Bisociative Knowledge Discovery via B-term Identification

[Odkrivanje bisociativnega znanja s pomočjo identifikacije B-termov]

Technical Report IJS-DP-10463

[Delovno poročilo IJS-DP-10463]

Matjaž Juršič, Igor Mozetič, Miha Grčar, Nada Lavrač
{matjaz.jursic, igor.mozetic, miha.grcar, nada.lavrac}@ijs.si

Department of Knowledge Technologies
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana
Slovenija

[Odsek za tehnologije znanja
Inštitut Jožef Stefan
Jamova 39, 1000 Ljubljana
Slovenija]

May 2010

[Maj 2010]

Bisociative Knowledge Discovery via B-term Identification

Matjaž Juršič[1], Igor Mozetič[1], Miha Grčar[1], Nada Lavrač[1, 2]

1 Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia; 2 University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia

{matjaz.jursic, igor.mozetic, miha.grcar, nada.lavrac}@ijs.si

Abstract. As the world has been moving into a new era of globalization over the last decades, an information overload has been rapidly gaining its significance among the problems of modern society. A difficulty for a person to comprehend overwhelming amounts of data is especially present in today's scientific community. A scientist could once stayed in touch with quite a broad spectrum of areas similar to his or hers field of research, now, because of increasing number of publications, he or she is often forced into pursuing only very narrow area around his specialization. This paper presents a new methodology, which tries to alleviate this crisis by helping the scientists with vast amount of data - especially scientific text data - to automate the extraction of the parts, which lie in their context of interest. Particularly we are dealing with the problem of the so-called closed discovery, where one already knows which two topics he wants to address jointly. In these cases, our methodology suggests the best terms that connect both areas of research and gives hints, which are the most optimistic pathways to follow in order to come to new discovery. The paper also shows some promising results by performing benchmarks with repeating two interdisciplinary discoveries made on the database of medical articles named PubMed.

Keywords: Bisociation, Knowledge Discovery, B-term Identification, PubMed, Text Mining, Network of Terms, Migraine-Magnesium, Autism-Calcineurin.

1 Introduction

Our approach to computational knowledge discovery is based on the concept of *bisociation*, as coined by Arthur Koestler [1]. Basically, a discovery in science (a situation, an idea), L, occurs when a scientist connects two habitually incompatible frames of reference (contexts, domains), M1 and M2. The event L is said to be bisociated with the two contexts.

When one develops computational support for knowledge discovery in science, one has to take into account at least two issues:

- The scientific landscape has changed considerably since 1964. Today, scientists are faced with large amounts of already available knowledge (heterogeneous and distributed sources on the Web), and large quantities of data from high-throughput experimental platforms. One goal of computational support for scientific discovery is to enhance human analytical capabilities and help her/him in connecting seemingly unrelated ideas.
- Computational support requires formal definitions of the above concepts (context, a discovery event, and bisociation), selection of appropriate knowledge and data representations, development of reasoning algorithms and user interfaces.

1.1 Knowledge representation

In the BISON project (European FP7 FET-Open project, 2008-2011, http://www.bisonet.eu/), we investigate possible computational realizations of bisociative reasoning. We decided to base the knowledge representation on a *bisociative information network*, called *BisoNet*. A BisoNet and related bisociation concepts can be defined as follows:

- A BisoNet is a large graph, where nodes are concepts and edges are probabilistic relations. The assumption of the project is that it is relatively easy to construct a BisoNet automatically, from available resources on the Web. Unlike semantic nets and ontologies, it carries little semantics and to a large extend encodes just circumstantial evidence that concepts are somehow related through edges with some probability.
- A context (frame of reference) is represented by a BisoNet subgraph. An assignment of different subgraphs to different contexts is subjective to the user (domain expert). This is to support a move from her/his 'normal' context to a habitually incompatible context, i.e. a creative jump 'out-of-the-frame'.
- A bisociation is an implicit link (to be discovered) between nodes or subgraphs from different contexts.
- Graph analysis algorithms can be used to compute links or structural similarities between distant nodes and subgraphs in a BisoNet.

Three patterns of bisociation were identified so far in the BISON project:

- **Bridging concepts** connect dense subgraphs from different contexts.
- **Bridging subgraphs** are subgraphs that connect concepts from different contexts.
- Bisociations based on **structural similarity** are represented by subgraphs of two different contexts with a similar structure**.**

1.2 Data Sources

There are roughly three types of resources, from which a BisoNet can be constructed:

- Structured data (databases, thesauri, ontologies, …). An example of a BisoNet created from several biomedical databases is Biomine [2].
- Textual data (scientific papers, web pages, news, …). An example resource in biomedicine are PubMed abstracts with MeSH annotations.
- Experimental data, for example, microarray data from wet lab experiments in biology or medicine.

In general, we expect the need to construct, use and combine BisoNets from different types of resources. We envision several processing phases which eventually lead to computational support for creative discoveries by humans:

- Phase 1: Preprocessing of textual and experimental data with text mining and data mining tools, respectively.
- Phase 2: Creation and combination of different BisoNets.
- Phase 3: User-habitual context specification and interactive exploration.

1.3 Context of this Work

In this paper we concentrate on discoveries from textual sources. Finding links between seemingly unrelated concepts from texts was already addressed by Swanson [2]. The Swanson's approach implements *closed discovery*, the so-called A-B-C process. Here, concepts A and C are given and one searches for intermediate B concepts.

A more challenging problem is open discovery, where only A is given, and target C concepts are proposed via intermediate B candidates (also referred as b-terms). An approach to open discovery was implemented in the RaJoLink system [3].

This paper presents a very limited approach to computational discovery with an emphasis on the re-creation of Swanson's and RaJoLink approaches, and an evaluation on two published benchmark scenarios. Specifically, we narrow down the broad spectrum of possibilities outlined above to the following points:

- Two problem domains: migraine-magnesium, and autism-calcineurin (A-C concept pairs in the closed discovery scenario).
- Goal: find intermediate B concepts.
- Textual sources given: PubMed titles (migraine-magnesium) and abstracts (autism-calcineurin).
- Two types of graphs:
    - Document-based graph, where nodes are documents (titles and abstracts, respectively) and edges are weighted by similarities between nodes.
    - Term-based graph (a BisoNet) where nodes are terms from documents, and edges are their co-occurrences, appropriately weighted.
- A two-phase approach to discovery of B concepts:
    - Phase 1: Preprocessing and text mining (clustering) of documents to identify candidate B concepts, called b-terms here.
    - Phase 2: Creation of a term-based BisoNet from all relevant terms, including 'interesting' b-terms.
- Two contexts: defined by the problem (A- and C-related documents), and not by the user. Contexts are degenerated to single nodes (A and C) to which, however, rich phase 1 derived features are attached.
- Bisociations: bridging concepts B, linking A and C.
- Evaluation: comparison to the golden standard, as extracted from publications [3] [2].

