Text Mining for Creative Cross-Domain Knowledge Discovery

Nada Lavrač
Jožef Stefan Institute, Ljubljana, Slovenia

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

Background and motivation
• Literature-based discovery
• Cross-domain literature mining approaches
  – Outlier detection for cross-domain knowledge discovery
  – Cross-domain knowledge discovery with CrossBee
• Summary and conclusions
The BISON project

• Explore the idea of bisociation (Arthur Koestler, The act of creation, 1964):

  – The mixture - in one human mind – of different contexts or different categories of objects, that are normally considered separate categories by the processes of the mind.

  – The thinking process that is the functional basis of analogical or metaphoric thinking as compared to logical or associative thinking.
The BISON project

- BISON: Bisociation Networks for Creative Information Discovery, European 7FP project, [www.bisonet.eu](http://www.bisonet.eu)
- 12 partners (2008-2011)
- Open access book (Springer 2012): Bisociative Knowledge Discovery edited by M. Berthold
Bisociation discovery in BISON

• BISON challenge:
  – Find new insights: new **bisociations**, i.e., interesting new links **across 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 yet unknown links in a graph, crossing different contexts (domains)

- Open problems:
  - Crossing different contexts (domains): Finding unexpected, previously unknown links between BisoNet nodes belonging to different contexts
  - Crossing different types of data and knowledge sources: Fusion of heterogeneous data/knowledge sources into a joint representation format - a large information network named BisoNet (consisting of nodes and relationships between nodes)
Complementary BISON approach

- Complementary approach: text mining
  - Find yet unknown terms in the intersection of documents, crossing different contexts (domains/literatures)
- Early related work: literature-based discovery (LBD)
  - Swanson (1988, 1990)
  - Smalheiser, Swanson (1998): ARROWSMITH
  - Weeber et al. (2001)
  - Hristovski et al. (2001): BITOLOA
- Recent work: cross-domain literature mining
  - Petrič et al. (2007, 2009): RaJoLink
  - Juršič et al. (2012): CrossBee
  - ...
Talk outline

• Background and motivation

Literature-based discovery

• Cross-domain literature mining approaches
  – Outlier detection for cross-domain knowledge discovery
  – Cross-domain knowledge discovery with CrossBee

• Summary and conclusions
Literature-based discovery

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

Literature about magnesium (A): 38,000 articles

Literature about migraine (C): 4,600 articles

Linking term B1
Linking term B2
Linking term B3
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, ...
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.
- . . .
Scientific literature as a source of knowledge

Example:

- 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!
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 amounts 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
• Literature-based discovery
• Cross-domain literature mining approaches
  – Outlier detection for cross-domain knowledge discovery
  – Cross-domain knowledge discovery with CrossBee
• Summary and conclusions
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
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
**NoiseRank on news articles**

