Biomedical Literature Exploration through Latent Semantics

Sérgio Matos[a], Hugo Araújo[a], José Luís Oliveira[a]
[a] DETI/IEETA, Universidade de Aveiro, Portugal

KEYWORD

Information Retrieval
Literature Exploration
Latent Semantic Analysis

ABSTRACT

The fast increasing amount of articles published in the biomedical field is creating difficulties in the way this wealth of information can be efficiently exploited by researchers. As a way of overcoming these limitations and potentiating a more efficient use of the literature, we propose an approach for structuring the results of a literature search based on the latent semantic information extracted from a corpus. Moreover, we show how the results of the Latent Semantic Analysis method can be adapted so as to evidence differences between results of different searches. We also propose different visualization techniques that can be applied to explore these results. Used in combination, these techniques could empower users with tools for literature guided knowledge exploration and discovery.

1 Introduction

In a thriving and evolving research area such as biomedicine, where the scientific literature is the main source of information, containing the outcomes of the most recent studies, being able to efficiently explore the literature or conduct a systematic literature search is fundamental. However, the fast increasing amount of articles published in this field creates difficulties in the way information can be searched and used by researchers [Shatkay, 2005; Lu, 2011]. Moreover, the inherent interrelations between concepts and the different perspectives, or themes, under which a given idea or concept may be studied are also important and introduce another level of complexity in exploring the literature associated with this specific domain. Given a disease, for example, researchers may be interested on different aspects, from the underlying genetics, to previous studies using a particular laboratory technique or experiment, to more clinically oriented information.

Biomedical researchers may use services such as PubMed, Google Scholar, or one of the many available tools specifically aimed at this domain [Lu, 2011]. The results of these tools are usually presented in the form of a list of documents, with no indication of how they are related between each other or to the search expression. Users are familiarized with this type of output, but exploring the information contained in the documents, and relating pieces of information obtained from various documents, is difficult and inefficient. Moreover, although it is trivial to perform simple literature searches using such tools, translating a more complex information need to a search expression that returns a satisfactory (from the user perspective) set of results is a much more difficult task. This means that, especially when conducting an exploratory literature search, researchers often have to sift through vast amounts of documents. In that process, they will iteratively check related references and reformulate their queries in view of the knowledge they gather at each point, looking for more specific or more relevant information [Kim and Rebholz-Schuhmann, 2008]. In the past years, different approaches have been proposed and evaluated as ways to facilitate simple literature searches, exploratory analysis and/or literature-based knowledge discovery. Some of these works have focused on literature-based methods for data analysis, generally extracting from the literature some sort of semantic descriptors to help compare or associate experimental data, such as sets of

genes [Homayouni et al., 2005; Chagoyen et al., 2006; Done et al., 2010; Xu et al., 2011]. In this work, we focus on facilitating literature searches and exploratory literature analysis. Our aim is to explore ways to overcome the limitations of current literature retrieval tools, helping researchers in searching and exploring the wealth of information enclosed in the scientific biomedical literature, by structuring the search results according to their latent semantics. In this work, we evaluate the use of Latent Semantic Analysis (LSA) for structuring the results of literature searches into high-level semantic divisions, or themes. LSA is a natural language processing technique that allows analysing the relations between a set of documents and the terms that belong to those documents, by representing them in a multi-dimensional semantic space [Landauer et al., 1998]. LSA is obtained by applying a Singular Value Decomposition (SVD) to the term-document matrix representation of a corpus, and then selecting the most important dimensions from this decomposition. Each dimension in this semantic space is then represented as a linear combination of words from a fixed vocabulary (the words that compose the documents in the collection), and is usually represented by the list of words with highest coefficient for that dimension. Since each dimension can be regarded as a different view of the results, we argue that looking at a given dimension corresponds to exploring the documents from a different perspective. This analysis allows organizing the documents according to the themes they include, providing an intuitive way for exploring the document collection. The next sections are organized as follows: related works are presented in Section 2, Section 3 describes the proposed methodology, Section 4 presents and discusses the results obtained. Final conclusions are made in Section 5.