A note worth mentioning: in our particular approach in this paper, a BisoNet creation is a major challenge, while BisoNet exploration is straightforward, once appropriate b-term candidates are identified and weighted. However, from the user's point of view, interactive exploration and appropriate presentation of the underlying features (extracted in phase 1 and attached to the nodes) seems very valuable in deciding which bridging concepts to pursue in further investigations.

1.4 Structure of this Paper

The purpose of this paper is to present a novel knowledge-from-text discovery methodology using term networks. At the beginning, we are given with the input to the whole procedure - raw texts. The goal and the output of the procedure is new knowledge about previously unknown relationships between entities, which are described in the input text. As we concentrate on bisociative knowledge, we put a strong emphasis on b-term identification phase. The structure of this paper closely follows the data though its development from raw text to the knowledge visualization enabled BisoNets.

Figure 1 shows the organisation of the following sections. Section 2 - Data Preprocessing, briefly describes the pre-processing step, which generates list of terms tagged with various statistical data extracted from raw text, and used to identify b-terms in the next phase. Section 3 - Identifying B-term Candidates is the main section of this paper. It presents an approach how one can retrieve important terms, b-terms, from a text using just statistical properties of the terms. For the proof of concept, we provide a case study on two datasets. Afterwards, in Section 4 - Selection of B-terms, we propose an expert guided technique for further refinement of the terms found in the previous step. Section 5 - Network Creation and Exploration sketches another simple UI tool, which helps an expert to actually find a new knowledge from the generated term network.

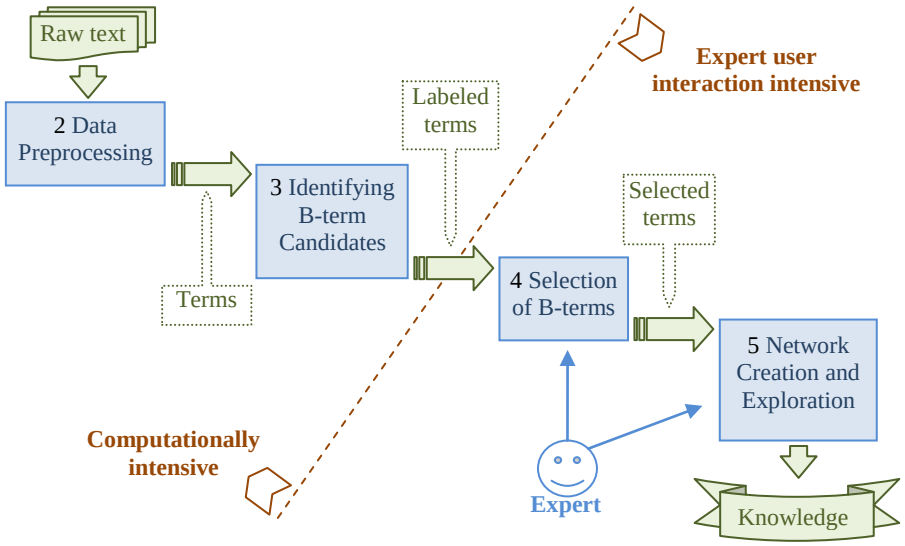

Figure 1: A process of transforming raw data to knowledge in four steps, which correspond to the four sections of this paper.

An alternative view on the structure of the paper is the division on two main parts: computationally intensive part, presented by sections 2 and 3; and expert-user-interaction intensive part presented by the sections: 4 and 5.

2 Data Preprocessing

The first step in the process of creating a network from text sources is text preprocessing. We do not want to burden a reader with too much details how we do that, since one can find these information in any introductory text-mining textbook. However, the two case studies we present in the next section use the following parameters to build the vector representations (bag of words) of the documents:

- Stopwords: depends on the dataset. However, in both cases we make sure that all words used also in b-term list are deleted out of stopwords list. Otherwise these words are filtered out and b-terms cannot be found.
- Stemmer: LemmaGen lemmatizer for English [5].
- Term length: max term length (N-gram length) is based on max b-term length and is set to three in our two examples (only a few b-terms, e.g.: "calcium channel blocker", contain three words).
- Min word frequency: is the minimum frequency (occurrence) of the words (in a text) which are to be included in the vocabulary. It is set to two.

When converting documents to standard text-mining representation, TF-IDF vectors, stopwords are not removed from the word-set. This non-standard modification is due to extremely short documents in the specific titles dataset. In case when a domain specific stopword list is present (or constructed by user on the fly) this information is used later in the processing of documents, but not in the preprocessing step.

The next step is to define how the network should be created - how does it look like, especially what the nodes in the network are. Since the purpose of this paper is to rediscover b-terms and to find out which rules make distinction between b-terms and all the other terms it seems reasonable, that all the terms (including b-terms) now become network nodes. The decision on how the terms will be represented is also non-trivial. We identified two basic approaches: Attached to the term is a (weighted) set of either:

- documents where it appears (document representation), or
- terms which co-appear with it (term representation).

The second option is equivalent to calculating (weighted) centroid of the documents in which the term appears. There is also an issue about weighting model to be used (if any). We tested many possible combinations while considering the quality of generated network. Therefore, we decided in favor of the next term/node representation: Each term is represented with the centroid (average) of TF-IDF (term frequency - inverse document frequency) vectors of documents. Additionally each vector in the centroid is weighted by the term-frequency of the term in the specific document.

To enable b-term discovery, we include as much of the statistical information as we recognize is required and is possible to be retrieved from a text. The output of the preprocessing step is therefore, not only a list of b-terms, but also accompanying statistics, which comes along as tags of parsed terms. This additional information is crucial for the calculation of heuristics, which we present and evaluate in the next section.

3 Identifying B-term Candidates

To separate b-terms from the (usual) non b-terms, we create a set of heuristics which seem promising for b-term discovery. Subsequently, we evaluate these heuristics and choose the one performing the best (named b-potential) as the weight of a node in a created network. This section firstly lists all the heuristics used in the evaluation, and then it presents the results of evaluation on two datasets, namely: migraine-magnesium and autism-calcineurin dataset.

