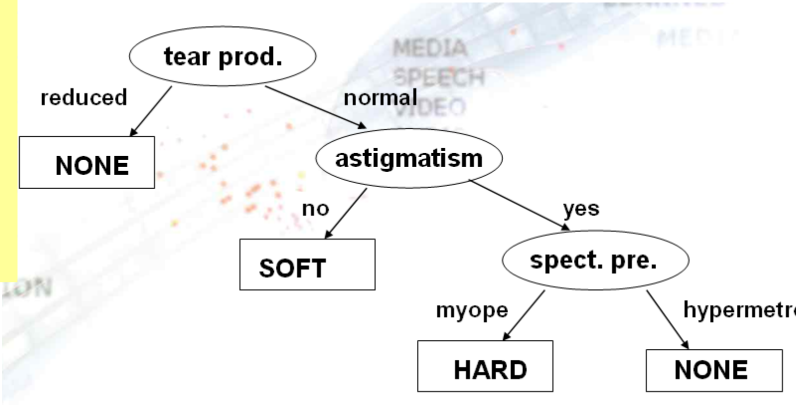
Given: transaction data table, a set of text documents, …

Find: a classification model, a set of interesting patterns

Data mining

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NONE
O2	23	myope	no	normal	SOFT
O3	22	myope	yes	reduced	NONE
O4	27	myope	yes	normal	HARD
O5	19	hypermetrope	no	reduced	NONE
O6-O13	...	...	...	...	...
O14	35	hypermetrope	no	normal	SOFT
O15	43	hypermetrope	yes	reduced	NONE
O16	39	hypermetrope	yes	normal	NONE
O17	54	myope	no	reduced	NONE
O18	62	myope	no	normal	NONE
O19-O23	...	...	...	...	...
O24	56	hypermetrope	yes	normal	NONE

Data Mining



lenses=NONE ← tear production=reduced

lenses=NONE ← tear production=normal AND astigmatism=yes AND
spect. pre.=hypermetrope

lenses=SOFT ← tear production=normal AND astigmatism=no

lenses=HARD ← tear production=normal AND astigmatism=yes AND
spect. pre.=myope

lenses=NONE ←

Data mining: Task reformulation

Person	Young	Myope	Astigm.	Reuced tea	Lenses
O1	1	1	0	1	NO
O2	1	1	0	0	YES
O3	1	1	1	1	NO
O4	1	1	1	0	YES
O5	1	0	0	1	NO
O6-O13	...	...	...	...	...
O14	0	0	0	0	YES
O15	0	0	1	1	NO
O16	0	0	1	0	NO
O17	0	1	0	1	NO
O18	0	1	0	0	NO
O19-O23	...	...	...	...	...
O24	0	0	1	0	NO

Binary features and class values

Text mining:
Words/terms as binary features

Document	Word1	Word2	…	WordN	Class
d1	1	1	0	1	NO
d2	1	1	0	0	YES
d3	1	1	1	1	NO
d4	1	1	1	0	YES
d5	1	0	0	1	NO
d6-d13	...	...	...	...	...
d14	0	0	0	0	YES
d15	0	0	1	1	NO
d16	0	0	1	0	NO
d17	0	1	0	1	NO
d18	0	1	0	0	NO
d19-d23	...	...	...	...	...
d24	0	0	1	0	NO

Instances = documents

Words and terms = Binary features

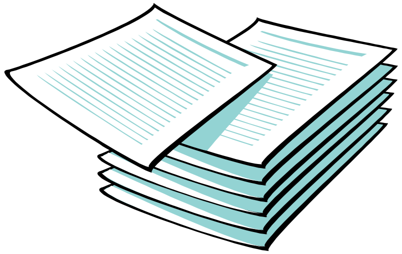
Text Mining from unlabeled data

Document	Word1	Word2	…	WordN	~~Class~~
d1	1	1	0	1	~~NO~~
d2	1	1	0	0	~~YES~~
d3	1	1	1	1	~~NO~~
d4	1	1	1	0	~~YES~~
d5	1	0	0	1	~~NO~~
d6-d13	...	...	...	...	~~...~~
d14	0	0	0	0	~~YES~~
d15	0	0	1	1	~~NO~~
d16	0	0	1	0	~~NO~~
d17	0	1	0	1	~~NO~~
d18	0	1	0	0	~~NO~~
d19-d23	...	...	...	...	~~...~~
d24	0	0	1	0	~~NO~~

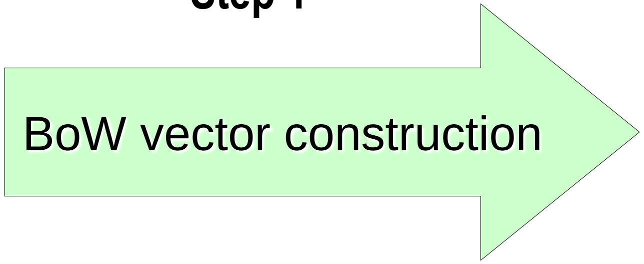
Unlabeled data - clustering: grouping of similar instances

- association rule learning

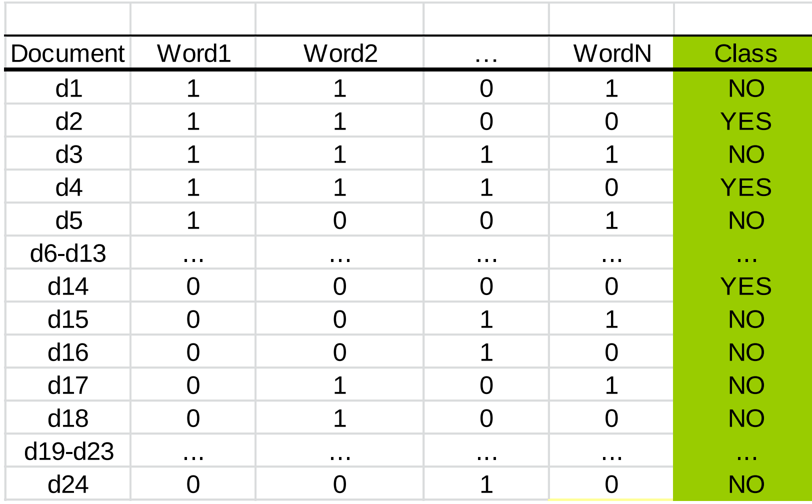
Text mining



Step 1

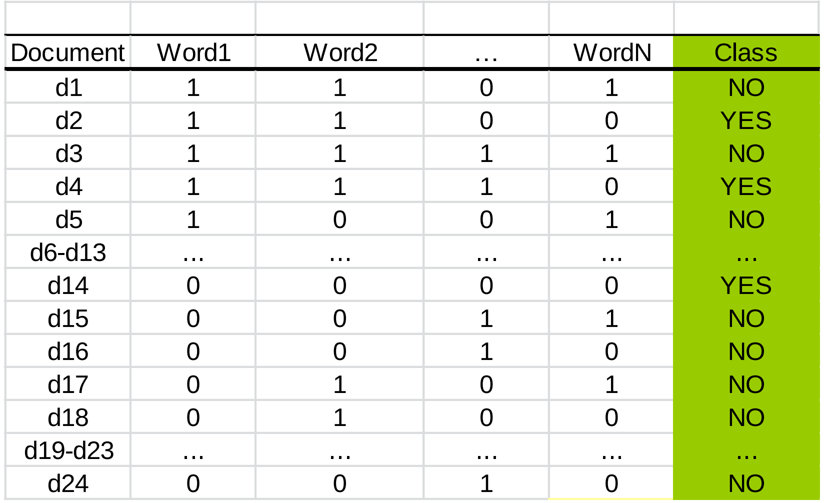


BoW vector construction



Document	Word1	Word2	…	WordN	Class
d1	1	1	0	1	NO
d2	1	1	0	0	YES
d3	1	1	1	1	NO
d4	1	1	1	0	YES
d5	1	0	0	1	NO
d6-d13	…	…	…	…	…
d14	0	0	0	0	YES
d15	0	0	1	1	NO
d16	0	0	1	0	NO
d17	0	1	0	1	NO
d18	0	1	0	0	NO
d19-d23	…	…	…	…	…
d24	0	0	1	0	NO

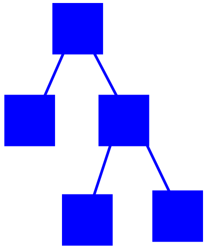
1. BoW features construction
2. Table of BoW vectors construction



Document	Word1	Word2	…	WordN	Class
d1	1	1	0	1	NO
d2	1	1	0	0	YES
d3	1	1	1	1	NO
d4	1	1	1	0	YES
d5	1	0	0	1	NO
d6-d13	…	…	…	…	…
d14	0	0	0	0	YES
d15	0	0	1	1	NO
d16	0	0	1	0	NO
d17	0	1	0	1	NO
d18	0	1	0	0	NO
d19-d23	…	…	…	…	…
d24	0	0	1	0	NO

Step 2

Data Mining



model, patterns, clusters,

…

Text Mining

- Feature construction
  - StopWords elimination
  - Stemming or lemmatization
  - Term construction by frequent N-Grams construction
  - Terms obtained from thesaurus (e.g., WordNet)

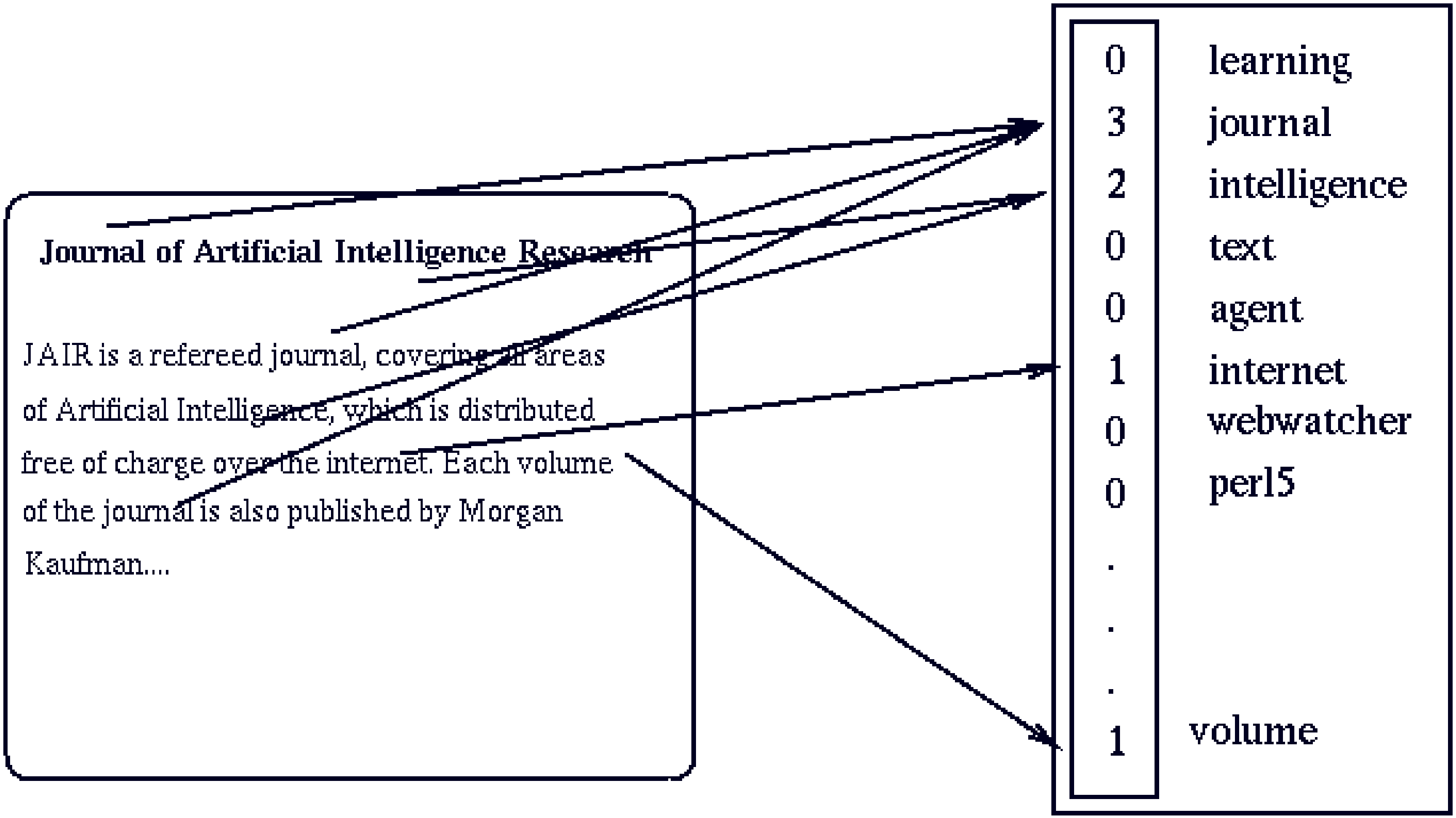
- BoW vector construction

- Mining of BoW vector table
  - Feature selection, Document similarity computation
  - Text mining: Categorization, Clustering, Summarization, …

Stemming and Lemmatization

- Different forms of the same word usually problematic for text data analysis
  - because they have different spelling and similar meaning (e.g. learns, learned, learning,…)
  - usually treated as completely unrelated words
- Stemming is a process of transforming a word into its stem
  - cutting off a suffix (eg., smejala -> smej)
- Lemmatization is a process of transforming a word into its normalized form
  - replacing the word, most often replacing a suffix (eg., smejala -> smejati)

Bag-of-Words document representation



Journal of Artificial Intelligence Research

JAIR is a refereed journal, covering all areas of Artificial Intelligence, which is distributed free of charge over the internet. Each volume of the journal is also published by Morgan Kaufman....