2 Related work

PubMed is the most popular and widely used biomedical literature retrieval system. It combines boolean and vector space models for document retrieval with expert assigned indexing terms from the Medical Subject Headings (MeSH) controlled vocabulary, and provides access to over 20 million citations [Lu, 2011]. However, as most information retrieval (IR) systems, PubMed uses query proximity models to search documents matching a user's query terms, returning results in the form of a list. Similarly, several other IR tools have been developed based on the MEDLINE literature database, including GoPubMed [Doms and Schroeder, 2005], XplorMed [Perez-Iratxeta et al., 2003], Chilibot [Chen and Sharp, 2004], FACTA [Tsuruoka et al., 2008], EBIMed [Rebholz-Schuhmann et al., 2007] and PolySearch [Cheng et al., 2008]. All of these tools allow for some form of literature exploration. GoPubMed categorizes the results according to Gene Ontology (GO) and MeSH terms mentioned in the documents. Users can select terms in the ontology tree to interactively refine the query and filter the results. XplorMed extracts the words most strongly associated with an initial set of abstracts, and associations between these words. Based on these word associations it is then possible to iteratively extend and refine the set of abstracts. Chilibot accepts two or more search terms and processes the results using linguistic analysis to identify sentences describing relations between the search terms, displaying the results as a list of sentences or as a graph, from which inter-concept relations can be inferred. FACTA and EBIMed show the results of a search as a table of the most relevant concepts mentioned in the returned documents, as well as the sentences where each of those concepts co-occurs with other terms. PolySearch also identifies concept associations, but the type of concept used as start and end points are defined initially, from a closed set of options. Although these tools allow exploring the results in one way or another, they all work at the term or concept level. None of these tools identifies higher-level semantics within the documents, which would allow identifying implicit relations and assist the exploration of the results by the user. More recently, the focus has been on the use of Latent Semantic Analysis (LSA) [Landauer et al., 1998; Deerwester et al., 1990] and probabilistic topic models such as Latent Dirichlet Allocation (LDA) [Blei et al., 2003]. These models allow identifying the relevant themes or concepts

associated to a document. Zheng et al. [Zheng et al., 2010] and Jahiruddin et al. [Jahiruddin et al., 2010] have proposed document conceptualization and clustering frameworks based on LSA and domain ontologies. Zheng et al. base their methods on a user-defined ontology, matching the terms that compose this ontology to phrase chunks extracted from the documents in a collection. LSA is then applied to the term-document matrix constructed from these matches. The authors demonstrated that the application of LSA considerably improves document conceptualization. Jahiruddin et al. integrate natural language processing (NLP) and semantic analysis to identify key concepts and relations between those concepts. Their method starts by selecting candidate terms from the noun phrases in the document collection. LSA is then applied to the matrix constructed from these terms in order to identify the most important ones. Relation extraction is also performed, by identifying relational verbs in the vicinity of biomedical entities and concepts. Validated concepts and interactions are then used to construct a semantic network, which can be used to navigate through the information extracted from the documents. In this work, we also use LSA to identify the latent semantics within a corpus. We then explore how the LSA results can be organized in order to structure the results of search queries, highlighting the most important topics related to each different query. Throughout the paper, we will interchangeably use the terms 'topic' and 'theme' to refer to the underlying theme(s) corresponding to a given LSA dimension. However, the use of the term 'topic' should not be confused with its meaning as used in (probabilistic) topic models.

3 Methods

As mentioned before, our aim is to structure the results of a literature search into high-level themes, or topics, in order to help researchers search and explore the information enclosed in the scientific biomedical literature. Following our rationale, this process should be user-centred and user-driven and should consider two interlinked factors: query reformulation and exploration of results. Therefore, instead of focusing on automatic knowledge discovery, we propose to give users the tools to search and explore the literature in an iterative and systematic manner, guided by their knowledge of the domain and supported by graphical and intuitive representations of the relations identified between the resulting documents and between concepts mentioned in those results. In order to test and validate our proposal, we applied our methods to a corpus related to neurodegenerative disorders, containing around 135 thousand Medline documents composed by the title and abstract of the publication. The PubMed query used to obtain the documents was: "Neurodegenerative Diseases"[MeSH Terms] OR "Heredodegenerative Disorders, Nervous System"[MeSH Terms]. Articles in languages other than English or not containing an abstract were discarded. The list of MeSH term assigned to each document was also obtained. Our approach consists of an offline phase followed by two online steps. In the offline phase we calculate the LSA transformation matrix and transform the corpus to the LSA space. This operation is performed once for the complete corpus, and the transformation matrix is kept for transforming the user queries into the semantic space. Given a query, the two online steps consist of: a) identifying the significant topics for that query and ranking the most relevant documents within each topic; and, b) obtaining a list of representative MeSH terms for each topic. These steps are described in the next sections.