3.1 Heuristics

The heuristics presented here are used in the evaluation of their capabilities to rank b-terms. All of them operate only on the data - statistics - which is retrievable from the text during the pre-processing. The exact evaluation procedure is discussed in detail in the following sections, at this point, a heuristic is a function that evaluates a goodness of a term and returns one numeric output as an estimate of term's "b-term quality". We are initially searching for such heuristic that sorting terms by its value, would result in finding all b-terms together either at the top or at the bottom of such sorted list (depending on whether higher is better or lower is better). This is an ideal case which we do not expect to find, however, we still want to find some, which sorting would bring much more b-terms compared to non b-terms either on the top or to the bottom.

Following list contains more detailed specifications and definitions of some terms (mainly statistics) used in the heuristics descriptions below:

- all terms and centroids are represented in "term representation" - if not otherwise noted,
- similarity: stands for cosine similarity - if not otherwise noted,
- migraine/magnesium cluster: stands for centroid of all documents from migraine/magnesium context in the dataset,
- migraine/magnesium centroid: stands for centroid of all documents containing word "migraine"/"magnesium",
- outliers: outliers are the documents that come from the migraine cluster but are (considering cosine similarity) closer to the magnesium cluster centroid – and vice versa.
- bissociation index: cosine similarity like measure to evaluate similarity between two vectors (defined in [6]).
- tpf, idpf: term pair frequency, inverse document pair frequency: pair frequency is a frequency of two terms together (they have to appear together in the same document, but not necessary consecutively in the text).

Since the next list of heuristics is quite extensive, a reader can typically safely skip it, and return to it in the case he or she is interested into some specific heuristic found in

the evaluation subsection below. Some of the (simpler) specifications are in the form of natural language descriptions, while another (more complicated) are in the form of equations, which usually combine the simpler ones.

- freq: num. of all appearances of a term in all documents,
- docFreq: num. of documents where a term appears,
- freqRatio: freq / docFreq,
- simMig: similarity of a term to the migraine cluster,
- simMag: similarity of s term to the magnesium cluster,
- simRatio: simMig / simMag,
- simFact: simMig × simMag,
- simMigCent: similarity of term to the migraine centorid,
- simMagCent: similarity of term to the magnesium centorid,
- simCentRatio: simMagCent / simMagCent,
- simCentFact: simMagCent × simMagCent,
- simMigDoc: similarity of term to the migraine centorid (document repres.),
- simMagDoc: similarity of term to the magnesium centorid (document repres.),
- simDocRatio: simMigDoc / simMagDoc,
- simDocFact: simMigDoc × simMagDoc,
- docCountMig: num. of doc. in the migraine cluster where a term occurs,
- docRatioMig: docCountMigraine / docFreq,
- docCountMag: num. of doc. in the magnesium cluster where a term occurs,
- docRatioMag: docCountMagnesium / docFreq,
- docCountFact: docCountMigraine × docCountMagnesium,
- docCountRatio: docCountMigraine / docCountMagnesium,
- missclassMig: docRatioMigraine × simMagCent,
- missclassMag: docRatioMagnesium × simMigCent,
- bisMig: bissociation index of a term to the migraine centroid,
- bisMag: bissociation index of a term to the magnesium centroid,
- bisFact: bisMig × bisMag,
- bisRatio: bisMig / bisMag,
- bisMinMigMag: min(bisMig, bisMag),
- bisMaxMigMag: max(bisMig, bisMag),
- occurInOutlyer: num. of occurrences of a term in outliers,
- occurInNotOutlyer: num. of occurrences of a term in non-outliers,
- percInOutlyer: occurInOutlyer / freq,
- percInNotOutlyer: occurInNotOutlyer / freq,
- avgTfIdf: average TF-IDF of a term (in all documents where a term appears),
- sumTfIdf: sum of TF-IDFs of a term (in all documents where a term appears),
- bisocPotential: term's TF-IDF in migraine centroid × term's TF-IDF in magnesium centorid,
- bisocPotentialSum: term's TF-IDF in migraine centroid + term's TF-IDF in magnesium centorid,
- tpfIdpfMig: tpf("migraine" term) × idpf("migraine", term),
- tpfIdpfMag: tpf("magnesium" term) × idpf("magnesium", term),
- tpfIdpfFact: tpfIdpfMig × tpfIdpfMag,
- tpfIdpfRatio: tpfIdpfMig / tpfIdpfMag,
- random: used as a baseline for comparison.

3.2 Benchmark

Given is a "golden standard" list of b-terms available for the domain of observation. Thus, we can label the terms and observe how well the heuristics are promoting the true b-terms compared to regular non b-terms.

We tested all heuristics on both datasets using analysis through ROC (Receiver Operating Characteristic curve) and AUROC (area under ROC). ROC curves were constructed in the following way:

- Sort all terms by observed heuristics (or reverse if reversing yields better area under ROC)
- In the case when a heuristics outputs the same estimate for many terms (for a block of terms) use random ordering for inner block sorting. Since this situation

is not rare in our datasets, we averaged AUROC through 20 different sorting of each list. We also list standard deviation of AUROC for each heuristic.

- Go from the start of the sorted term list and for each term determine if term is: b-term: draw vertical line (up) on the ROC curve, non b-term: draw horizontal line (right) on the ROC curve.

Like this, we get one (averaged) ROC curve for each heuristic. The ROC's vertical axis scale is from 0 to number of b-terms and the horizontal is from 0 to number of non-b-terms. AUROC is defined by percentage (area under the curve divided by area under best possible curve). If a heuristic is perfect (it detects all b-terms and ranks them the top of the ordered list) we get a curve that goes firstly just up and then just left with an AUROC of 100%. The worst possible heuristic sorts all terms randomly regardless of being b-term or not and achieves AUROC of 50% in average.

Next two subsections contain experimental results for two domains, namely migraine-magnesium dataset [2] and autism-calcineurin dataset [3], used for the evaluation of chosen b-term discovery indicators