0	learning
3	journal
2	intelligence
0	text
0	agent
1	internet
0	webwatcher
0	perl5
.	
.	
.	
1	volume

Word weighting

- In bag-of-words representation each word is represented as a separate variable having numeric weight.
- The most popular weighting schema is normalized word frequency TFIDF:

$$tfidf(w) = tf.\log(\frac{N}{df(w)})$$

  - Tf(w) – term frequency (number of word occurrences in a document)
  - Df(w) – document frequency (number of documents containing the word)
  - N – number of all documents
  - Tfidf(w) – relative importance of the word in the document

The word is more important if it appears several times in a target document

The word is more important if it appears in less documents

Cosine similarity between document vectors

- Each document D is represented as a vector of TF-IDF weights
- Similarity between two vectors is estimated by the similarity between their vector representations (cosine of the angle between the two vectors):

$$Similarity\ (D_1, D_2) = \frac{\sum_i x_{1i} x_{2i}}{\sqrt{\sum_j x_j^2}\ \sqrt{\sum_k x_k^2}}$$

Talk outline

- Background and motivation
- Background technologies
  - Literature-based discovery
  - Text mining
- Cross-domain literature mining approaches
  - Outlier detection for cross-domain knowledge discovery
  - Cross-domain knowledge discovery with CrossBee
- Summary and conclusions
- CrossBee demo by Bojan Cestnik

Outlier detection

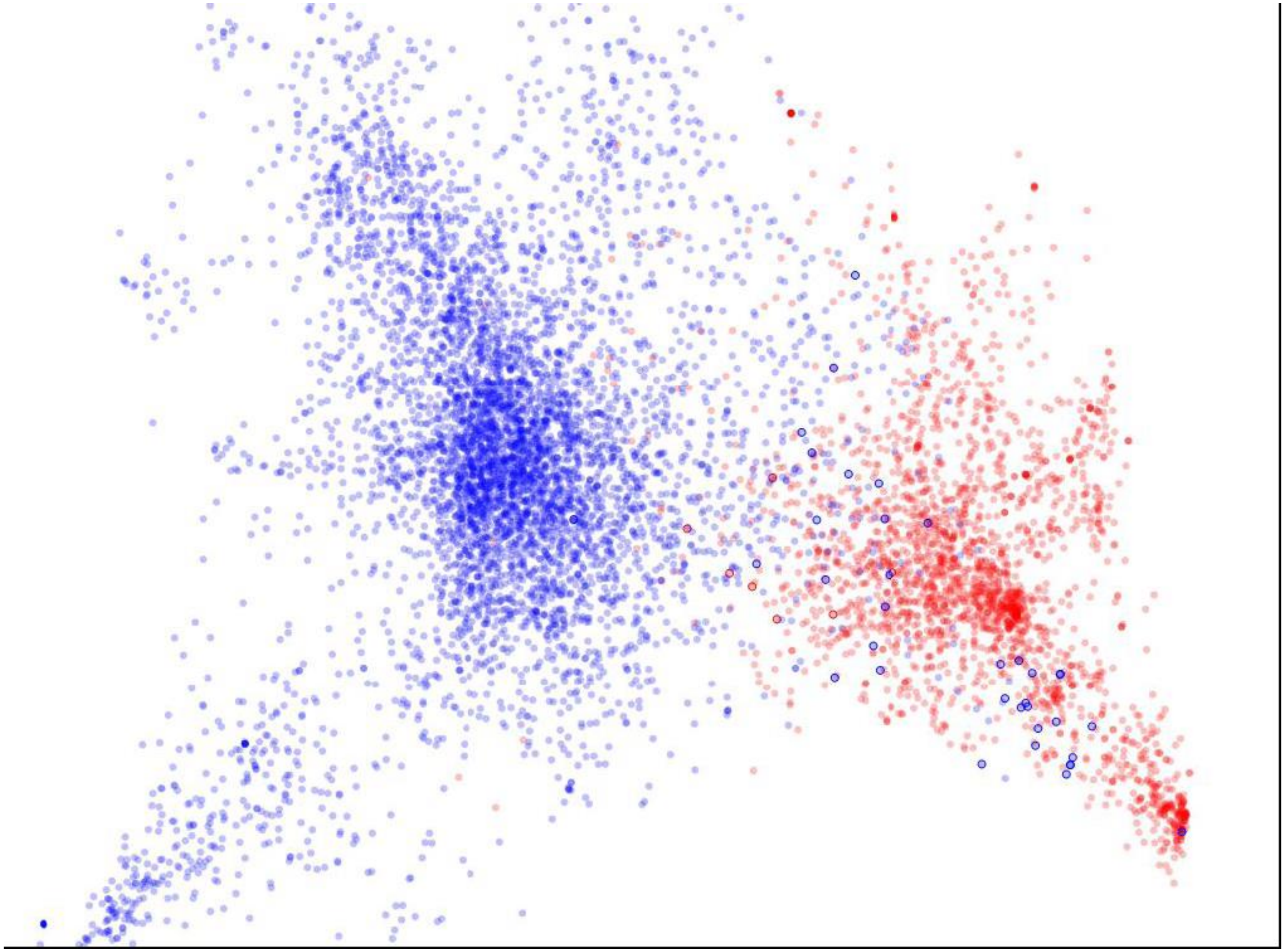
Outlier detection for cross-domain knowledge discovery

- The goal is to identify interesting **terms** or **concepts** which relate or link separate domains.

  ⇒ bridging terms (b-terms) / bridging concepts

- We explore the utility of outlier detection in the task of cross-domain bridging term discovery

Outlier detection for cross-domain knowledge discovery



2-dimensional projection of documents (about autism (red) and calcineurin (blue). Outlier documents are bolded for the user to easily spot them.

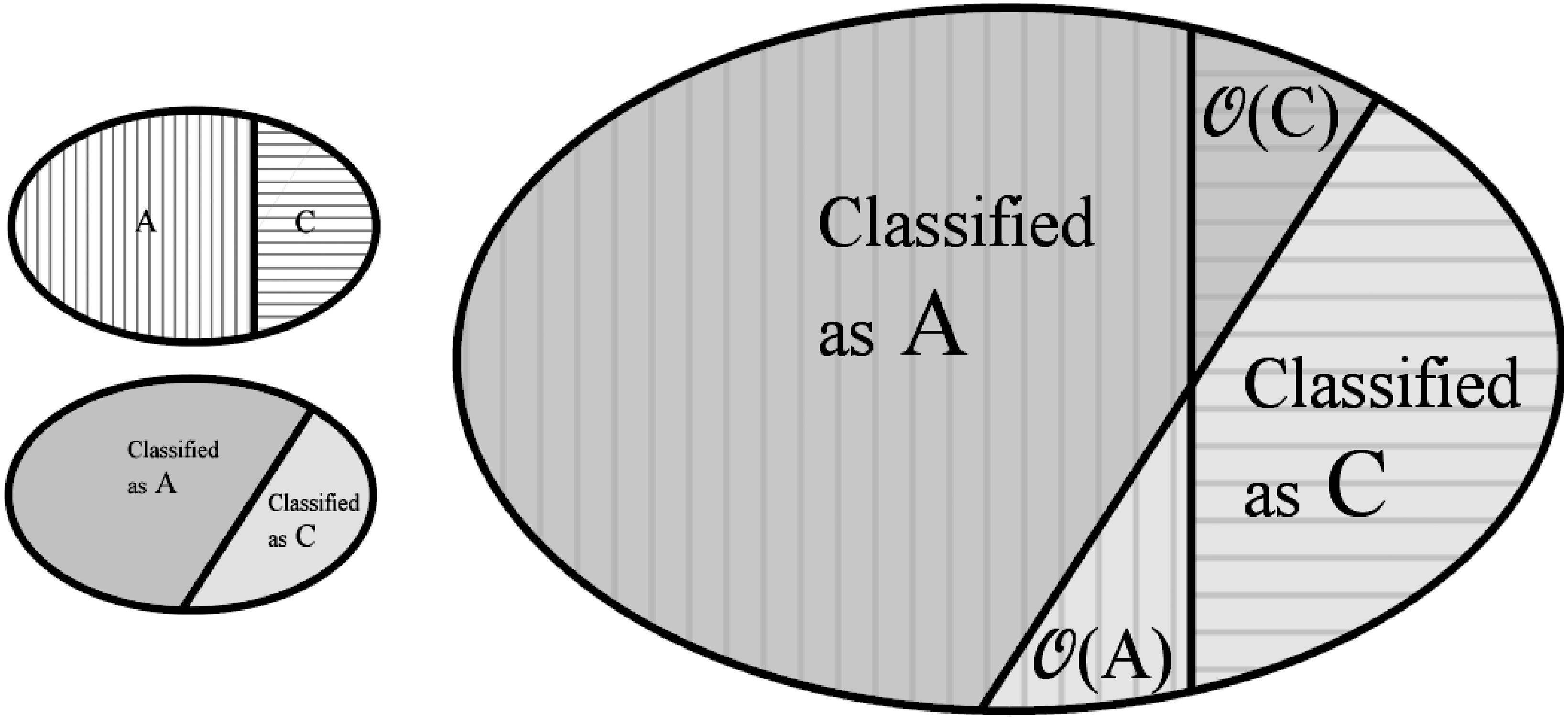
Our research has shown that most domain bridging terms appear in outlier documents.
(Lavrač, Sluban, Grčar, Juršič 2010)

Outlier detection for cross-domain knowledge discovery

- Outlier document and bridging term detection
- Three approaches
  - Outlier detection through noise/outlier detection and ranking with NoiseRank
  - Outlier document detection through document clustering with OntoGen
  - Outlier document and outlier term detection using Banded matrices (current work, out of scope of this presentation)

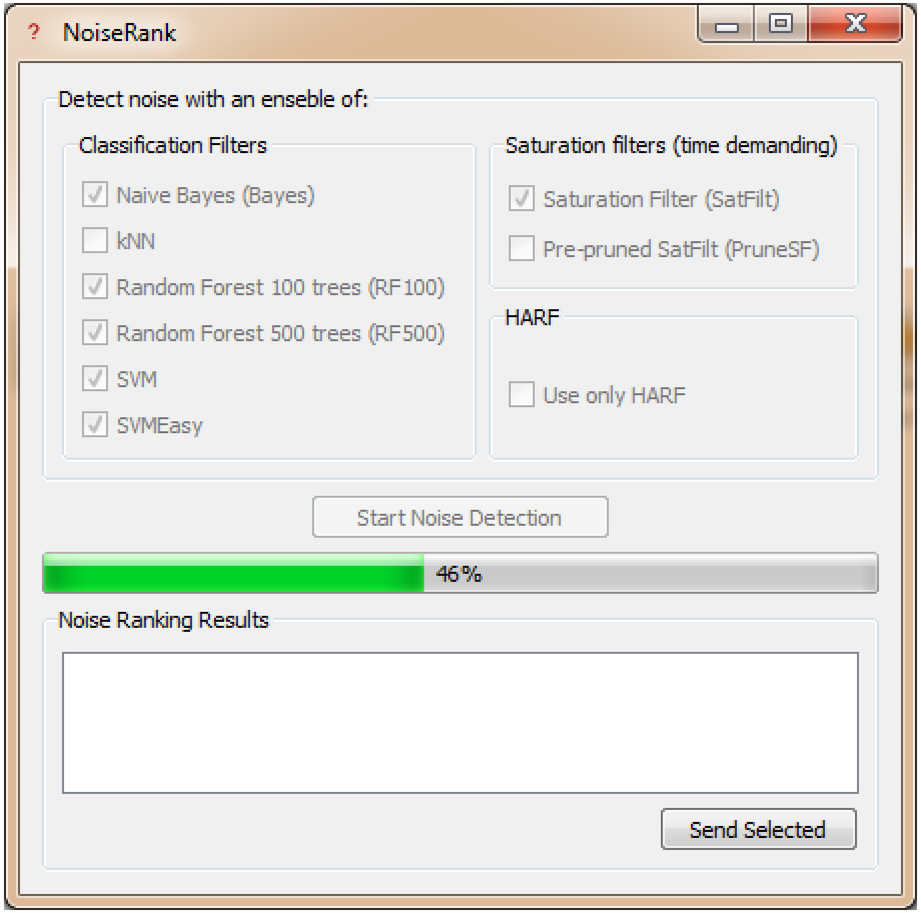
Detecting outlier documents

- By classification noise detection on a domain pair dataset, assuming two separate document corpora A and C



NoiseRank: Ensemble-based noise and outlier detection

- Misclassified document detection by an ensemble of diverse classifiers (e.g., Naive Bayes, Random Forest, SVM, … classifiers)
- Ranking of misclassified documents by “voting” of classifiers



Detect noise with an enseble of:

Classification Filters

☑ Naive Bayes (Bayes)
☐ kNN
☑ Random Forest 100 trees (RF100)
☑ Random Forest 500 trees (RF500)
☑ SVM
☑ SVMEasy

Saturation filters (time demanding)

☑ Saturation Filter (SatFilt)
☐ Pre-pruned SatFilt (PruneSF)