3.1 Document representation

Before applying LSA, the corpus was processed in order to identify vocabulary terms representing concepts from the biomedical domain. The vocabulary was created based on the UMLS Metathesaurus [NLM, 2013]. The Neji framework[1] was used to pre-process the corpus and perform concept recognition. A total of around 45 thousand distinct concepts were identified in the documents, producing a concept-document matrix with 135 thousand lines and 45 thousand columns.

[1] http://bioinformatics.ua.pt/neji/

3.2 Latent semantic analysis

Following the concept recognition step, the documents are represented by the set of concepts occurring in them, and the term-document matrix used for calculating the LSA is constructed from this representation instead of through the most common token level (bag-of-words) representation.

The Gensim framework [Rehurek and Sojka, 2010] was used for calculating LSA. This framework implements an incremental method for performing the SVD step in the LSA calculation, making it more efficient and allowing more documents to be added to the LSA space.

3.3 Ranking relevant documents

This step starts by selecting only the most relevant LSA dimensions (topics) for the query, given the query representation in the LSA space. A threshold is applied to eliminate those dimensions to which the query is less related, i.e. has a smaller coefficient after transforming it to the LSA space. This is an important step since the LSA transformation is applied to the complete corpus and will in general include many dimensions that are not relevant for a given query. Next, we proceed to ranking the relevant documents within each of the selected topics.

Given the selected dimensions, a first approach for finding the most similar documents is simply to project the documents and query values in each dimension and calculate the distance between them (Fig. 1). This, however, does not consider the relevance of the documents for each topic. Inspecting Fig. 1, although the distance between the query Q and document A is similar to the distance between Q and document C for dimension 1, C is much more related to this dimension than A and should therefore have a higher score. To account for this, we calculate a score for each document in each dimension, given by the product of the document's LSA coefficient for that dimension and the similarity between the query and the document, calculated as the cosine similarity on the LSA space:

$$Score(D_k, T_j) = \left| Sim(D_k, Q) \times V(D_k, T_j) \right| \quad (1)$$

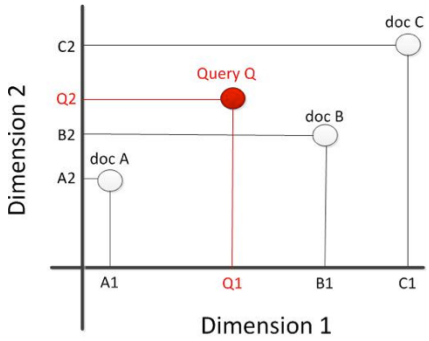

Fig. 1. Document and query projection into the LSA dimensions.

where $V(D_k, T_j)$ is the LSA coefficient for document D_k in dimension T_j and $Sim(D_k, Q)$ is the cosine similarity between document D_k and the query Q, in the LSA space.

Finally, we use a second threshold to filter these scores, obtaining the most similar documents for the query regarding each of the considered topics.

3.4 Representative terms

In order to represent each identified topic and facilitate the exploration of results by users we make use of the MeSH terms assigned to the documents by expert annotators. These indexing terms represent the major concepts in each Medline article, and therefore provide a semantic representation of the documents. The important aspect to consider is that we expect that each cluster of results represent a distinct topic or theme. Therefore, the documents assigned to each dimension, after the previous step, should not only be different but should also be focused on different themes. To evaluate this, we can compare the set of MeSH terms assigned to the documents in different topics to see how different they are. In order to do so, we first define an association score for each MeSH term in each topic, creating a vector representation that we then use to compare the topics.