3.2.1 Migraine-Magnesium Dataset Results (Training Dataset)

Heuristic	Cr	AUROC	Stdev	...			
simDocFact	(+)	93,78%	0,00%	bisMag	(-)	79,46%	0,00%
tpfIdpfFact	(+)	93,61%	0,06%	simMag	(-)	76,31%	0,00%
bisocPotential	(+)	93,57%	0,06%	tpfIdpfRatio	(+)	75,90%	0,07%
docCountFact	(+)	93,52%	0,06%	simMagCent	(-)	74,54%	0,00%
docCountMig	(+)	87,55%	0,20%	occurInNotOutlyer	(+)	71,15%	1,05%
simMigDoc	(+)	86,74%	0,00%	docFreq	(+)	71,08%	1,05%
simFact	(+)	85,06%	0,00%	freq	(+)	71,02%	1,06%
simMig	(+)	84,40%	0,00%	sumTfIdf	(+)	71,02%	1,06%
tpfIdpfMig	(+)	84,14%	0,22%	tpfIdpfMag	(+)	64,58%	0,85%
simMigCent	(+)	84,04%	0,00%	missclassMag	(+)	63,78%	0,00%
missclassMig	(+)	83,87%	0,00%	simMagDoc	(-)	63,47%	0,00%
simCentRatio	(+)	82,52%	0,00%	bisFact	(+)	61,90%	0,00%
simCentFact	(+)	82,39%	0,00%	bisMaxMigMag	(+)	60,75%	0,00%
bisMig	(+)	81,80%	0,00%	bisMinMigMag	(+)	56,41%	0,00%
simRatio	(+)	80,84%	0,00%	occurInOutlyer	(+)	56,18%	3,78%
simDocRatio	(+)	80,60%	0,00%	percInOutlyer	(+)	56,14%	3,77%
docCountRatio	(+)	80,56%	0,06%	percInNotOutlyer	(-)	56,14%	3,77%
docRatioMig	(+)	80,56%	0,06%	freqRatio	(+)	52,97%	4,66%
docRatioMag	(-)	80,56%	0,06%	avgTfIdf	(+)	52,97%	4,66%
bisRatio	(+)	80,40%	0,00%	docCountMag	(+)	51,85%	0,85%
bisocPotentialSum	(+)	80,32%	0,06%	**random**	(+)	50,96%	4,70%

Table 1: Comparison results of all heuristics defined for b-term retrieval (ordered by quality - AUROC). First column is the name of the heuristic; second displays what is the correlation of heuristic: (+) positive - high value of heuristic means higher probability of term being a b-term and (-) negative - means just the opposite; third column is a percent of area under ROC curve; and the last is the standard deviation of AUROC.

The dataset for this benchmark consists of two sets of PubMed [5] article titles. First set is about concept migraine and the second is about concept magnesium. All documents used in the analysis are retrieved using PubMed article search which is limited with the condition that an article was published before 1988 and using keywords "migraine" and "magnesium" – one for each concept. The date limitation is enforced due to Swanson's discovery of this specific bisociation in this year. For the stopwords and b-term list we use the lists that Swanson defined in his work. Stopwords are modified according to the discussion in the section 3.2. Also all the

other settings are used as defined in that section. The basic properties of the parsed and preprocessed dataset are:

- 8058 documents, 13433 distinct terms found,
- 2425 documents on migraine, 5633 documents on magnesium,
- 42 b-terms (97.7%) found among extracted terms (there are 43 b-terms in total defined by Swanson).

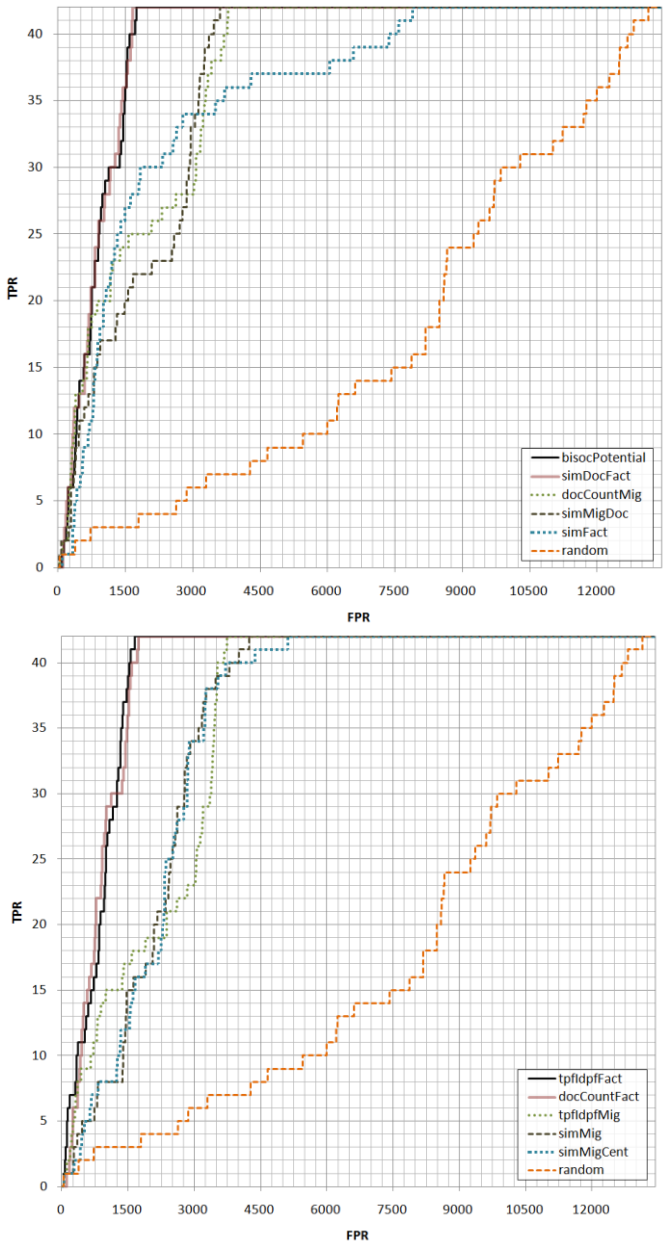

Figure 2: ROC curves of 10 best heuristics on migraine-magnesuim dataset. ROC curves are divided to two separate plots since some curves are almost the same (e.g.: bisocPotential and docCountFact) and are not separable on the same plot. Subsection 3.2 discribes the charts and the meaning of axels in more details.

Table 1 and Figure 2 show the results of the benchmark. On one hand Table 1 shows the performance (quality-wise) of all heuristics, while on the other hand, Figure 2 offers the detailed perspective into the ROC curves just for the 10 best performing heuristics. We discuss and compare the heuristics in section 3.3 - Comparison of the Heuristics, at this point we would just like to point out to the property that the

majority of created heuristics are significantly better on separating b-terms from non b-terms, compared than the random decision maker. This result confirms the basic proposition of this paper, which suggests there are differences between b-terms and non b-terms already at the level of basic statistics indicators retrievable from the texts.