Articles on Kenyan elections: local vs. Western media

<table>
<thead>
<tr>
<th>Rank</th>
<th>Class</th>
<th>ID</th>
<th>Detected by:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WE</td>
<td>352</td>
<td>_<em>Bayes___RF100___RF500___SVM___SVMEasy___SatFilt</em></td>
</tr>
<tr>
<td>2</td>
<td>LO</td>
<td>25</td>
<td>_<em>Bayes___RF100___RF500___SVM___SVMEasy</em></td>
</tr>
<tr>
<td>3</td>
<td>LO</td>
<td>101</td>
<td>_<em>Bayes___RF100___RF500___SVM___SVMEasy</em></td>
</tr>
<tr>
<td>4</td>
<td>LO</td>
<td>173</td>
<td>_<em>Bayes___RF100___RF500___SVM___SVMEasy</em></td>
</tr>
<tr>
<td>5</td>
<td>WE</td>
<td>348</td>
<td>_<em>Bayes___RF100___RF500___SVM___SVMEasy</em></td>
</tr>
<tr>
<td>6</td>
<td>WE</td>
<td>326</td>
<td>_<em>Bayes___RF100___RF500___SVM___SVMEasy</em></td>
</tr>
<tr>
<td>7</td>
<td>WE</td>
<td>357</td>
<td>_<em>Bayes___RF100___RF500___SVM___SVMEasy___SatFilt</em></td>
</tr>
<tr>
<td>8</td>
<td>WE</td>
<td>410</td>
<td>_<em>Bayes___RF100___RF500___SVM___SVMEasy</em></td>
</tr>
<tr>
<td>9</td>
<td>LO</td>
<td>21</td>
<td>_<em>RF100___RF500___SVM___SVMEasy</em></td>
</tr>
<tr>
<td>10</td>
<td>LO</td>
<td>4</td>
<td>_<em>Bayes___RF500___SVM___SVMEasy</em></td>
</tr>
<tr>
<td>11</td>
<td>LO</td>
<td>68</td>
<td>_<em>RF100___RF500___SVM___SVMEasy</em></td>
</tr>
<tr>
<td>12</td>
<td>LO</td>
<td>162</td>
<td>_<em>Bayes___RF500___SVM___SVMEasy</em></td>
</tr>
<tr>
<td>13</td>
<td>WE</td>
<td>358</td>
<td>_<em>Bayes___RF100___RF500___SVM___SVMEasy</em></td>
</tr>
<tr>
<td>14</td>
<td>WE</td>
<td>464</td>
<td>_<em>RF100___RF500___SVM___SVMEasy</em></td>
</tr>
<tr>
<td>15</td>
<td>LO</td>
<td>153</td>
<td>_<em>Bayes___SVM___SVMEasy</em></td>
</tr>
<tr>
<td>16</td>
<td>LO</td>
<td>201</td>
<td>_<em>RF100___RF500___SatFilt</em></td>
</tr>
<tr>
<td>17</td>
<td>WE</td>
<td>238</td>
<td>__RF100___RF500___SVM</td>
</tr>
<tr>
<td>18</td>
<td>WE</td>
<td>364</td>
<td>__Bayes___RF500___SVM</td>
</tr>
<tr>
<td>19</td>
<td>WE</td>
<td>370</td>
<td>__Bayes___RF100___SVM</td>
</tr>
<tr>
<td>20</td>
<td>WE</td>
<td>379</td>
<td>_<em>RF100___RF500___SVMEasy</em></td>
</tr>
</tbody>
</table>
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.
Talk outline

- Background and motivation
- Literature-based discovery
- Cross-domain literature mining approaches
  - Outlier detection for cross-domain knowledge discovery
  - Cross-domain knowledge discovery with CrossBee
- Summary and conclusions
Outlier detection by clustering of PubMed articles

Slide adapted from D. Mladenić, JSI
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

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

Work by Petrič et al. 2010
Talk outline

• Background and motivation
• Literature-based discovery
• Cross-domain literature mining approaches
  – Outlier detection for cross-domain knowledge discovery
  – Cross-domain knowledge discovery with CrossBee
• Summary and conclusions
**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 b-term term term b-term term term term term term term term term term term term |
| b-term term term term b-term term term term term term term term term term term term term term term term |
| b-term term term term b-term term term term term term term term term term term term term term term term |

---
Closed discovery setting
CrossBee: Methodology overview

Data Acquisition and Preprocessing

- Document Acquisition
- Document Preprocessing
- Background Knowledge

Candidate Term Extraction

Term Bisociativity Calculation

Term Sorting

Term Ranking

Incorporating available background knowledge

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

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

Heuristic

\[ s = f(t, d) \]
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

<table>
<thead>
<tr>
<th>term 1</th>
<th>term 2</th>
<th>term 3</th>
<th>term 4</th>
<th>term 5</th>
<th>term 6</th>
<th>term 7</th>
<th>term 8</th>
<th>term 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>term 1</td>
<td>0.149</td>
<td>0.680</td>
<td>0.071</td>
<td>0.175</td>
<td>0.637</td>
<td>0.429</td>
<td>0.175</td>
<td>0.637</td>
</tr>
<tr>
<td>term 2</td>
<td>0.759</td>
<td>0.311</td>
<td>0.071</td>
<td>0.175</td>
<td>0.637</td>
<td>0.429</td>
<td>0.175</td>
<td>0.637</td>
</tr>
<tr>
<td>term 3</td>
<td>0.900</td>
<td>0.429</td>
<td>0.775</td>
<td>0.175</td>
<td>0.637</td>
<td>0.429</td>
<td>0.175</td>
<td>0.637</td>
</tr>
<tr>
<td>term 4</td>
<td>0.666</td>
<td>0.311</td>
<td>0.429</td>
<td>0.175</td>
<td>0.637</td>
<td>0.429</td>
<td>0.175</td>
<td>0.637</td>
</tr>
<tr>
<td>term 5</td>
<td>0.311</td>
<td>0.429</td>
<td>0.175</td>
<td>0.175</td>
<td>0.637</td>
<td>0.429</td>
<td>0.175</td>
<td>0.637</td>
</tr>
<tr>
<td>term 6</td>
<td>0.071</td>
<td>0.175</td>
<td>0.637</td>
<td>0.175</td>
<td>0.637</td>
<td>0.429</td>
<td>0.175</td>
<td>0.637</td>
</tr>
<tr>
<td>term 7</td>
<td>0.175</td>
<td>0.637</td>
<td>0.429</td>
<td>0.175</td>
<td>0.637</td>
<td>0.429</td>
<td>0.175</td>
<td>0.637</td>
</tr>
<tr>
<td>term 8</td>
<td>0.637</td>
<td>0.429</td>
<td>0.175</td>
<td>0.175</td>
<td>0.637</td>
<td>0.429</td>
<td>0.175</td>
<td>0.637</td>
</tr>
<tr>
<td>term 9</td>
<td>0.429</td>
<td>0.175</td>
<td>0.637</td>
<td>0.175</td>
<td>0.637</td>
<td>0.429</td>
<td>0.175</td>
<td>0.637</td>
</tr>
</tbody>
</table>
**Ensemble heuristic**