Table 1. Top results in topics 12 and 25 for the query term "Dopamine"

Topic 12

20926973	Intense dopamine innervation of the subventricular zone in Huntington's disease
10838590	Neuronal cell death in Huntington's disease: a potential role for dopamine
9822765	Dopamine modulates the susceptibility of striatal neurons to 3-nitropropionic acid (…)
10829080	Severe deficiencies in dopamine signaling in presymptomatic Huntington's disease (…)
17065224	Dopamine enhances motor and neuropathological consequences of polyglutamine (…)

Topic 25

20926973	Polymorphisms of dopamine receptor and transporter genes and Parkinson's disease
10838590	Higher nigrostriatal dopamine neuron loss in early than late onset Parkinson's disease?
9822765	Brain dopamine receptors: 20 years of progress
10829080	Involvement of ventrolateral striatal dopamine in movement initiation and execution
17065224	Recent discoveries on the function and plasticity of central dopamine pathways

To define the association score between a given topic and a MeSH term, we consider the LSA coefficient of each document in that topic that contains that MeSH term, as well as the relative ranking of the document in that dimension. The ranking is an important factor since the value of the coefficients for the most relevant documents may vary significantly across different dimensions, and we want to force that if a MeSH term is assigned to the top ten documents in a topic, it should have a higher score for that topic than for a topic where it is assigned to lower ranking documents. Therefore, we define the association score as the sum of the document coefficients, normalized by the ranking of that document within that topic:

$$Score\left(M_i, T_j\right) = \sum_{k=1}^{N} \frac{V\left(D_k, T_j\right)}{Rank\left(D_k, T_j\right)}, \quad M_i \in D_k, D_k \in T_j \qquad (2)$$

where $V\left(D_k, T_j\right)$ is LSA coefficient of document D_k for topic T_j and $Rank\left(D_k, T_j\right)$ is the ranking of document D_k in topic T_j, and N is the number of documents in topic T_j.

4 Results and discussion

Using the methods proposed in the previous section, it is possible to organize the literature search results into separate lists, each associated to a certain theme. Different numbers of LSA dimensions were considered, varying from 100 to 400. There was no significant variation in the results, and here results are shown for 100 dimensions. The documents retrieved by LSA similarity will be distributed across these topics, allowing an easier navigation. Also, although a given document may occur in more than one topic, which is expected since articles discuss interrelated subjects, it will appear in different ranking positions in each result list. Therefore, users looking at two different topics will see two different sets of results. This also justifies using the document ranking when calculating the score for each MeSH term and topic pair, since the most important results for the users are the top ones in each result list. Table 1 shows the first five results for the query "Dopamine", in topics 12 and 25. As can be noticed, the results lists are significantly different between topics, illustrating how the retrieved results are organized around separate themes.

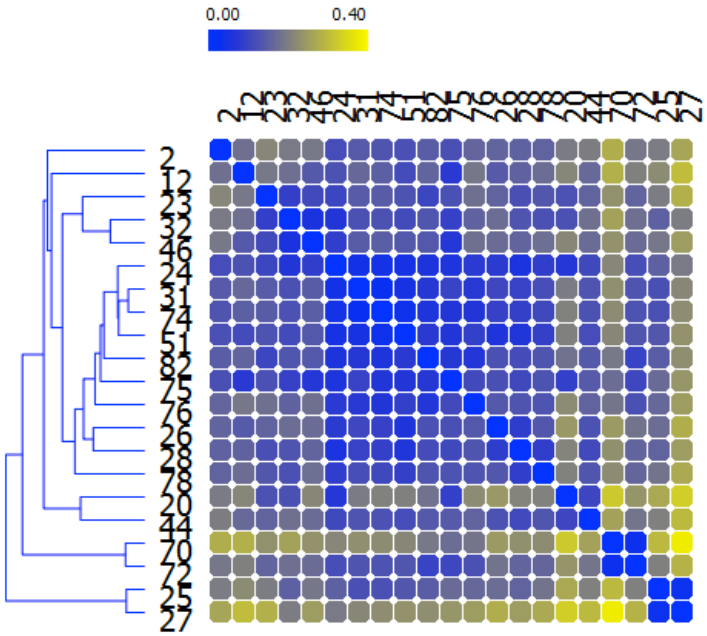

Fig. 2. Inter-topic distances for the query "Dopamine". Each cell in the heatmap represents the distance between a pair of topics.