3.2.2 Autism-calcineurin Dataset Results (Testing Dataset)

Heuristic	Cr	AUROC	Stdev	...			
bisocPotential	(+)	84,36%	3,17%	docRatioMag	(+)	63,62%	1,26%
docCountFact	(+)	83,96%	3,17%	simMagCent	(+)	59,92%	0,00%
docCountMag	(+)	78,32%	0,48%	tpfIdpfMig	(-)	57,54%	3,39%
occurInOutlyer	(+)	77,89%	4,36%	simMagDoc	(+)	57,45%	4,76%
percInOutlyer	(+)	77,33%	4,45%	bisRatio	(-)	56,65%	2,67%
percInNotOutlyer	(-)	77,33%	4,45%	simCentFact	(-)	56,51%	0,00%
bisocPotentialSum	(+)	76,17%	0,02%	simDocRatio	(-)	56,38%	4,91%
freq	(+)	75,66%	1,49%	tpfIdpfMag	(+)	56,22%	4,84%
sumTfIdf	(+)	75,66%	1,49%	simRatio	(-)	55,94%	2,67%
docFreq	(+)	73,38%	0,20%	docCountMig	(+)	55,49%	1,41%
occurInNotOutlyer	(+)	72,78%	0,23%	tpfIdpfRatio	(-)	55,49%	4,91%
bisMaxMigMag	(-)	66,71%	0,00%	bisMag	(+)	55,01%	2,67%
simMigCent	(-)	66,53%	0,00%	simMag	(+)	54,11%	2,67%
simMig	(-)	66,22%	0,00%	bisFact	(+)	53,77%	2,71%
bisMig	(-)	64,63%	0,00%	simMigDoc	(-)	53,50%	3,34%
freqRatio	(+)	64,45%	6,64%	bisMinMigMag	(+)	53,23%	2,71%
avgTfIdf	(+)	64,45%	6,64%	simFact	(+)	53,08%	2,71%
simCentRatio	(-)	63,90%	0,00%	missclassMig	(-)	52,41%	1,24%
missclassMag	(+)	63,67%	0,00%	simDocFact	(+)	50,79%	8,30%
docCountRatio	(-)	63,62%	1,26%	tpfIdpfFact	(+)	49,67%	8,30%
docRatioMig	(-)	63,62%	1,26%	**random**	(-)	48,62%	11,03%

Table 2: Comparison results of all heuristics (ordered by quality - AUROC) for autism-calcineurin dataset. Columns are described in detail under Table 1.

In this benchmark the goal was initially based on open discovery - just A concept (autism) is known. However, since the main goal of this paper is the b-term discovery we set an experiment as if both A and C concepts are known. As a result we again get closed discovery setting as in migraine-magnesium dataset. Both concepts (autism and calcineurin) of this benchmark are also retrieved using PubMed keyword search. Settings of the preprocessing phase are as discussed in the Subsection 3.2, with the slight modification, since authors in [2] do not use stopwords list but rather a vocabulary (which is in fact inverse of stopwords list - allowed words list).

The basic overall statistics of the parsed and preprocessed dataset is:

- 22262 documents, 17514 distinct terms found,
- 14890 documents on autism, 7372 documents on calcineurin,
- 8 b-terms (66.6%) found among extracted terms (there are 12 b-terms in total identified by Authors).

Table 2 and Figure 3 show the results of the benchmark on autism-calcineurin in the same way as Table 1 and Figure 2 show for the magnesium-migraine dataset. The differences between figures are effect of different numbers of b-terms in both datasets - just 8 b-terms are found among the extracted terms in the autism-calc. case.

Note that the names of the heurstics are not correctly named in this benchmark. Names are again referencing magnesium-migraine dataset. Therefore a reader is asked to interpret the names by replacing migraine and magnesium with autism and calcineurin respectively. This will be corrected in the final version of this paper.

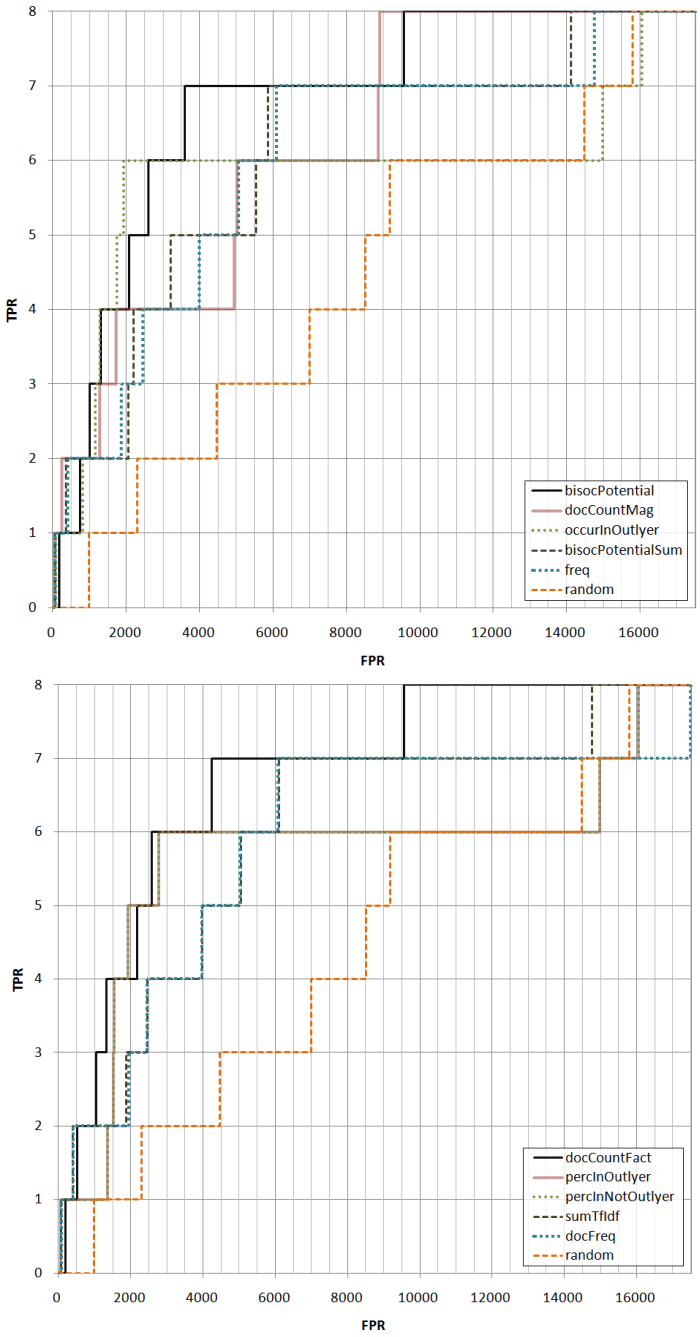

Figure 3: ROC curves of 10 best heuristics on autism-calcineurin dataset. ROC curves are, similarly as on first dataset, divided to two separate plots since some curves are so similar that they are not separable if on the same plot.

3.3 Comparison of the Heuristics

Overall, the results of the two benchmarks are positive. They confirm the initial idea for this paper, which assumes that b-terms and non b-terms are separable already at the low level of statistical properties of the terms (like co-appearance, similarity, etc.). Both tests confirmed the hypothesis, as defined statistics are able to quite efficiently separate b-terms from the rest.