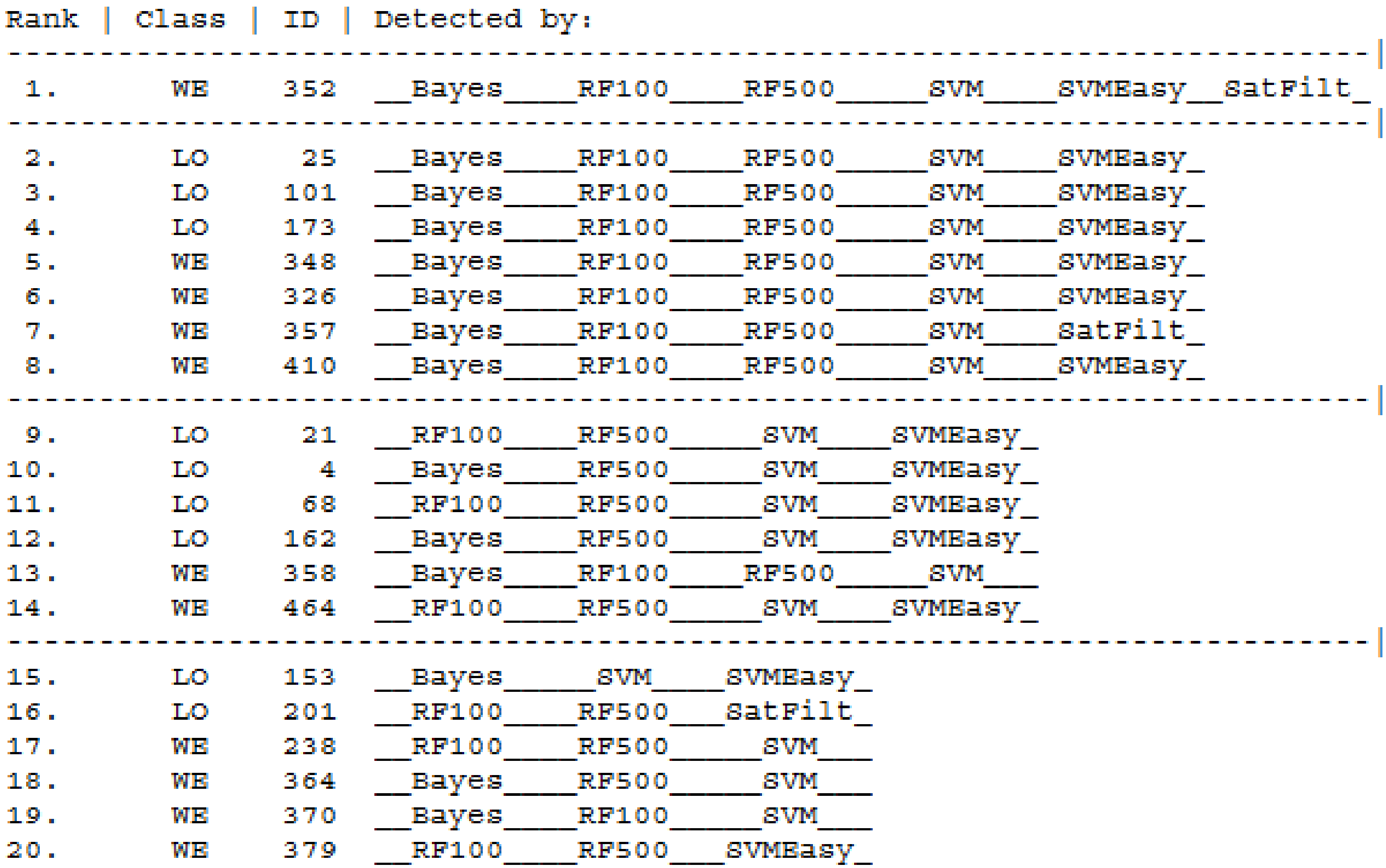
HARF

☐ Use only HARF

Noise Ranking Results

NoiseRank on news articles

Articles on Kenyan elections: local vs. Western media



```
Rank | Class | ID | Detected by:
-------------------------------------------------------------------------|
 1.      WE    352  __Bayes____RF100____RF500_____SVM____SVMEasy__SatFilt_
-------------------------------------------------------------------------|
 2.      LO     25  __Bayes____RF100____RF500_____SVM____SVMEasy_
 3.      LO    101  __Bayes____RF100____RF500_____SVM____SVMEasy_
 4.      LO    173  __Bayes____RF100____RF500_____SVM____SVMEasy_
 5.      WE    348  __Bayes____RF100____RF500_____SVM____SVMEasy_
 6.      WE    326  __Bayes____RF100____RF500_____SVM____SVMEasy_
 7.      WE    357  __Bayes____RF100____RF500_____SVM____SatFilt_
 8.      WE    410  __Bayes____RF100____RF500_____SVM____SVMEasy_
-------------------------------------------------------------------------|
 9.      LO     21  __RF100____RF500_____SVM____SVMEasy_
10.      LO      4  __Bayes____RF500_____SVM____SVMEasy_
11.      LO     68  __RF100____RF500_____SVM____SVMEasy_
12.      LO    162  __Bayes____RF500_____SVM____SVMEasy_
13.      WE    358  __Bayes____RF100____RF500_____SVM___
14.      WE    464  __RF100____RF500_____SVM____SVMEasy_
-------------------------------------------------------------------------|
15.      LO    153  __Bayes_____SVM____SVMEasy_
16.      LO    201  __RF100____RF500___SatFilt_
17.      WE    238  __RF100____RF500_____SVM___
18.      WE    364  __Bayes____RF500_____SVM___
19.      WE    370  __Bayes____RF100_____SVM___
20.      WE    379  __RF100____RF500___SVMEasy_
```

NoiseRank on news articles

- **Article 352: Out of topic** The article was later indeed removed from the corpus used for further linguistic analysis, since it is not about Kenya(ns) or the socio-political climate but about British tourists or expatriates’ misfortune.

- **Article 173: Guest journalist** Wrongly classified because it could be regarded as a “Western article” among the local Kenyan press. The author does not have the cultural sensitivity or does not follow the editorial guidelines requiring to be careful when mentioning words like tribe in negative contexts. One could even say that he has a kind of “Western” writing style.

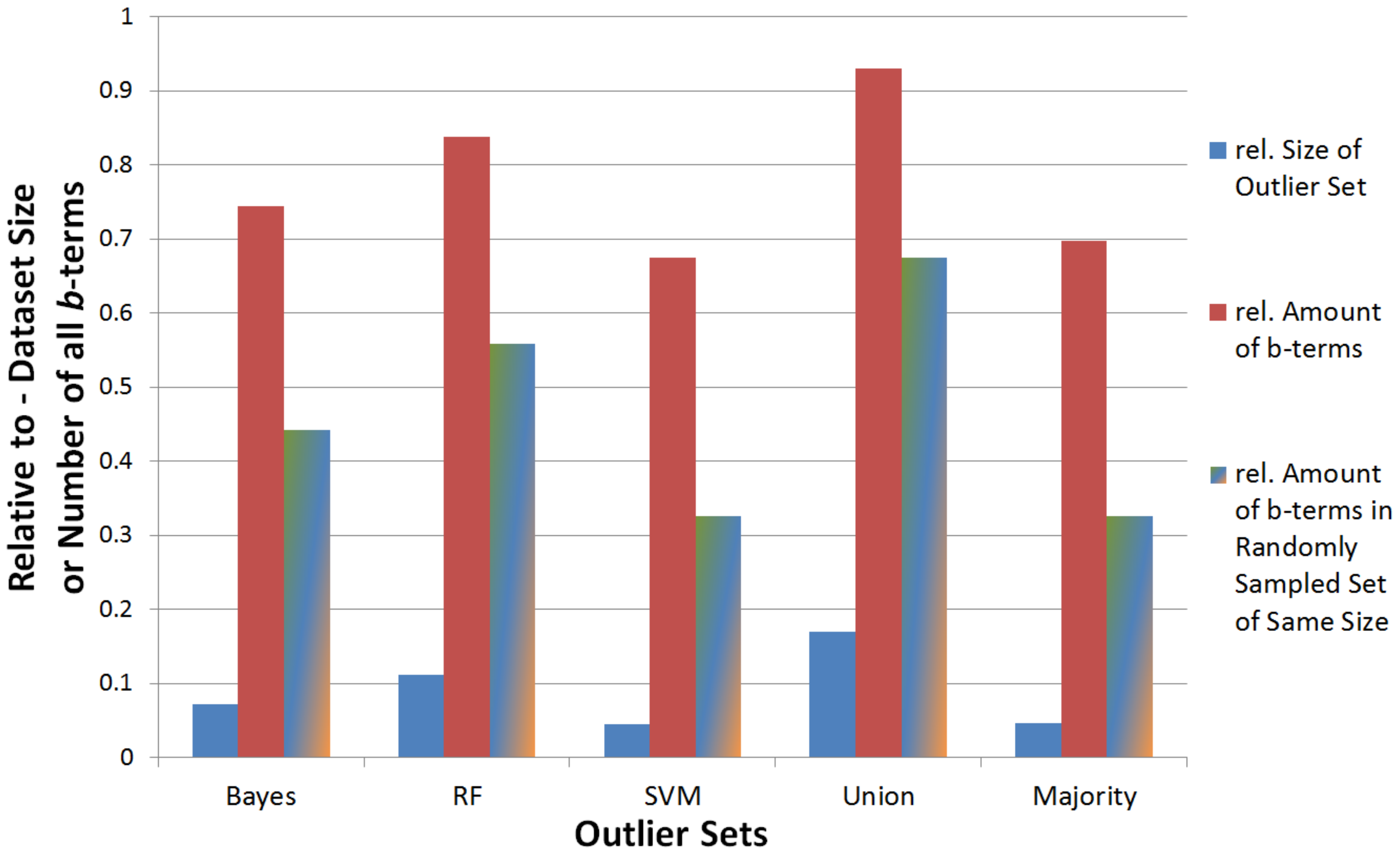
Experimental evaluation

- 2 datasets retrieved form the *PubMed database**
  - Migraine-Magnesium (8,058 docs, 43 known *b*-terms)
  - Autism-Calcineurine (15,243 docs, 13 known *b*-terms)
- Ensemble consisting of three elementary classifiers
- Evaluating the cross-domain linking potential of outlier documents by:
  - Number of *b*-terms appearing in the detected outlier document sets
  - Ratio of *b*-terms in an outlier set against its size
  - Increase in relative frequency of *b*-terms in outlier document sets