4.1 Inter-topic distances

Given the topic description as a vector of MeSH terms and scores, we can visually analyze the differences or similarities between pairs of topics. Since MeSH terms are assigned to articles based on the full-text content, and represent the themes of the article, using this as a measure of semantic similarity between the topics is a logical choice. We define the inter-topic distance as

$$Dist(T_j, T_k) = 1 - \frac{A \cdot B}{\|A\|\|B\|} = 1 - \frac{\sum_{i=1}^{M} a_i \times b_i}{\sqrt{\sum_{i=1}^{M} a_i^2} \cdot \sqrt{\sum_{i=1}^{M} b_i^2}} \quad (3)$$

where ***A*** and ***B*** are the MeSH term vector representations of topics *j* and *k*, respectively, with coefficients given by Eq. 2, and *M* is the number of distinct MeSH terms. Fig. 2 shows the inter-topic distances corresponding to the query term "Dopamine", illustrated as a heatmap. The dendogram is also shown, allowing identifying the most similar topics. Such a representation helps identifying topics that are similar and those that are more distant. For example, topics 25 and 27 and topics 70 and 72 are very similar between each other, and at the same time each pair is very different from the remaining topics. These similarities could arise from LSA dimensions which are somewhat overlapping or, more interestingly, to distinct but somehow related topics.

4.2 Multidimensional scaling

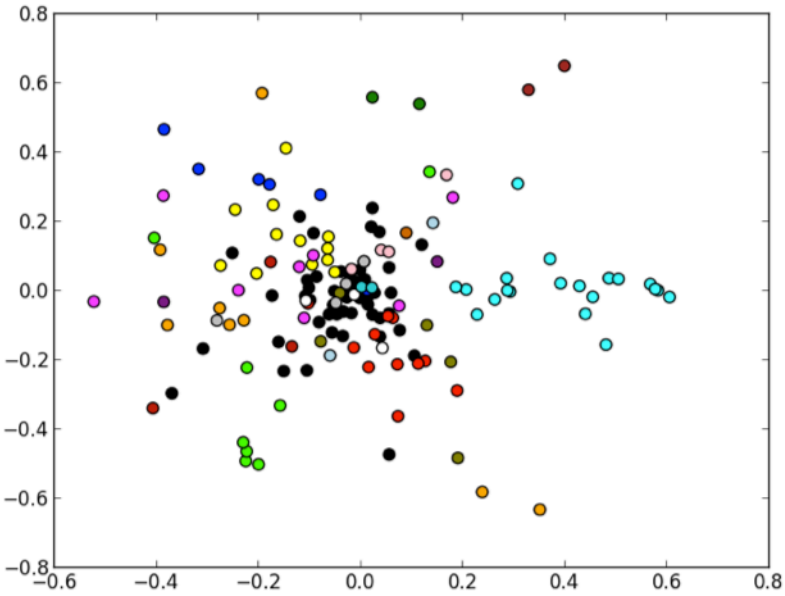

Fig. 3. Distribution of documents in the geometrical space created by MDS. The top 20 documents for each topic relevant for the query "Dopamine" are shown. The colour of the circle indicates the topic to which the document is assigned; black circles indicate documents assigned to more than one topic.

An intuitive way to represent the different topics in the results is by using multidimensional scaling (MDS), an exploratory technique used to visualize proximities in a low dimensional space [Van Deun et al., 2007]. It uses a set of data analysis methods to detect underlying relations between entities from a correlation matrix and represents them in a geometrical space, being possible to visualize those relations. The MDS algorithm starts by determining the specified topics and initial coordinate matrix, followed by calculating the distances between the several entities in the matrix. The next step relies on optimizing the matrix scale using a measure of goodness-of-fit referred to as stress. After this evaluation the coordinates are updated and the scale is optimized once again. This process is repeated until the stress is minimized. Using MDS, the documents or the topics resulting from a search can be displayed on a two-dimensional space, where they appear distributed according to their similarities. Fig. 3 shows the result of MDS for the query "Dopamine", using the LSA cosine similarity between each pair of documents. Only the top 20 documents in each topic were considered. Each document is represented by a circle, coloured according to the topic that document belongs to; black circles represent documents assigned to more than one topic. Although some separation is lost due to the reduction to two dimensions, some clusters of documents from the same topic are still apparent. This representation, used within a literature retrieval system, would allow users to navigate the results while visualizing the relations or similarities between the resulting documents.

4.3 Word clouds

Another way of visualizing the results of LSA is by representing the topics by word clouds. Word clouds are popularly used as a visual representation for textual data, with each word differing in size, colour, position or font, depending on a weight, usually defined by its frequency in a document or collection of documents. For LSA, one could use the coefficient for each word in a LSA dimension, illustrating the words that better describe that dimension. However, for our proposed aim, this would not suffice, as this representation would not be associated with the searched query.