When we examine Figure 2 in detail, we see that various heuristics have very diverse b-term retrieval characteristics. The best four (*simDocFact, tpfIdpfFact,*

bisocPotential, and *docCountFact*) are very similar by performance and also by characteristic ROC curve. They all discover first half of 43 b-terms after approximately 900 retrieved terms and all of b-terms before reaching 1,800 retrieved terms. The ROC curve is about linear with around one discovered b-term per every 42 terms retrieved. Although this does not seem very promising, we have to take into consideration that there are still another 11,000 terms in the "term-pool" after all the b-terms are retrieved.

Figure 3 shows similar, but somehow less clear picture. The main problem of this dataset is low number of b-terms found by the original paper authors [2]. As a consequence the resolution along y-axis is low, which partly confuses the picture. However, when we analyze it thoroughly, it reveals similar situation as Figure 2 - but with two exceptions. There are two b-terms which are found very late by all heuristics (around after a half of retrieved terms). It seems that the statistical properties of these two do not match the others. Nevertheless, excluding these two and taking into account that there are more terms in this dataset compared to the previous one, the numbers match up - all b-terms retrieved among the first 2.000-3.000 terms / out of 18.000 terms, while retrieval is about linear.

What does this outcome mean for the b-term selection step? The outcome can be viewed from two perspectives:

- The first is more theoretical and compares how many terms would an expert (with an eye to spot a b-term if he or she sees it) have to traverse before one finds them all. In our cases we conclude that only 15% of all terms have to be seen to retrieve all b-terms which results in 85% of work saved.
- The more practical aspect is an automatic generation of a list of terms, which is offered to an expert for the purpose of selection of b-terms. In this case, a list generated using by our heuristic ordered terms is 7 times more probable to contain a b-term than a randomly created list of terms.

As the most promising heuristics we picked the one, which is found among the best in first scenario (training dataset from Table 1), namely *bisocPotential.* Our decision is not based on the AUROC performance alone, but we also employed Occam's razor principle and selected the simplest heuristics among similarly performing ones. The second experiment (testing dataset in Table 2) confirmed our choice, since the chosen measure is again among the best performing ones. The definition of *bisocPotential* is the following:

$$bisocPotential_T = tfidf_T(domain1\ centroid) \times tfidf_T(domain2\ centroid)$$

Where $tfidf_T(domainX\ centroid)$ is a value of the tfidf component of the term T in the X domain's centroid.

From now on we refer to the new measure as **b-potential** or **BP** (shortened for b-term-high-potential-measure). The following sections are building mainly on the result of this section and extensively use BP measure to offer the expert user the most relevant terms.

4 Selection of B-terms

Now, as we have the measure to assess the quality of terms, namely b-potential, we are all set to include the expert into the research. We envision b-term selection in the closed discovery scenario in the following way (we briefly subsume the sections 2 and 3 in the process overview below):

- We acquire documents from two domains, e.g. migraine and magnesium, by querying PubMed with keywords "migraine" and "magnesium" for documents published before 1979 (i.e. approximation of the Swanson's dataset).
- We preprocess documents in a typical text-mining way: we remove domain-independent English stop words (such as a, the, they, is), apply stemming, and compute L2-normalized TF-IDF vectors corresponding to documents.
- We project the TF-IDF vectors onto 2-dimensional canvas (See Figure 4) by employing the least-squares meshes projection technique. The projected documents visually form two clusters (i.e. the migraine and magnesium cluster.
- We allow the user to explore the space of documents. By clicking on the point representing a document, the user is given additional information about the document. The most relevant information is the ranking of document's terms according to b-potential. BP in this usage, measures how much each term contributes to the similarity between its document and the opposite centroid vector. If the document is from the domain of migraine, for example, we compute the component-wise product of the document's TF-IDF vector and the magnesium centroid vector. It turns out that if we rank terms according to BP in this way, relevant b-terms tend to be ranked higher than other non-stop terms as discussed in previous section.

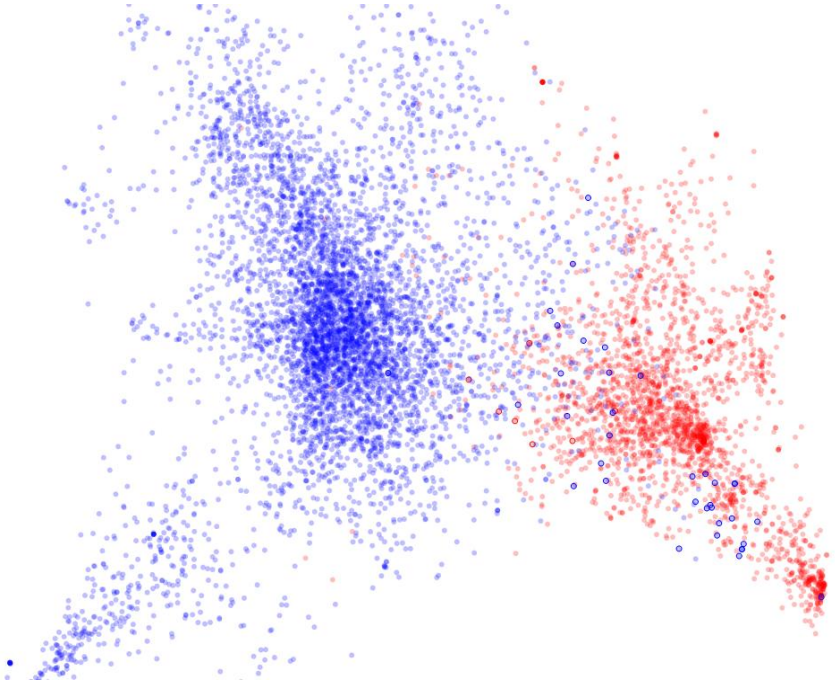

Figure 4: 2-dimensional projection of documents (titles only) about magnesium (blue) and migraine (red). Outlier documents[1] are bolded for a user to easily spot them.

[1] The outliers do not seem to contain more b-terms than other documents. Empirical experiments to prove or reject this claim still need to be conducted. The outliers are defined to be those documents that lie closer to the opposite centroid than to their own centroid (e.g. documents about magnesium that lie closer to the migraine centroid than to the magnesium centroid). The identified outliers are emphasized (i.e. outlined) in the visualization

Figure 5 shows an example of the user guided application for b-terms selection. The user clicks on a particular point in the visualization. The documents in the proximity of the point are shown in the right panel. Each document is described with its label (either "magnesium" or "migraine"), its title (e.g. "The non-epileptiform basilar artery migraine"), and the list of terms that contribute to the similarity between the document and the opposite centroid vector, sorted descending according to their BP (e.g. "artery*, basilar artery*, basilar, migraine"). The terms marked with "*" are the b-terms identified by the user (in this screenshot, these are the Swanson's golden standard b-terms). Those put into brackets are the domain-specific stop words identified by the user (in this screenshot, these are the stop words available with the Swanson's dataset).