* PubMed: http://www.ncbi.nlm.nih.gov/pubmed

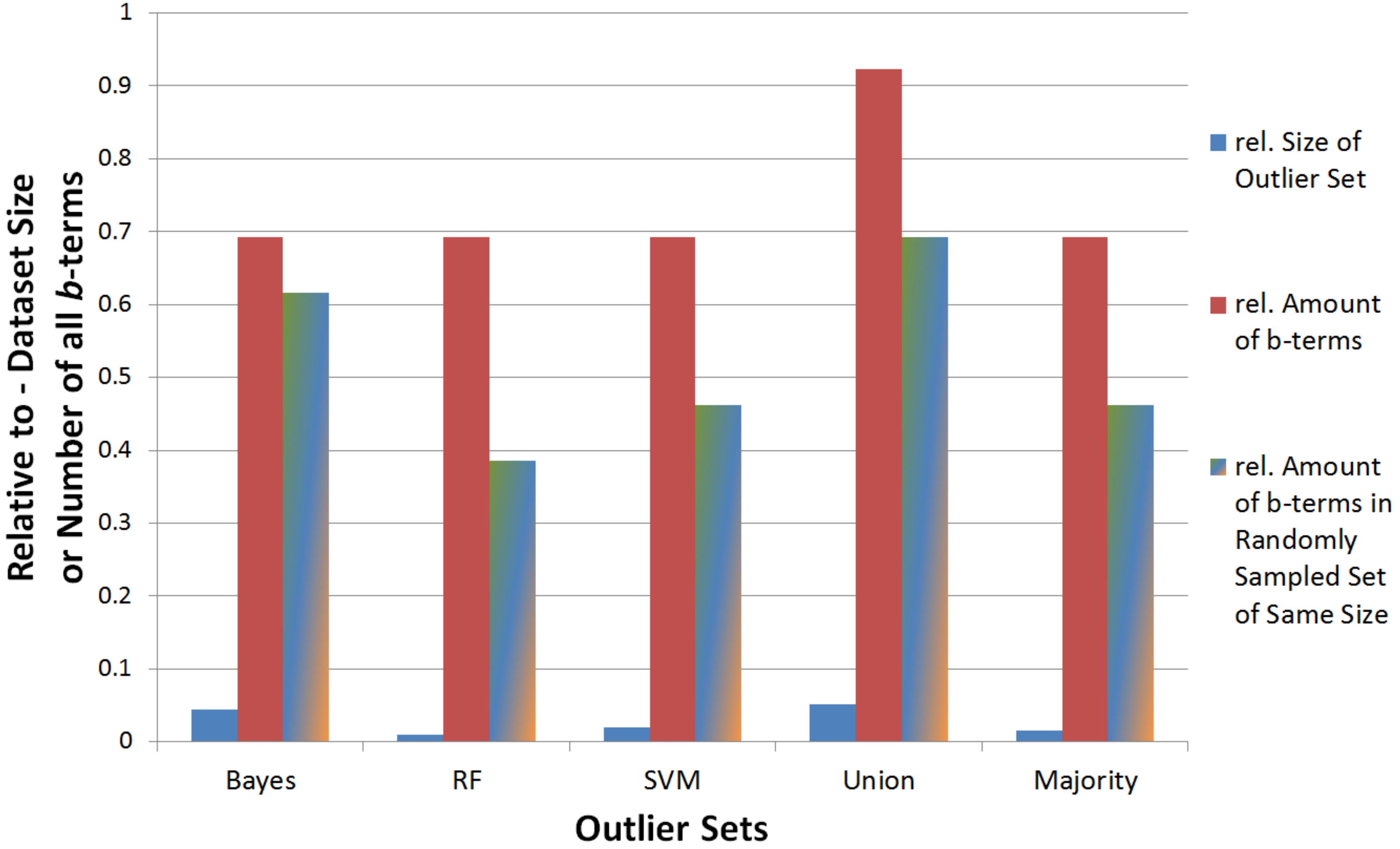
b-terms in outlier sets

- On the Migraine-Magnesium domain pair



b-terms in outlier sets

- On the Autism-Calcineurine domain pair



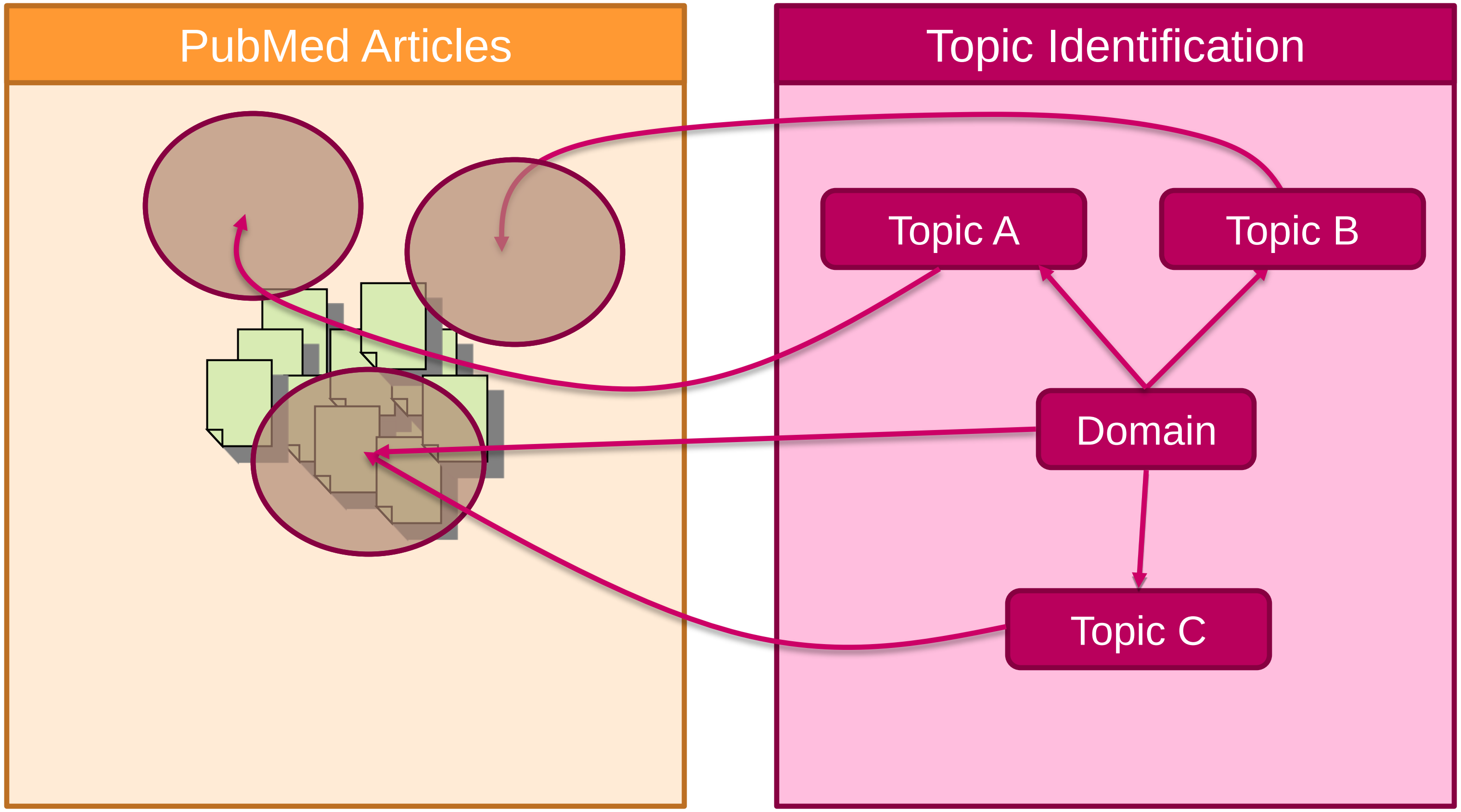
Outlier detection for cross-domain knowledge discovery

- Outlier document and bridging term detection
- Three approaches
  - Outlier detection through noise/outlier detection and ranking with NoiseRank
  - Outlier document detection through document clustering with OntoGen
  - Outlier document and outlier term detection using Banded matrices (current work, out of scope of this presentation)

Document clustering

- Clustering is a process of finding natural groups in data in a unsupervised way (no class labels pre-assigned to documents)
- Document similarity is used
- Most popular clustering methods:
  - K-Means clustering
  - Agglomerative hierarchical clustering
  - EM (Gaussian Mixture)
  - …

Document clustering with OntoGen



Slide adapted from D. Mladenić, JSI

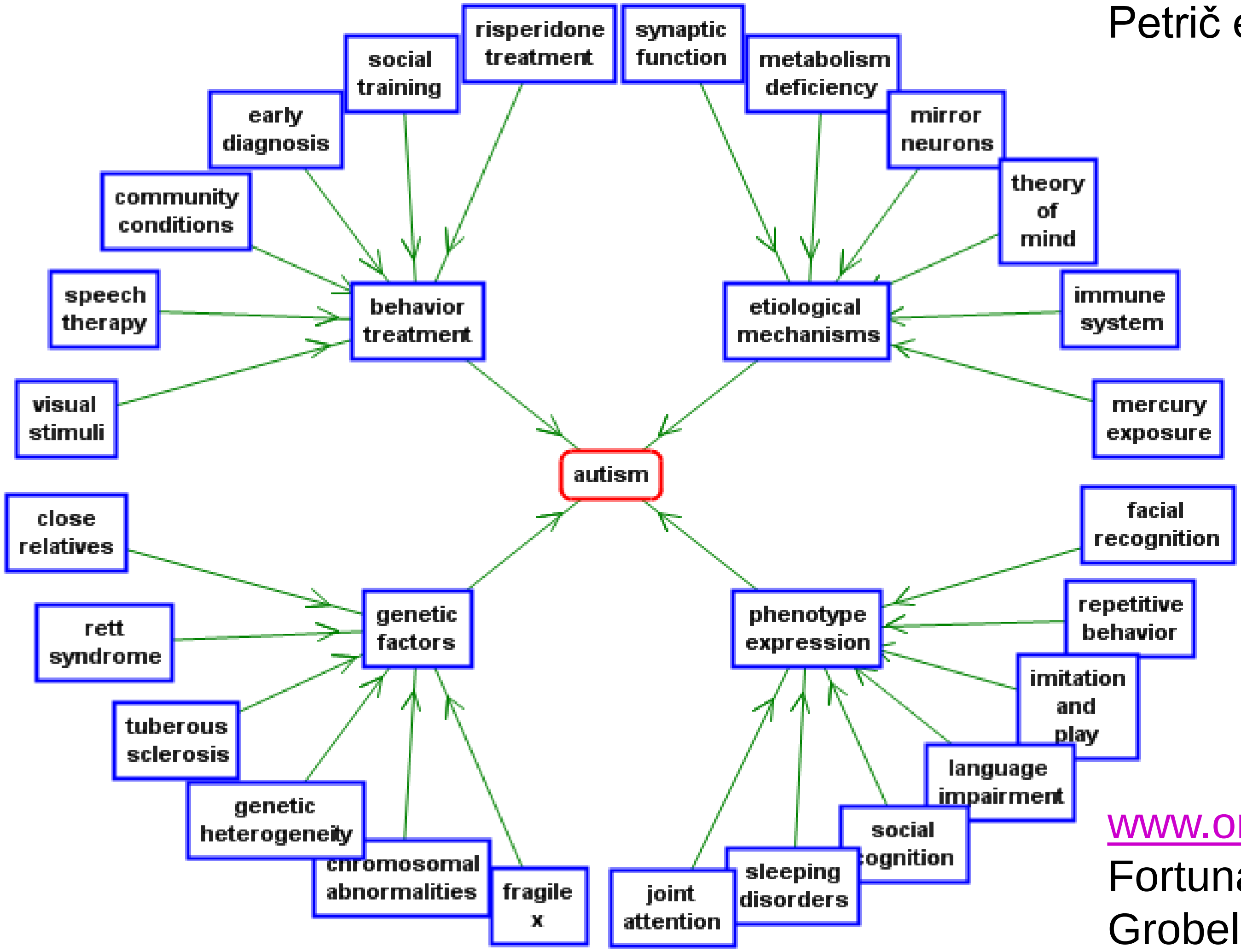
K-Means clustering in OntoGen

OntoGen uses k-Means clustering for semi-automated topic ontology construction

- Given:
  - set of documents (eg., word-vectors with TFIDF),
  - distance measure (eg., cosine similarity)
  - K - number of groups
- For each group initialize its centroid with a random document
- While not converging
  - each document is assigned to the nearest group (represented by its centroid)
  - for each group calculate new centroid (group mass point, average document in the group)

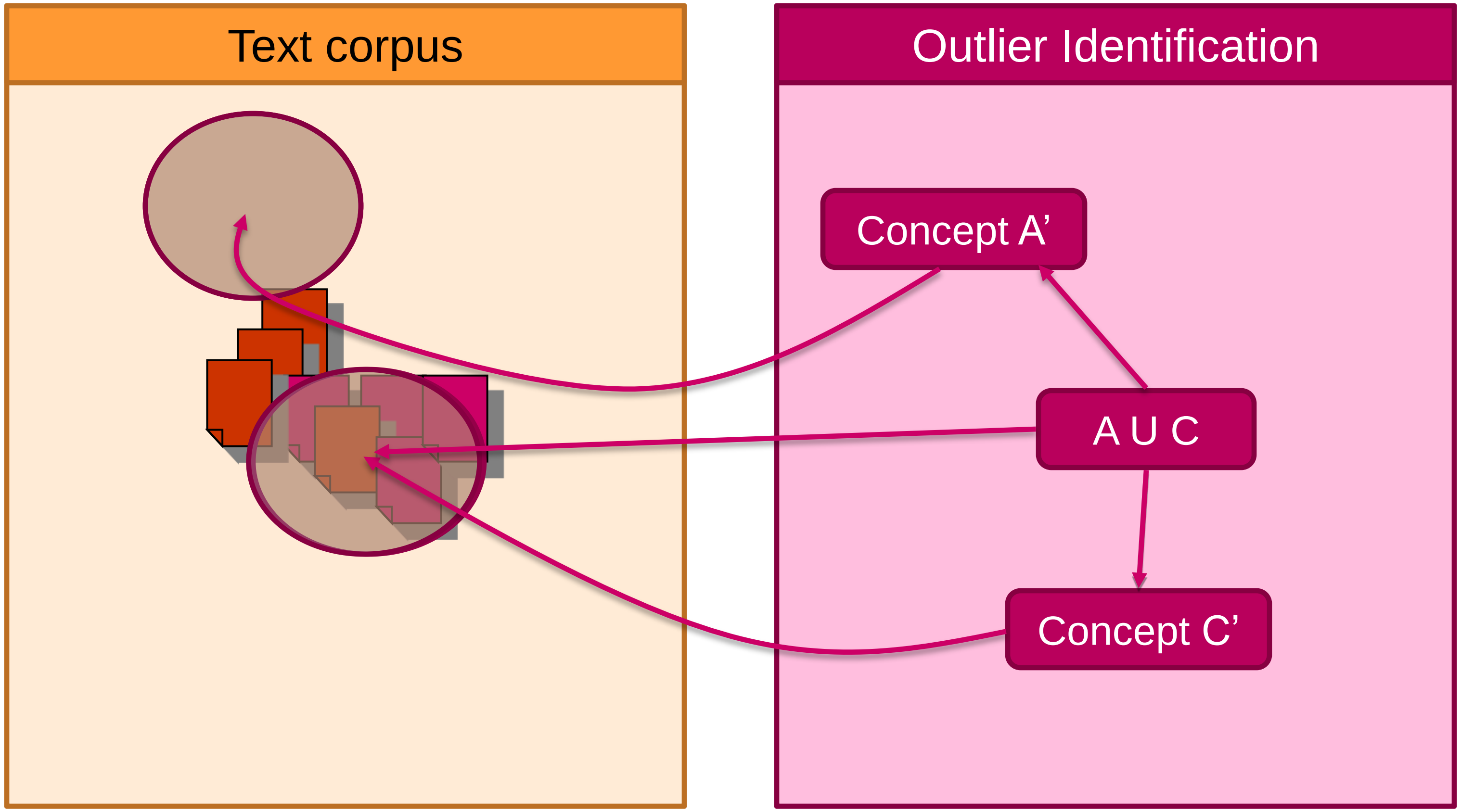
Using OntoGen for clustering PubMed articles on autism