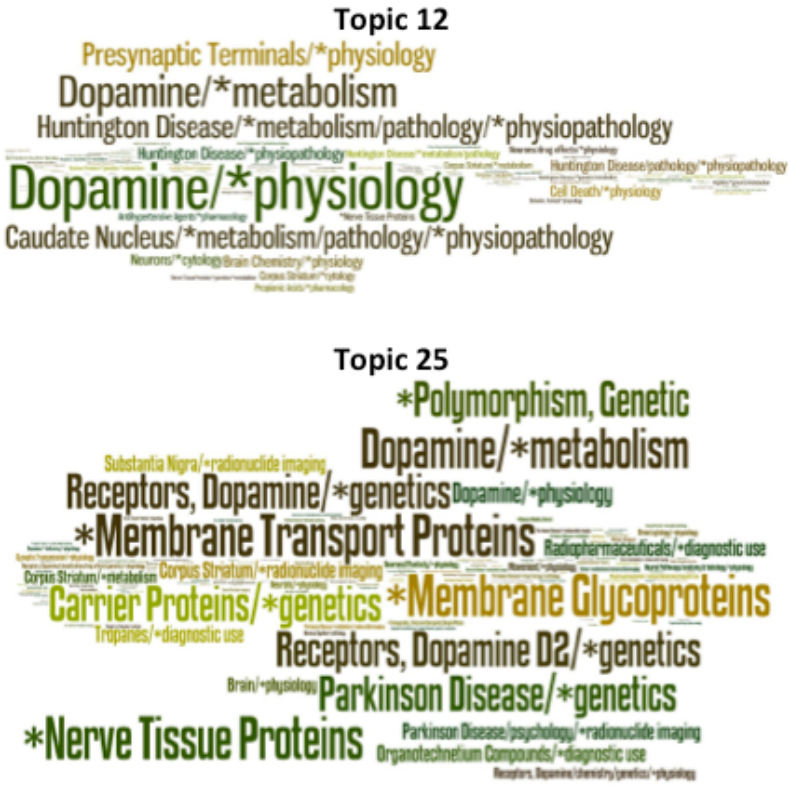

Fig. 3. MeSH term cloud for topics 12 and 25 showing the most relevant terms in these topics for the query "Dopamine". The size of the font is proportional to the score of that term for the topic and query combination. The cloud was created in the Wordle website (http://www.wordle.net/)

Instead, we want that the word cloud for a topic reflect the most important terms for that topic given the specific query. Therefore, we follow the same approach as before, and represent the most significant MeSH terms for resulting documents assigned to that topic. In this case, we want the word cloud for a topic to reflect the most important terms for that topic given the specific query. Therefore, we use the most significant MeSH terms for resulting documents assigned to that topic. Fig. 4 illustrates the MeSH term cloud corresponding to topics 12 and 25 for the query ``Dopamine''. From the most prominent terms, one can identify that topic 12 is about physiology and physiopathology in Huntington's disease, while topic 25 is mostly about receptors, transport and metabolism of Dopamine. It is important to emphasize that, although the LSA dimensions are defined for the entire corpus and are kept constant, the word cloud for this same topic would be different for a different query, as given by the score defined in Eq. 2. Combining this type of visualization with the MDS result described in the previous section would give users of a literature retrieval system a rich view of the information contained in their results, facilitating its exploration.

5 Conclusions

We have described an approach for structuring the results of a literature search based on the latent semantic information extracted from the documents in a corpus, as expressed by LSA.

Moreover, we show how the results of LSA can be adapted so as to evidence differences between results of different queries and propose several visualization techniques that can be applied to explore these results. Further work is required for evaluating how users would benefit from the proposed solutions. Although objective evaluation of methods such as the one proposed here is usually very difficult, the results presented indicate that methods for structuring literature search results, used in combination within a literature retrieval system, could empower users with tools for literature guided knowledge exploration and discovery.

6 Acknowledgment

This research work was partially funded by FEDER through the COMPETE programme and by national funds through FCT – "Fundação para a Ciência e a Tecnologia" under project number PTDC/EIA-CCO/100541/2008 (FCOMP-01-0124-FEDER-010029). Sérgio Matos is funded by FCT under the Ciência2007 programme.

7 References

[Blei et al., 2003] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research, 3 (2003) 993–1022.