<table>
<thead>
<tr>
<th>heuristic 1</th>
<th>heuristic 2</th>
<th>heuristic 3</th>
<th>ensemble heuristic</th>
</tr>
</thead>
<tbody>
<tr>
<td>term 3</td>
<td>term 7</td>
<td>term 7</td>
<td>term 1</td>
</tr>
<tr>
<td>term 2</td>
<td>term 6</td>
<td>term 8</td>
<td>term 2</td>
</tr>
<tr>
<td>term 1</td>
<td>term 5</td>
<td>term 1</td>
<td>term 3</td>
</tr>
<tr>
<td>term 8</td>
<td>term 8</td>
<td>term 5</td>
<td>term 4</td>
</tr>
<tr>
<td>term 9</td>
<td>term 1</td>
<td>term 6</td>
<td>term 5</td>
</tr>
<tr>
<td>term 5</td>
<td>term 9</td>
<td>term 2</td>
<td>term 4</td>
</tr>
<tr>
<td>term 7</td>
<td>term 4</td>
<td>term 4</td>
<td>term 7</td>
</tr>
<tr>
<td>term 4</td>
<td>term 2</td>
<td>term 7</td>
<td>term 1</td>
</tr>
<tr>
<td>term 6</td>
<td>term 3</td>
<td>term 9</td>
<td>term 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>term 8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>term 9</td>
</tr>
</tbody>
</table>
## Ensemble heuristic

**final ensemble heuristic**

<table>
<thead>
<tr>
<th>term 8</th>
<th>heuristic 1, heuristic 2, heuristic 3</th>
<th>term 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>term 1</td>
<td>heuristic 1, heuristic 3</td>
<td>term 1</td>
</tr>
<tr>
<td>term 5</td>
<td>heuristic 2, heuristic 3</td>
<td>term 5</td>
</tr>
<tr>
<td>term 7</td>
<td>heuristic 2, heuristic 3</td>
<td>term 7</td>
</tr>
<tr>
<td>term 2</td>
<td>heuristic 1</td>
<td>term 2</td>
</tr>
<tr>
<td>term 3</td>
<td>heuristic 1</td>
<td>term 3</td>
</tr>
<tr>
<td>term 6</td>
<td>heuristic 2</td>
<td>term 6</td>
</tr>
<tr>
<td>term 7</td>
<td>-</td>
<td>term 7</td>
</tr>
<tr>
<td>term 9</td>
<td>-</td>
<td>term 9</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>
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

- 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

Graphs showing the performance of different models on the training data set, comparing the number of non-b-terms against the number of b-terms. The models include:
- hvr (outFreqRelRF, 64.43%)
- hvr (outFreqRelSVM, 63.12%)
- hvr (freqDomnRatioMin, 57.81%)
- hvr (freqRatio, 50.34%)
- hvr (random, 50%)

Another graph shows:
- hvr (outFreqRelRF, 64.43%)
- ens (allBestAUROC_VoteCountAndPositionSum_1/3, 71.07%)
- ens (allBestincr_VoteCountAndPositionSum_1/3, 68.9%)
- hvr (random, 50%)
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
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.
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)


Selected readings