Work by
Petrič et al. 2009



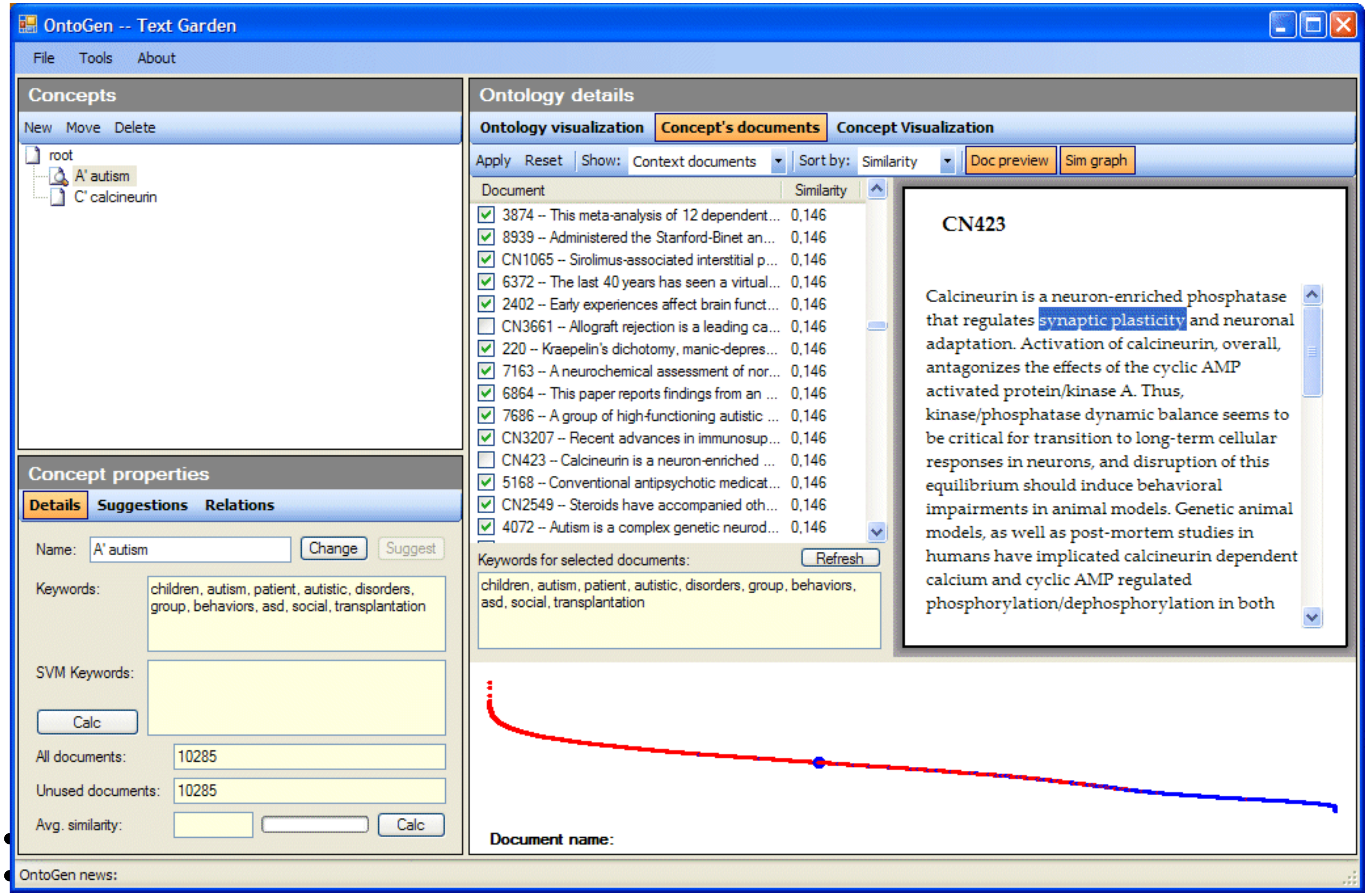
www.ontogen.si
Fortuna, Mladenić,
Grobelnik 2006

Using OntoGen for outlier document identification



Slide adapted from D. Mladenić, JSI

Results on autism-calcineurin: Outlier calcineurin document CN423



Show: Context documents

Sort by: Similarity

	Document	Similarity
☑	3874 -- This meta-analysis of 12 dependent...	0,146
☑	8939 -- Administered the Stanford-Binet an...	0,146
☑	CN1065 -- Sirolimus-associated interstitial p...	0,146
☑	6372 -- The last 40 years has seen a virtual...	0,146
☑	2402 -- Early experiences affect brain funct...	0,146
☐	CN3661 -- Allograft rejection is a leading ca...	0,146
☑	220 -- Kraepelin's dichotomy, manic-depres...	0,146
☑	7163 -- A neurochemical assessment of nor...	0,146
☑	6864 -- This paper reports findings from an ...	0,146
☑	7686 -- A group of high-functioning autistic ...	0,146
☑	CN3207 -- Recent advances in immunosup...	0,146
☐	CN423 -- Calcineurin is a neuron-enriched ...	0,146
☑	5168 -- Conventional antipsychotic medicat...	0,146
☑	CN2549 -- Steroids have accompanied oth...	0,146
☑	4072 -- Autism is a complex genetic neurod...	0,146

Keywords for selected documents: children, autism, patient, autistic, disorders, group, behaviors, asd, social, transplantation

CN423

Calcineurin is a neuron-enriched phosphatase that regulates synaptic plasticity and neuronal adaptation. Activation of calcineurin, overall, antagonizes the effects of the cyclic AMP activated protein/kinase A. Thus, kinase/phosphatase dynamic balance seems to be critical for transition to long-term cellular responses in neurons, and disruption of this equilibrium should induce behavioral impairments in animal models. Genetic animal models, as well as post-mortem studies in humans have implicated calcineurin dependent calcium and cyclic AMP regulated phosphorylation/dephosphorylation in both

Name: A' autism

Keywords: children, autism, patient, autistic, disorders, group, behaviors, asd, social, transplantation

SVM Keywords:

All documents: 10285

Unused documents: 10285

Avg. similarity:

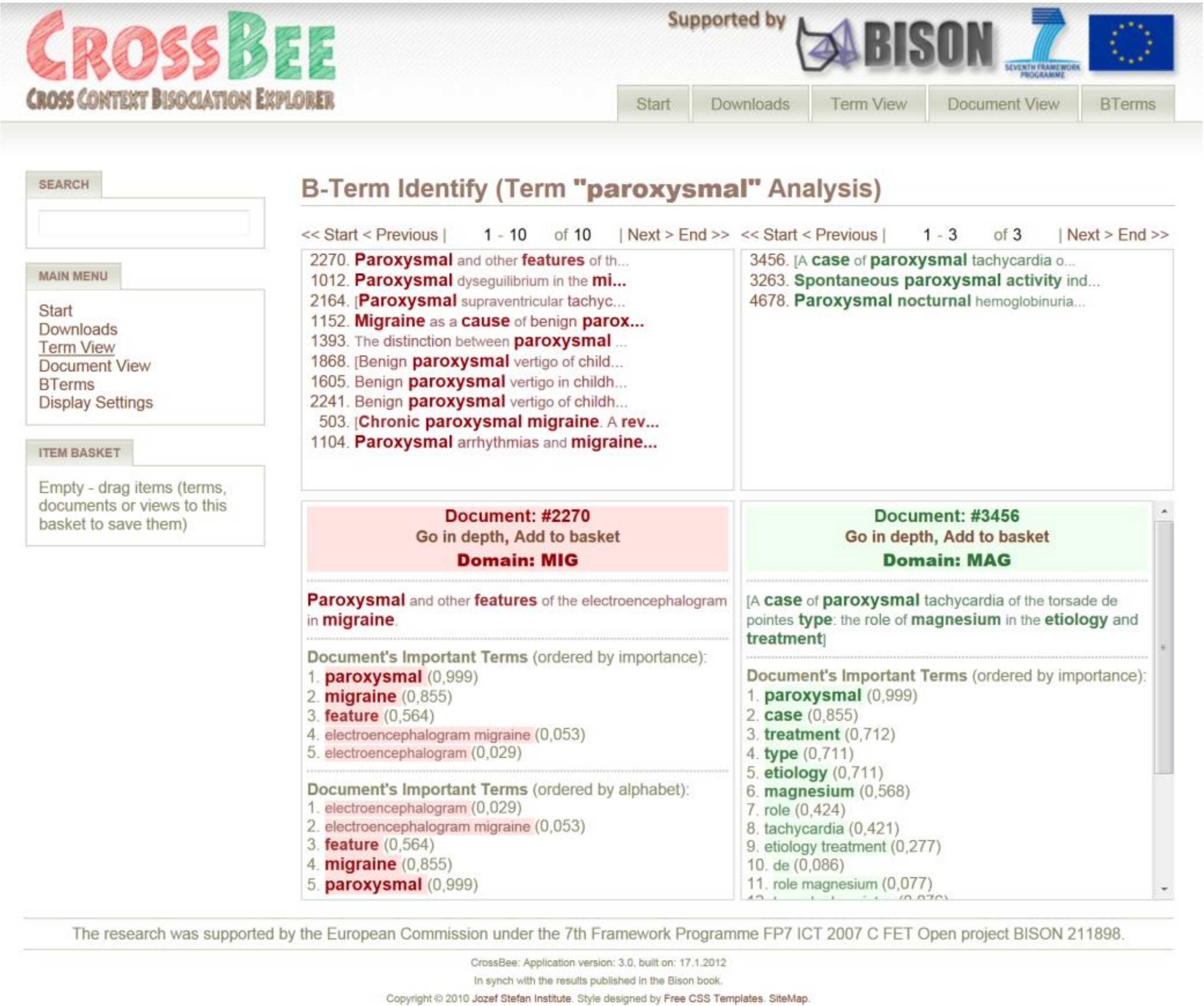
Document name:

Work by
Petrič et al. 2010

Talk outline

- Background and motivation
- Background technologies
  - Literature-based discovery
  - Text mining
- Cross-domain literature mining approaches
  - Outlier detection for cross-domain knowledge discovery
  - Cross-domain knowledge discovery with CrossBee
- Summary and conclusions
- CrossBee demo by Bojan Cestnik

CrossBee: Cross Context Bisociation Explorer



SEARCH

MAIN MENU

Start
Downloads
Term View
Document View
BTerms
Display Settings

ITEM BASKET

Empty - drag items (terms, documents or views to this basket to save them)

B-Term Identify (Term "paroxysmal" Analysis)

<< Start < Previous | 1 - 10 of 10 | Next > End >>

2270. Paroxysmal and other features of th...
1012. Paroxysmal dyseguilibrium in the mi...
2164. [Paroxysmal supraventricular tachyc...
1152. Migraine as a cause of benign parox...
1393. The distinction between paroxysmal ...
1868. [Benign paroxysmal vertigo of child...
1605. Benign paroxysmal vertigo in childh...
2241. Benign paroxysmal vertigo of childh...
503. [Chronic paroxysmal migraine. A rev...
1104. Paroxysmal arrhythmias and migraine...

<< Start < Previous | 1 - 3 of 3 | Next > End >>

3456. [A case of paroxysmal tachycardia o...
3263. Spontaneous paroxysmal activity ind...
4678. Paroxysmal nocturnal hemoglobinuria...

Document: #2270
Go in depth, Add to basket
Domain: MIG

Paroxysmal and other features of the electroencephalogram in migraine.

Document's Important Terms (ordered by importance):
1. paroxysmal (0,999)
2. migraine (0,855)
3. feature (0,564)
4. electroencephalogram migraine (0,053)
5. electroencephalogram (0,029)

Document's Important Terms (ordered by alphabet):
1. electroencephalogram (0,029)
2. electroencephalogram migraine (0,053)
3. feature (0,564)
4. migraine (0,855)
5. paroxysmal (0,999)

Document: #3456
Go in depth, Add to basket
Domain: MAG

[A case of paroxysmal tachycardia of the torsade de pointes type: the role of magnesium in the etiology and treatment]

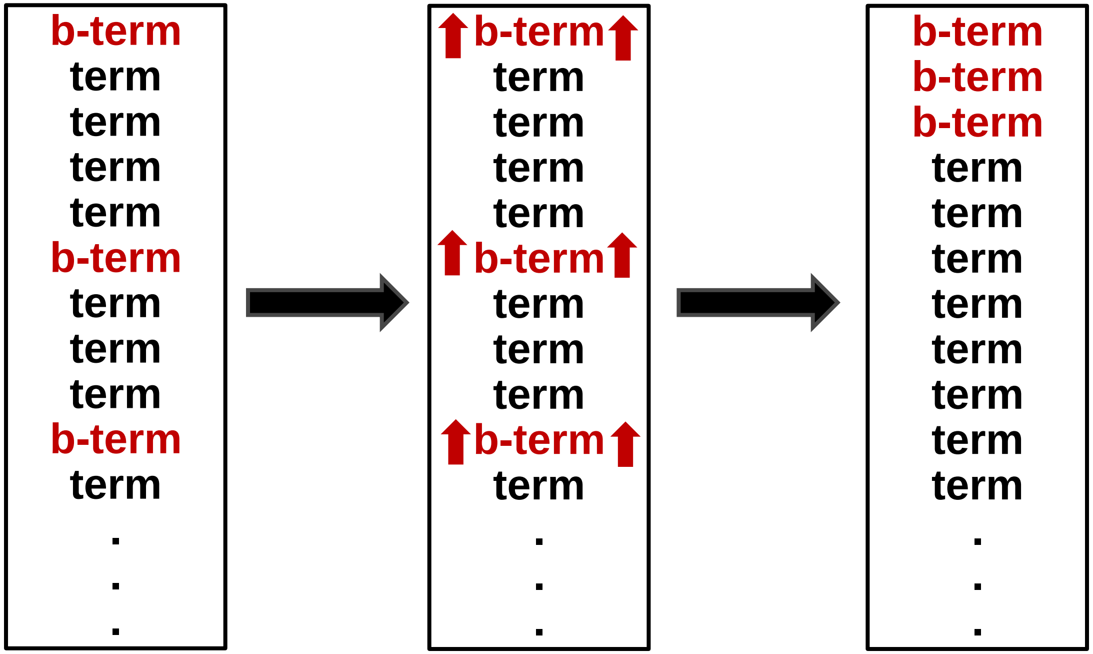
Document's Important Terms (ordered by importance):
1. paroxysmal (0,999)
2. case (0,855)
3. treatment (0,712)
4. type (0,711)
5. etiology (0,711)
6. magnesium (0,568)
7. role (0,424)
8. tachycardia (0,421)
9. etiology treatment (0,277)
10. de (0,086)
11. role magnesium (0,077)

The research was supported by the European Commission under the 7th Framework Programme FP7 ICT 2007 C FET Open project BISON 211898.