[Chagoyen et al., 2006] M. Chagoyen, P. Carmona-Saez, H. Shatkay, J.M. Carazo, A. Pascual-Montano, Discovering semantic features in the literature: a foundation for building functional associations, BMC Bioinformatics 7 (2006) 41.

[Chen and Sharp, 2004] H. Chen, B.M. Sharp, Content-rich biological network constructed by mining PubMed abstracts, BMC Bioinformatics 5 (2004) 147.

[Cheng et al., 2008] D. Cheng, C. Knox, N. Young, P. Stothard, S. Damaraju, D.S. Wishart, PolySearch: a webbased text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites, Nucleic Acids Resesarch, 36 (suppl 2) (2008) W399–W405.

[Deerwester et al., 1990] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science, 41 (1990) 391–407.

[Doms and Schroeder, 2005] A. Doms, M. Schroeder, GoPubMed: exploring PubMed with the Gene Ontology, Nucleic Acids Research 33 (Web Server issue) (2005) W783–W786.

[Done et al., 2010] B. Done, P. Khatri, A. Done, S. Draghici, Predicting novel human Gene Ontology annotations using semantic analysis, IEEE/ACM Trans. On Computational Biology and Bioinformatics 7 (1) (2010) 91–99.

[Homayouni et al., 2005] R. Homayouni, K. Heinrich, L. Wei, M.W. Berry, Gene clustering by latent semantic indexing of Medline abstracts, Bioinformatics 21 (1) (2005) 104–115.

[Jahiruddin et al., 2010] Jahiruddin, M. Abulaish, L. Dey, A concept-driven biomedical knowledge extraction and visualization framework for conceptualization of text corpora, Journal of Biomedical Informatics 43 (2010) 1020–-1035.

[Kim and Rebholz-Schuhmann, 2008] J.J. Kim, D. Rebholz-Schuhmann, Categorization of services for seeking information in biomedical literature: a typology for improvement of practice, Briefings in Bioinformatics 9 (6) (2008) 452–465.

[Landauer et al., 1998] T.K. Landauer, P.W. Foltz, D. Laham, An introduction to latent semantic analysis. Discourse Processes 25 (2-3) (1998) 259–284.

[Lu, 2011] Z. Lu, PubMed and beyond: a survey of web tools for searching biomedical

literature, Database 2011 (2011) baq036.

[NLM, 2013] UMLS Metathesaurus Fact Sheet. Available at http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

[Perez-Iratxeta et al., 2003] A. Perez-Iratxeta, A.J. Pérez, P. Bork, M.A. Andrade, Update on XplorMed: a web server for exploring scientific literature, Nucleic Acids Research 31 (2003) 3866–3868.

[Rebholz-Schuhmann et al., 2007] D. Rebholz-Schuhmann, H. Kirsch, M. Arregui, S. Gaudan, M. Riethoven, P. Stoehr, EBIMed—text crunching to gather facts for proteins from Medline, Bioinformatics 23 (2) (2007) e237–e244.

[Rehurek and Sojka, 2010] R. Rehurek, P. Sojka, Software Framework for Topic Modelling with Large Corpora, in: Proceedings of the LREC 2010 workshop New Challenges for NLP Frameworks, Valetta, Malta, 2010, pp. 46–50.

[Shatkay, 2005] H. Shatkay, Hairpins in bookstacks: information retrieval from biomedical text, Briefings in Bioinformatics 6 (3) (2005) 222–238.

[Tsuruoka et al., 2008] Y. Tsuruoka, J. Tsujii, S. Ananiadou, FACTA: a text search engine for finding associated biomedical concepts, Bioinformatics 24 (21) (2008) 2559–2560.

[Van Deun et al., 2007] K. Van Deun, W.J. Heiser, L. Delbeke, Multidimensional unfolding by nonmetric multi-dimensional scaling of Spearman distances in the extended permutation polytope, Multivariate Behavioral Research 42 (1) (2007) 103–132.

[Xu et al., 2011] L. Xu, N. Furlotte, Y. Lin, K. Heinrich, M.W. Berry, E.O. George, R. Homayouni, Functional cohesion of gene sets determined by latent semantic indexing of Pubmed abstracts, PLOS One 6 (4) (2011) e18851.

[Zheng et al., 2010] H.-T. Zheng, C. Borchert, Y. Jiang, A knowledge-driven approach to biomedical document conceptualization, Artificial Intelligence in Medicine 49 (2010) 67–78.