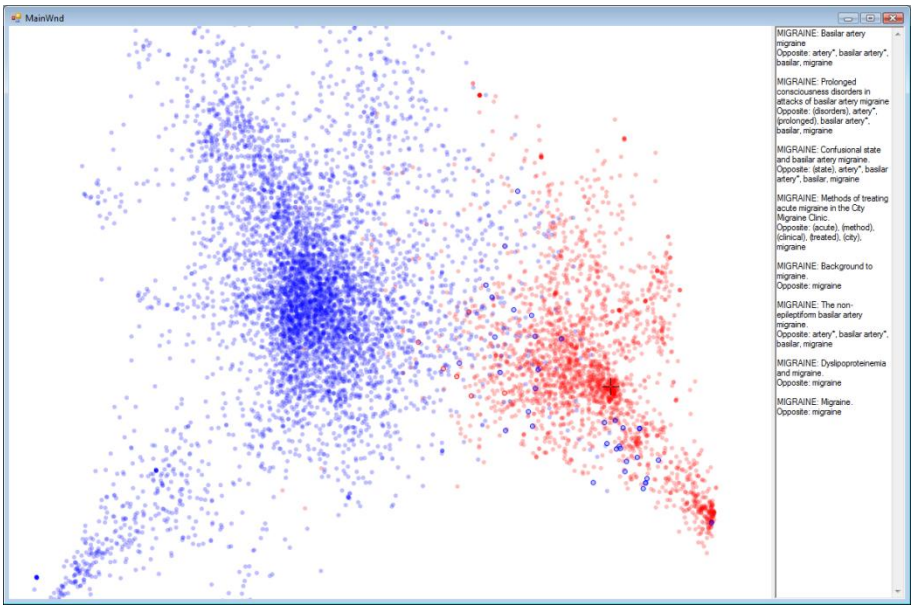

Figure 5: Proposed user interaction oriented application for b-term selection

The user is able to maintain two lists of terms: the list of domain-specific stop terms (such as patient, disease, treatment) and the list of b-terms.

- The list of stop terms can be initially populated with domain-specific stop terms provided with the domain (if available).
- The user can specify stop terms and b-terms through a dialog window, without resorting to visualization. In the dialog window, the user is presented with a sorted list of terms.
- The terms are sorted according to the BP weight. This tends to rank relevant b-terms higher than other non-stop terms, which means that relevant b-terms are presented at the top of the list, mixed with domain-specific stop terms. The user is expected to explore the list, from top to bottom, and to mark terms either as being stop terms, b-terms, or neither.
- Besides BP we also tested bisociativity index measure (BI[2]). At first glance, this measure is comparable to our new BP measure. Further experiments need to be conducted to assess the quality of the two measures (e.g. compute area under ROC curve).

[2] Bisociativity index (BI) as described in [7] is

$$BI(t^A, t^B) = \sum_{i=0}^{n} \left(\sqrt[k]{tf_i^{\mathrm{A}} \cdot tf_i^{\mathrm{B}}} \cdot \left(1 - \frac{|\tan^{-1}(tf_i^{\mathrm{A}}) - \tan^{-1}(tf_i^{\mathrm{B}})|}{\tan^{-1}(1)}\right)\right), tf_i^{\mathrm{A}}, tf_i^{\mathrm{B}},$$

Where $tf_i^{\mathrm{A}}, tf_i^{\mathrm{B}} \in [0,1]$. Note that by definition the BI is a property of a link (between term t^{A} and term t^{B}) while BP measure is a property of a term. Therefore, we compute, for each term t, two BI values, namely bisociativity index between "migraine" centroid and the term, BI(t, mig.), and bisociativity index between "magnesium" centroid and the term, BI(t, mag.). To assign BI to a term, we decided to use the following formula: $BI(t) = min(BI(t, mig), BI(t, mag))$.