CrossBee: Application version: 3.0, built on: 17.1.2012
In synch with the results published in the Bison book.
Copyright © 2010 Jozef Stefan Institute. Style designed by Free CSS Templates. SiteMap.

Problem definition

Goal: Develop a term ranking methodology that ranks high all the terms which have high bisociation potential (denoted as *bridging* terms or *b-terms*)

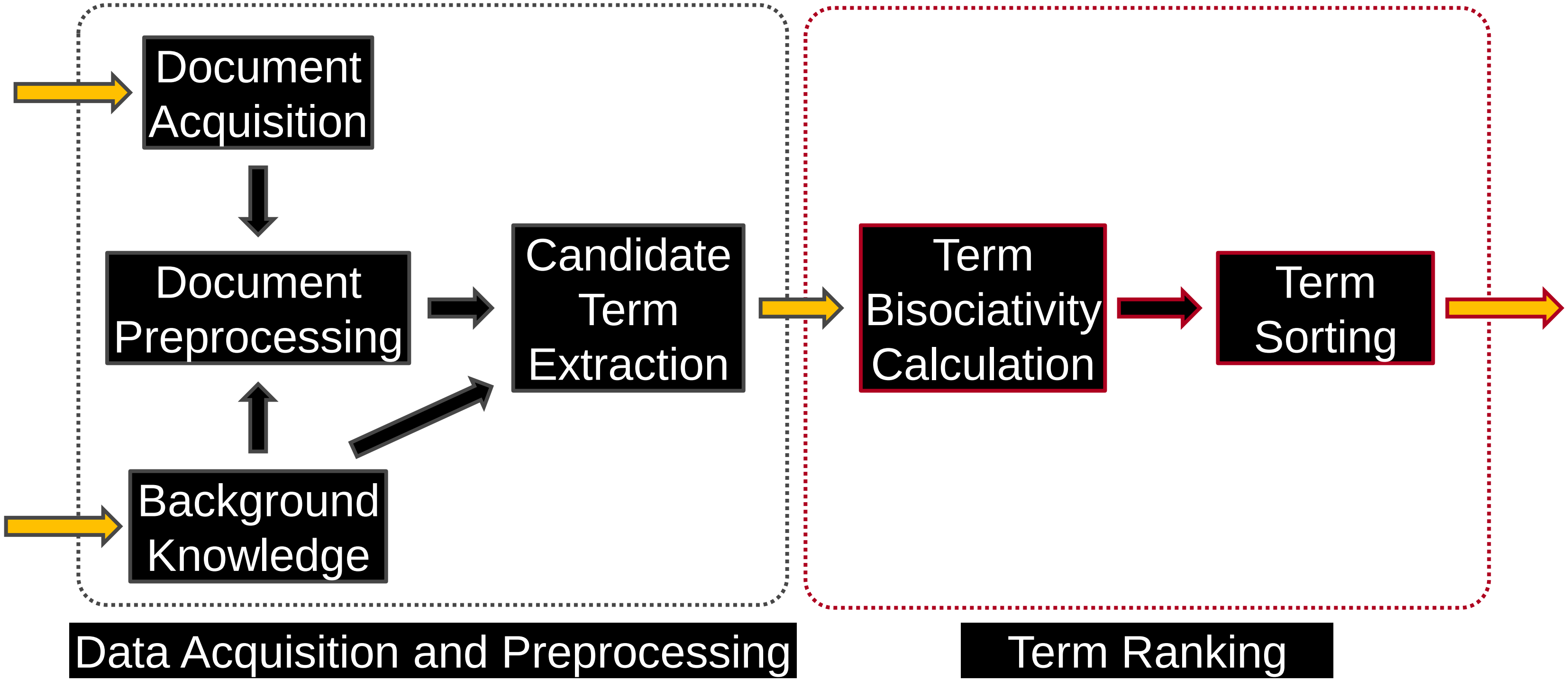


b-term
term
term
term
term
b-term
term
term
term
b-term
term
.
.
.

⬆b-term⬆
term
term
term
term
⬆b-term⬆
term
term
term
⬆b-term⬆
term
.
.
.

b-term
b-term
b-term
term
term
term
term
term
term
term
term
.
.
.

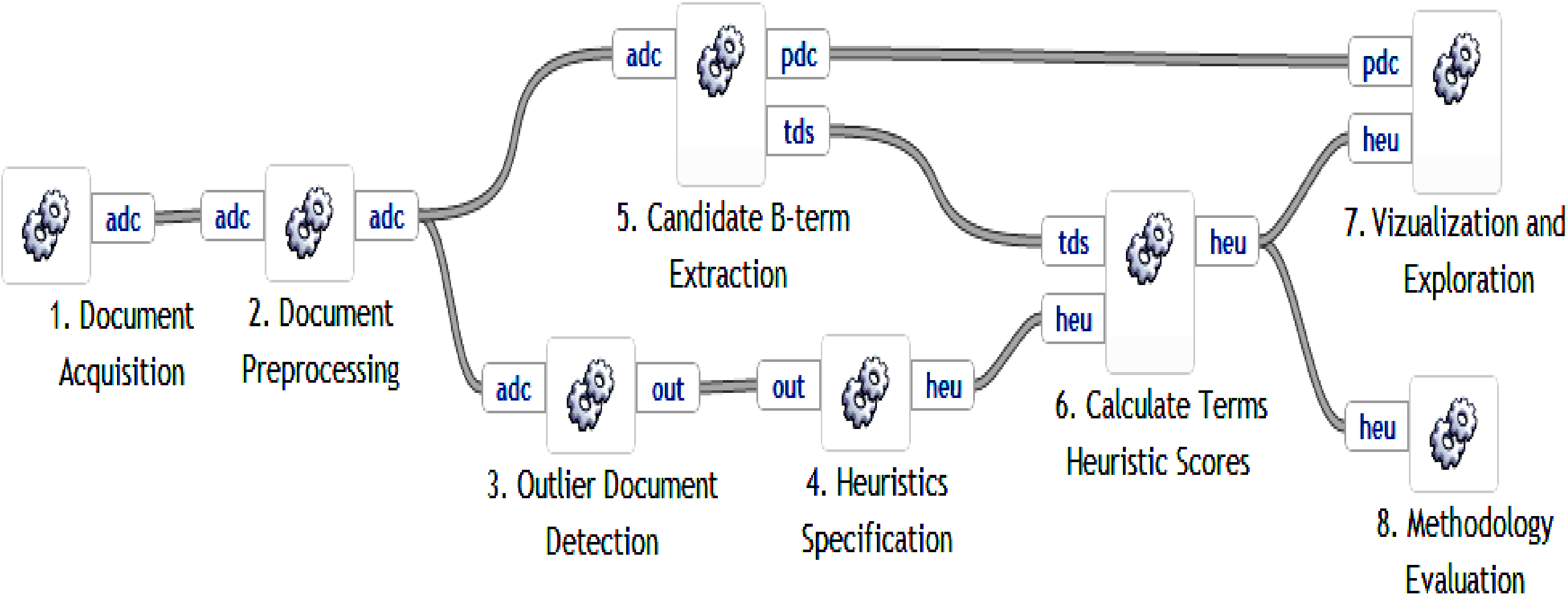
CrossBee: Methodology overview


Document Acquisition
Document Preprocessing
Background Knowledge
Candidate Term Extraction
Term Bisociativity Calculation
Term Sorting
Data Acquisition and Preprocessing
Term Ranking

Incorporating available background knowledge

- Vocabularies: e.g. for word/term filtering
- Ontologies: e.g. for enriching documents term sets

Methodology implementation


adc
adc
adc
adc
pdc
tds
pdc
heu
5. Candidate B-term Extraction
7. Vizualization and Exploration
1. Document Acquisition
2. Document Preprocessing
tds
heu
heu
6. Calculate Terms Heuristic Scores
adc
out
out
heu
heu
3. Outlier Document Detection
4. Heuristics Specification
8. Methodology Evaluation

Methodology implementation in ClowdFlows browser based service oriented data mining platform, clowdflows.net

Data acquisition and preprocessing

- Document acquisition from the Web
  - Acquiring documents from. PubMed
  - Snippets returned from web search engines
  - Crawling the Internet and gathering documents from web pages
- Document preprocessing
  - Tokenization
  - Stopwords removal
  - Stemming or lemmatization: LemmaGen
  - Part of speech tagging or syntactic parsing
- Candidate term extraction
  - Frequent n-grams in preprocessed documents

Term ranking

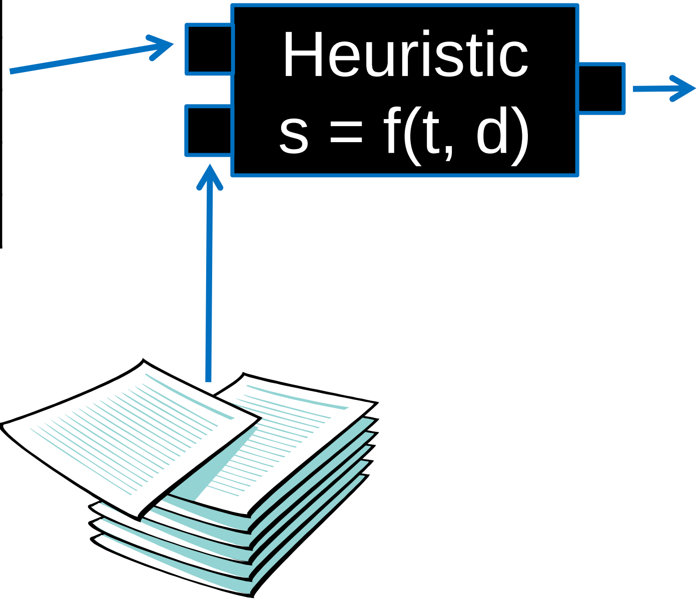
- Term ranking:
  - Assign scores to all the terms
  - Sort the terms according to the assigned scores
- How to assign scores to terms?
  - Using a heuristic function that estimates the probability that a term is b-term
- How to construct the “optimal” heuristic using training data?
  1. Create several promising heuristics
  2. Evaluate the constructed heuristics on a training dataset
  3. Construct the ensemble heuristic using the best individual heuristics
  4. Use the ensemble heuristic for scoring the terms

Heuristic function

- Input: a term with its statistic properties calculated from texts
- Output: a number [0,1] which ranks the term (its probability of being a b-term)

Ideal heuristic: such that ranks all true b-terms very high and all the others lower

combination
associate
cortical spread depression
efficacy safety
5 hydroxytryptamine receptor
accumulate

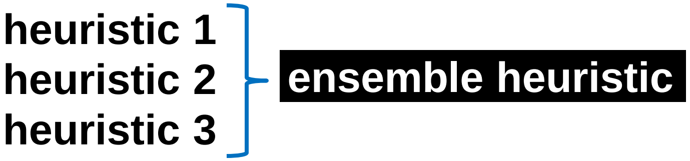


combination	0.154
associate	0,759
cortical spread depression	0,666
efficacy safety	0,311
5 hydroxytryptamine receptor	0,900
accumulate	0,486

Bisociation potential heuristics

- Heuristics can be grouped based on:
  - frequency (variations of the term occurrences)
    - $freqTerm(t) = countTerm_{D_u}(t)$: term frequency across both domains
  - tf-idf (combinations of tf-idf weights of a term)
    - $tfidfDomnProd(t) = tfidf_{D_1}(t) \cdot tfidf_{D_2}(t)$: product of a term’s importance in both domains
  - similarity (similarity of a term to the average terms)
  - outliers (frequency of a term in documents at the border of the two domains)
    - $outFreqRelRF(t) = \frac{countTerm_{D_{RF}}(t)}{countTerm_{D_u}(t)}$: relative frequency in RF outlier set

Ensemble heuristic



heuristic 1
heuristic 2
heuristic 3

ensemble heuristic

heuristic 1

term 1	0,149
term 2	0,759
term 3	0,900
term 4	0,666
term 5	0,311
term 6	0,071
term 7	0,175
term 8	0,637
term 9	0,429
.	.
.	.
.	.

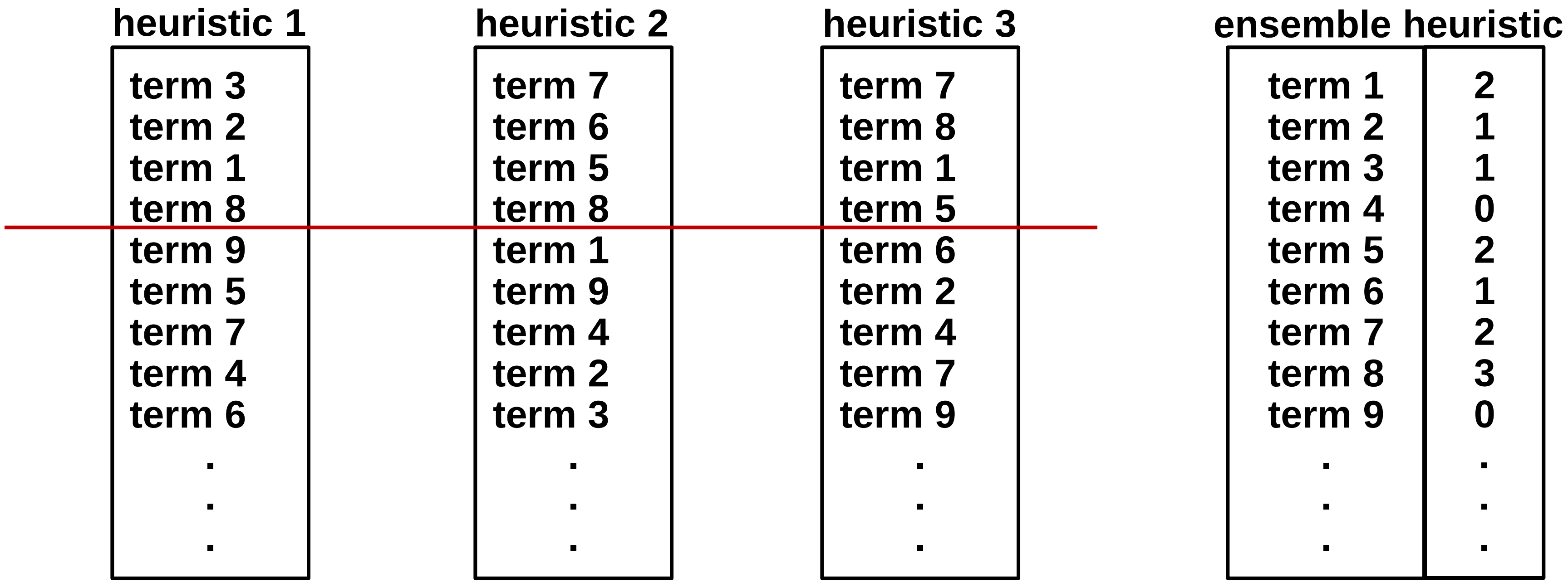
heuristic 2

term 1	0,429
term 2	0,149
term 3	0,071
term 4	0,175
term 5	0,637
term 6	0,759
term 7	0,970
term 8	0,636
term 9	0,311
.	.
.	.
.	.

heuristic 3

term 1	0,680
term 2	0,311
term 3	0,071
term 4	0,175
term 5	0,637
term 6	0,429
term 7	0,149
term 8	0,759
term 9	0,980
.	.
.	.
.	.

Ensemble heuristic



heuristic 1	heuristic 2	heuristic 3	ensemble heuristic	
term 3	term 7	term 7	term 1	2
term 2	term 6	term 8	term 2	1
term 1	term 5	term 1	term 3	1
term 8	term 8	term 5	term 4	0
term 9	term 1	term 6	term 5	2
term 5	term 9	term 2	term 6	1
term 7	term 4	term 4	term 7	2
term 4	term 2	term 7	term 8	3
term 6	term 3	term 9	term 9	0
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Ensemble heuristic

final ensemble heuristic

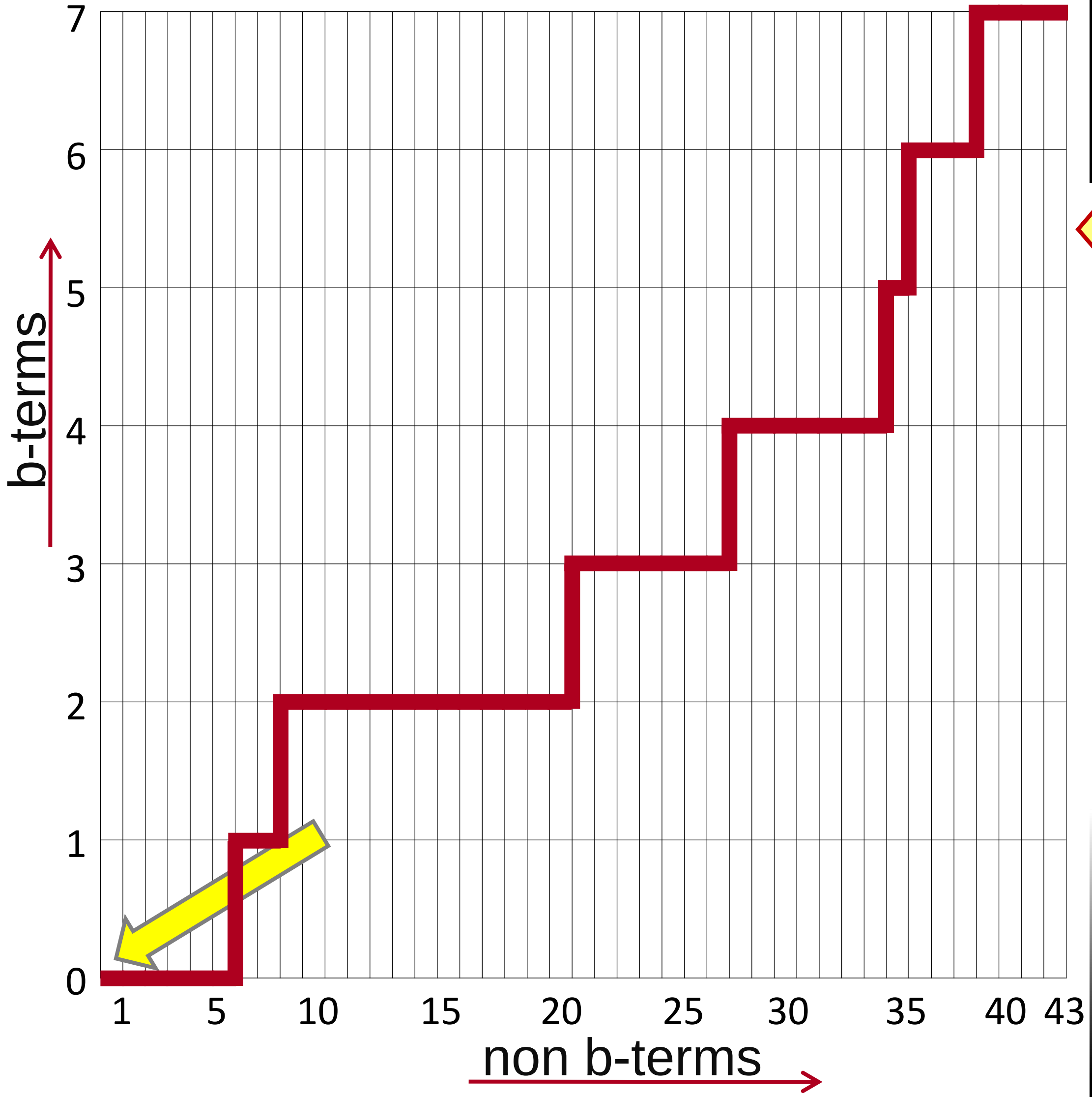
term 8	heuristic 1, heuristic 2, heuristic 3
term 1	heuristic 1, heuristic 3
term 5	heuristic 2, heuristic 3
term 7	heuristic 2, heuristic 3
term 2	heuristic 1
term 3	heuristic 1
term 6	heuristic 2
term 7	-
term 9	-
.	.
.	.
.	.

term 8
term 1
term 5
term 7
term 2
term 3
term 6
term 7
term 9
.
.
.

Domains and datasets

- Training dataset: migraine-magnesium
  - 8,058 documents (2,425- 5,633), 13,433 distinct terms
  - 43 expert identified b-terms (work by Swanson, D. R., Smalheiser, N. R., Torvik, V. I.: Ranking indirect connections in literature-based discovery : The role of Medical Subject Headings (MeSH))
- Test dataset: autism-calcineurin
  - 22,262 documents (14,890-7,372), 17,514 distinct terms
  - 12 expert identified b-terms (work by Petric, I., Urbancic, T., Cestnik, B., Macedoni-Luksic, M.: Literature mining method RaJoLink for uncovering relations between biomedical concepts)

Evaluation ROC curve construction

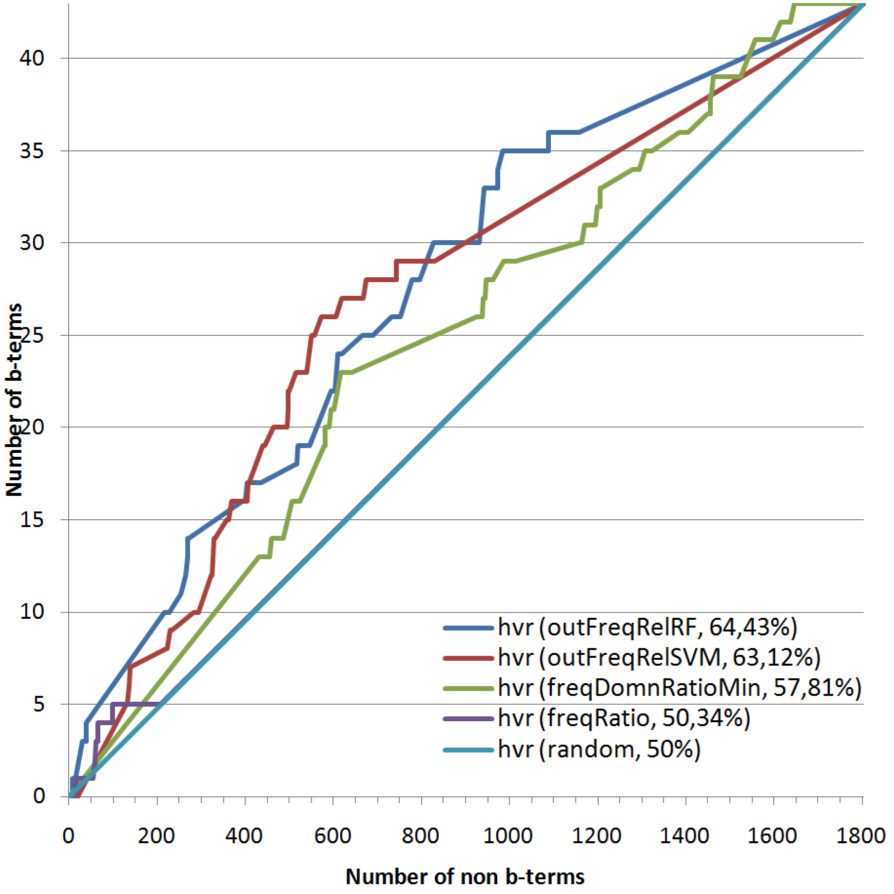


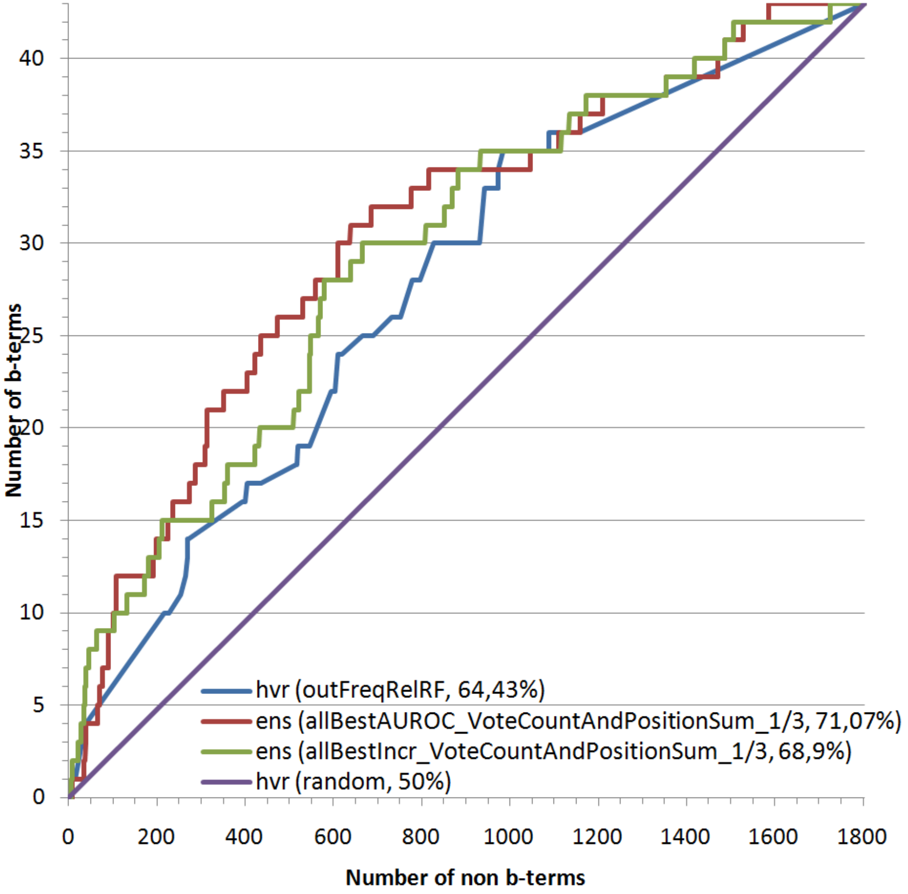
Ranked term list:

50 terms = 7 b-terms + 43 non b-terms

400
animal human
anti inflammatory agent
basal
bruxism
biochemical aspect
brain serotonin
arteriopathy
cerebral artery
cerebral vasospasm
child treatment
clinical comparative
clinical form
clinical statistical
combination treatment
comparative double
comparative double blind