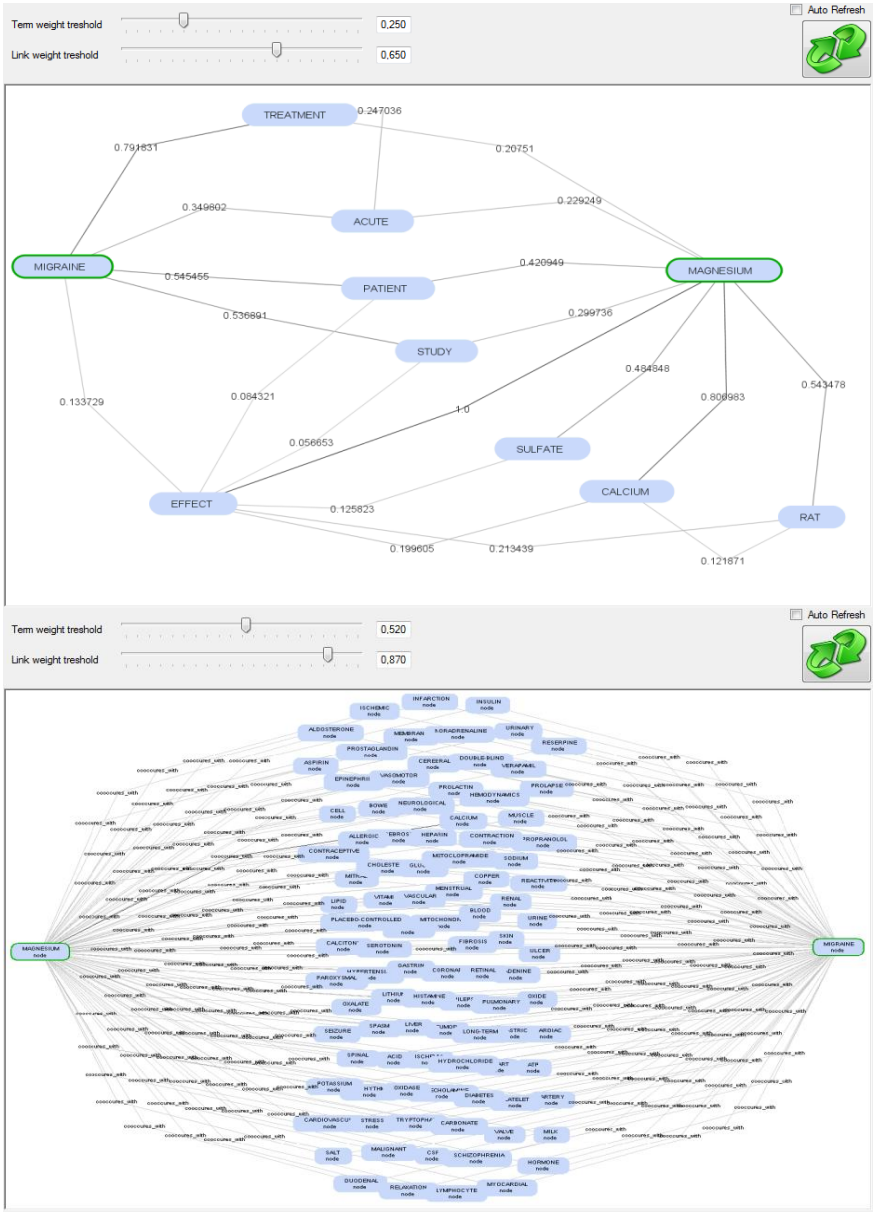

Figure 6: Bisociation visualizer application: user can display a network of terms and interactively browse through it. One also has the option to filter out terms and/or links with weights lower than some specified value (the slider controls at the top of the application). Both images show the application on the same data, with the only difference that the top image shows higher threshold setting than the bottom one. It is clear that increasing threshold manifests in more nodes and links displayed.

5 Network Creation and Exploration

The list all the terms with their b-potential values, the list of stop terms, and the list of b-terms (marked by user) are passed to this phase from the selection step.

In this step, an expert goes through the data visualised as a graph and tries to deduct some new knowledge out of it. All the information derived in the previous steps - most notably b-potential weights ant b-terms labels play crucial role here by enabling

the expert to see deeper into the data, since the system is displaying data that are more relevant.

The actual graph displayed to the user is constructed out of the terms using their BP weights in the following way:

- Nodes are terms (all terms including b-terms but without stop terms). Only those nodes are used that have their BP greater than some preset node weight threshold.
- Links between nodes are calculated using bisociativity index measure (BI). Only those links are generated that have BI weight greater than some preset link weight threshold.

Graph can be dynamically displayed and explored. Currently, the debate, which system to use for visualization is still open. The possibilities which to employ include:

- BMVIS: Biomine visualization,
- CET (Creative Exploration Toolbox),
- Latino
- Orange for WS

Besides standard functions for interactively navigating through graph (moving, zooming, unzooming, ...) the exploration step contains also two filtering controls. With them, the expert controls the size of the graph displayed and implicitly sets the ratio of importance between b-potential and bisociativity index measures. These two controls are be implemented as:

- Node weights (BP) threshold slider – only those nodes are displayed that have weight higher than the weight set on this control (see Figure 6).
- Link weights (BI) threshold (slider) – only those links are displayed that have weight higher than the weight set on this control. (see Figure 6).

Conclusion and Further Work

This work presents a new methodology designed to help bridging the contexts in the domains where there is an overflow of data present. The usage is envisioned as a heavily user interactive tool, which helps an expert to find the interesting terms and concepts. These are found by connecting his domain with some other field of research, in order to come to new, breakthrough discoveries. Even better usage scenario is when there are two experts from different fields interested in cooperation. In this case, the system is an ideal concept generator, which helps both scientists to start communicating in the terms that are familiar with, and meanwhile generating the promising ideas for research.

In this paper, we illustrate the four-step methodology consisting of: preprocesing the data, identifying b-term candidates, selecting b-terms, and network exploration. The main research contribution of this work is concentrated around finding and evaluating the measure for identification of candidate b-terms. In two benchmarks, we provide preliminary evidence that using b-potential (heuristics designed for b-term identification) we achieve approximately 7 times boost in b-term identification compared to not using any strategy. In practice, this means that lists offered to the user for selection of the b-terms are 7 times more probable to contain b-terms compared to showing him or her just the random list of terms.

Although we are satisfied with initial results of our approach, we still need to conduct much more testing, improve the methodology at certain points and implement user interface part of the system. Currently we see next promising ways to follow:

- Improve the technique for discovery of bisociative knowledge from networks. Currently we propose very simple model of expert-system user interface in the

last section of this paper. This model should be enhanced with some stronger functionality, which would enable the expert to see "deeper" in the data.

- Better b-potential heuristic. On ROC curve diagrams, we spotted that some heuristics express themselves with interesting and not anticipated function shapes. Even though we select the best heuristic for b-potential, the strange shapes are suggesting us that there might still be some hidden knowledge inside. Hence, by combining different heuristics (maybe via some voting models) it is perhaps possible to unveil this hidden knowledge and even further improve b-term detection rate.
- Implement the user interface to the level, where we would be able to test the whole methodology, including real data and collaborative expert trying to find some novel knowledge. In case of success, this would be the ultimate proof of our concept.

Acknowledgements

The work presented in this paper was supported by the European Commission under the 7th Framework Programme FP7-ICT-2007-C FET-Open project BISON-211898, and by the Slovenian Research Agency grants Knowledge Technologies (P2-0103), Systems Biology (J4-2228), and Semantic SoKD (J2-2353).

References

1. Koestler, A.: The Act of Creation. The Macmillan Co. (1964)
2. Sevon, P., Eronen, L., Hintsanen, P., Kulovesi, K., Toivonen, H.: Link discovery in graphs derived from biological databases. In : Proc. of 3rd International Workshop on Data Integration in the Life Sciences (DILS'06) (July 2006)
3. Swanson, D. R., Smalheiser, N. R., Torvik, V. I.: Ranking indirect connections in literature-based discovery : The role of Medical Subject Headings (MeSH). JASIST 57(11), 1427-1439 (2006)
4. Petric, I., Urbancic, T., Cestnik, B., Macedoni-Luksic, M.: Literature mining method RaJoLink for uncovering relations between biomedical concepts. Journal of Biomedical Informatics 42(2), 219--227 (2009)
5. Juršič, M., Mozetič, M., Lavrač, N.: Learning Ripple Down Rules for Efficient Lemmatization. In : Information Society - IS 2007 - Data Mining and Data Warehouses, Ljubljana, vol. A, pp.206-209 (2007)
6. PubMed. (Accessed March 2010) Available at: http://www.ncbi.nlm.nih.gov/pubmed/
7. Segond, M., Borgelt, C.: "BisoNet" Generation using Textual Data. In : Proceedings of Workshop on Explorative Analytics of Information Networks at ECML PKDD (2009)