Results on training data set





CrossBee system

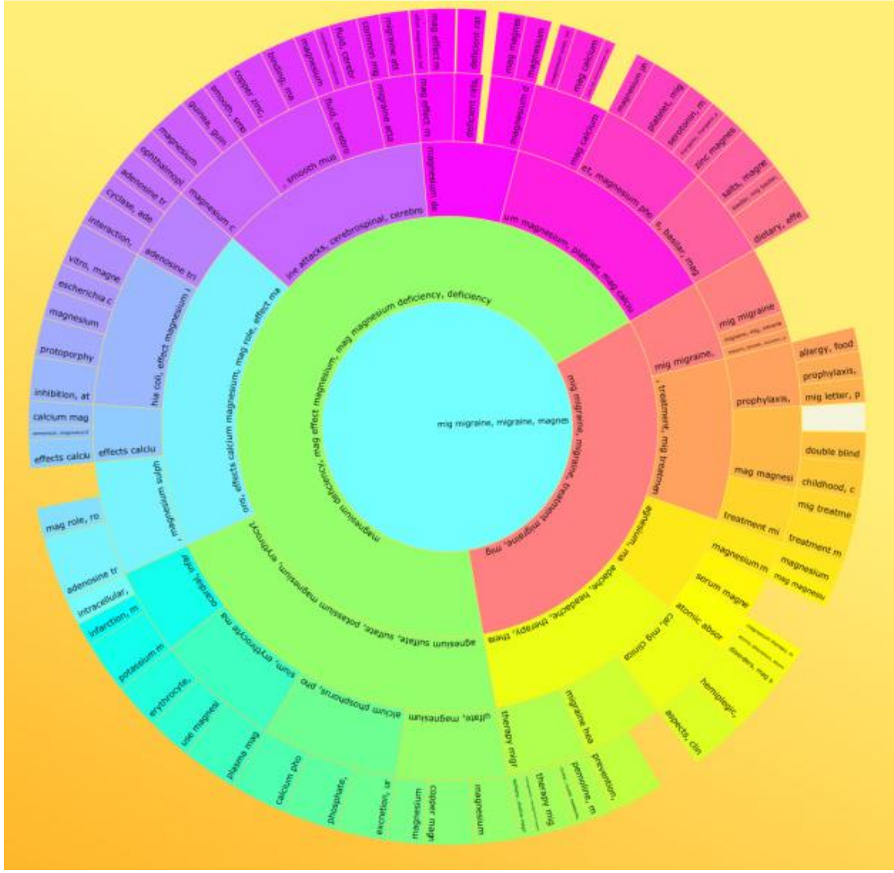
- Cross Context Bisociation Explorer

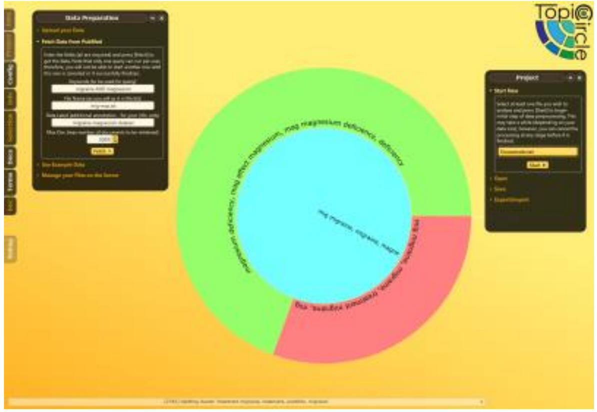
- What is CrossBee?
- Web user interface which fuses multiple approaches developed for discovering bisociations in text

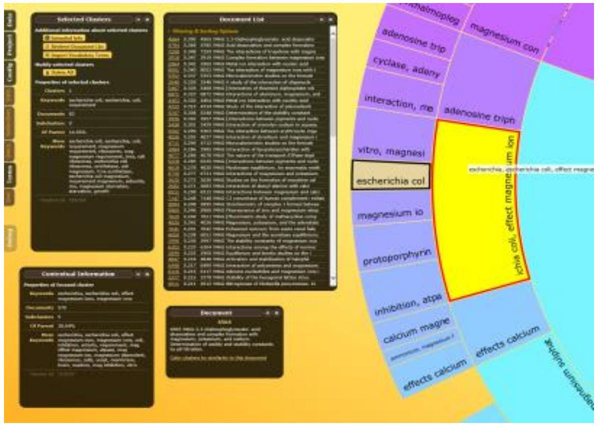
- Why CrossBee?
- Collaborating with domain experts on their data in real time on user friendly system (and thus evaluating their and our hypotheses)

Additional CrossBee functionality

CrossBee Topic Circle for top-down document clustering

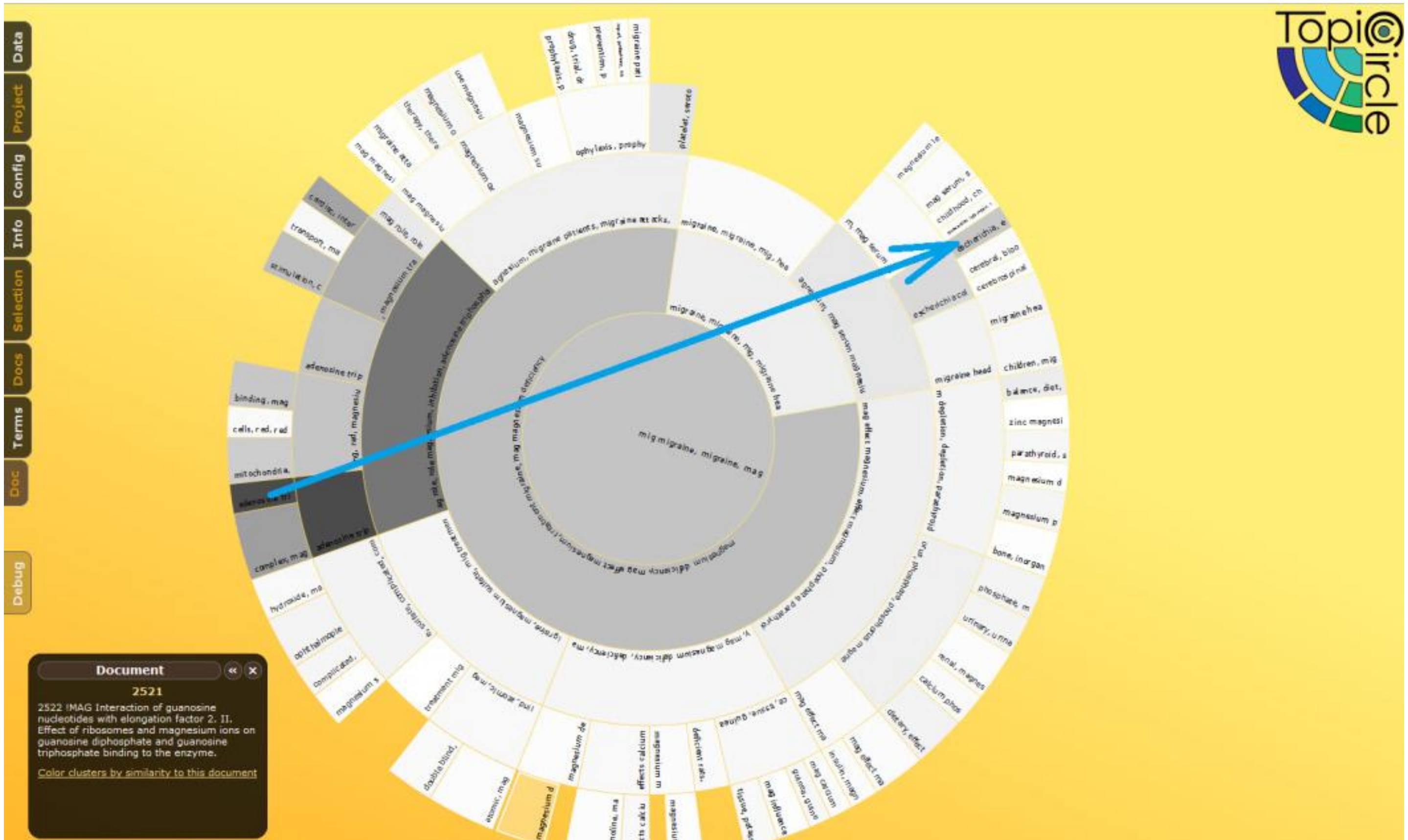



Topic Circle
Data Preparation
Project
mig migraine, migraine, magne
mig effect magnesium, mag magnesium deficiency, deficiency



Additional CrossBee functionality

Cluster colors can show e.g., cluster’s similarity to a single selected document. The arrow shows similar clusters in two different domains, potentially indicate to a novel bisociative link between the two domains.



Document

2521

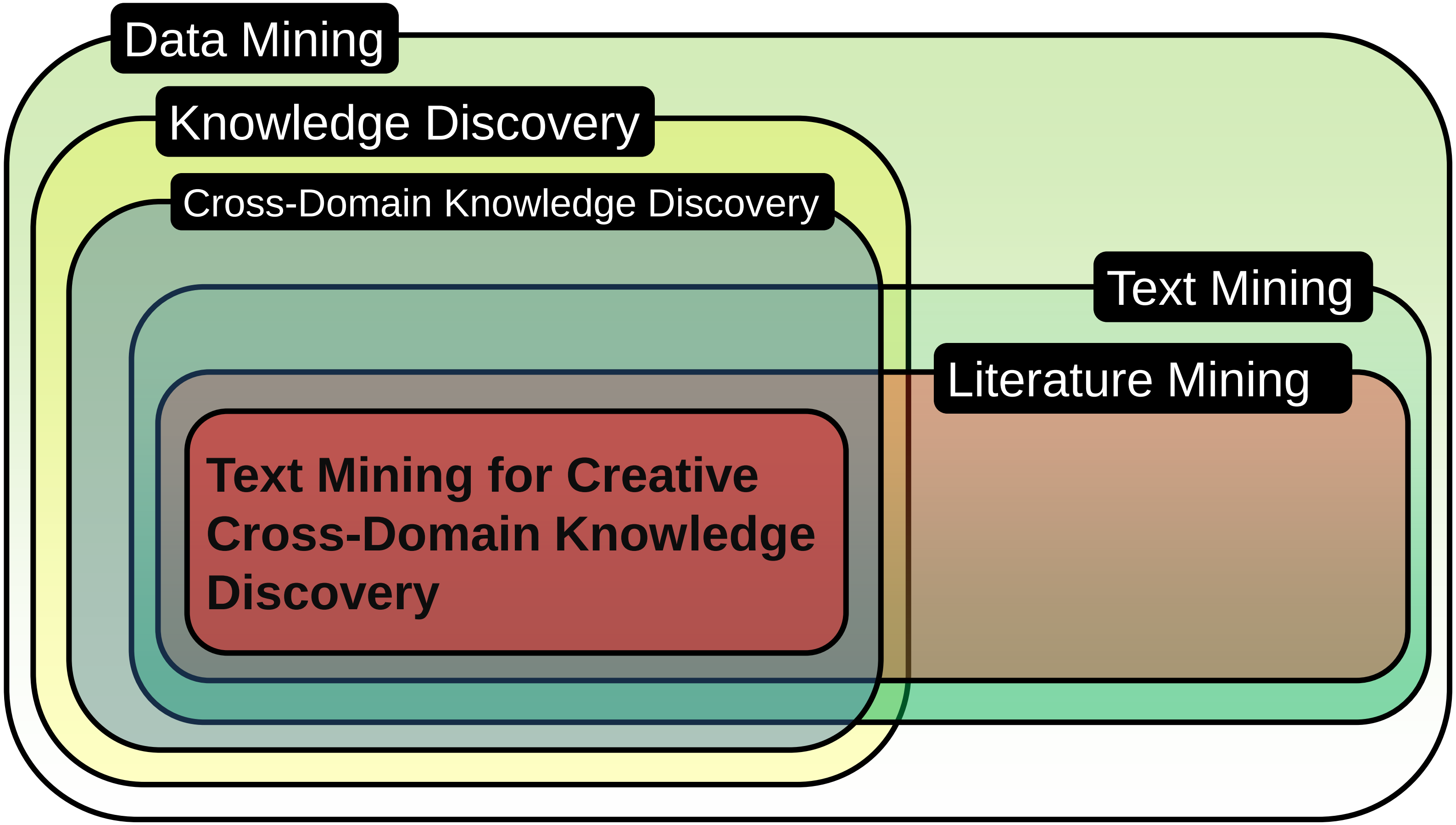
2522 !MAG Interaction of guanosine nucleotides with elongation factor 2. II. Effect of ribosomes and magnesium ions on guanosine diphosphate and guanosine triphosphate binding to the enzyme.

Color clusters by similarity to this document

Summary and conclusions

- Current literature-based approaches mostly depend on simple associative information search
- Potential of outlier detection for b-term discovery
  - Document outlier detection and ranking by NoiseRank
  - Document outlier detection by OntoGen
- CrossBee: improving computational creativity by supporting the expert in the task of cross-domain literature mining (novelty: ensemble-based bridging term ranking)

Summary and conclusions


Data Mining
Knowledge Discovery
Cross-Domain Knowledge Discovery
Text Mining
Literature Mining
Text Mining for Creative Cross-Domain Knowledge Discovery

Selected readings

- M. Berthold (2012): Bisociative Knowledge Discovery, Springer (open access)
- Juršič, M., Cestnik, B., Urbančič, T., Lavrač, N.: Cross-domain literature mining: Finding bridging concepts with CrossBee. In: Proc. 3rd International Conference on Computational Creativity (2012)
- Juršič, M., Cestnik, B., Urbančič, T., Lavrač, N.: HCI empowered literature mining for cross-domain knowledge discovery. In: Proc. HCI-KDD, pp. 124-135, Springer (2013)
- Petrič, I., Urbančič, T., Cestnik, B., Macedoni-Lukšič, M.: Literature mining method RaJoLink for uncovering relations between biomedical concepts. Journal of Biomedical Informatics. vol. 42/2, pp. 219–227 (2009)

Selected readings

- Petrič, I., Cestnik, B., Lavrač, N., Urbančič, T.: Outlier Detection in Cross-Context Link Discovery for Creative Literature Mining. Computer Journal 55/1, pp. 47–61 (2012)
- Sluban, B., Gamberger, D., Lavrač, N. Ensemble-based noise detection : noise ranking and visual performance evaluation. Data mining and knowledge discovery (2013)
- Swanson, D. R.: Medical literature as a potential source of new knowledge. Bull Med Libr Assoc. vol. 78/1, pp. 29–37 (1990)
- Weeber, M., Vos, R., Klein, H., de Jong-van den Berg, L. T. W.: Using concepts in literature-based discovery: Simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries. J. Am. Soc. Inf. Sci. Tech. vol. 52/7, pp. 548–57 (2001)