Structuring and Exploring the Biomedical Literature Using Latent Semantics

Sérgio Matos, Hugo Araújo, and José Luís Oliveira

DETI/IEETA, University of Aveiro, 3810-193 Aveiro, Portugal
{aleixomatos,hugo.rafael,jlo}@ua.pt

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

Being able to conduct a systematic literature search is an essential skill for researchers in any field. 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, this becomes even more important. However, the fast increasing amount of articles published in this field is creating difficulties in the way information can be efficiently searched and used by researchers [8, 6].

Another important aspect is the inherent interrelations between concepts. Additionally, researchers may be interested in studying a given idea or concept from a particular perspective. Given a disease, for example, they 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.

Although many literature retrieval tools have been developed for this particular domain, many limitations are still present, specially in the form the results are presented, forcing users to continually reformulate their queries in view of information they gather at each point, looking for more specific or more relevant information [4].

In this work, we evaluate the use of Latent Semantic Analysis (LSA) for structuring the results of a literature search into high-level semantic divisions, or themes. LSA is a natural language processing technique that allows analysing the relations

S. Omatu et al. (Eds.): *Distrib. Computing & Artificial Intelligence,* AISC 217, pp. 609–616.
DOI: 10.1007/978-3-319-00551-5_72

between a set of documents and the terms that belong to those documents, by representing them in a multi-dimensional semantic space [5]. Each dimension in this semantic space is 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 value for that dimension. Since each dimension can be regarded as a different view of the results, 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 Medical Subject Headings (MeSH) categories, giving researchers access to over 20 million citations [6] . 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 based on the MEDLINE literature database have been developed (see [6] for a comprehensive list of tools).

More recently, the focus has been on the use of Latent Semantic Analysis (LSA) [5, 2] and probabilistic topic models such as Latent Dirichlet Allocation (LDA) [1] . These models allow identifying the relevant themes or concepts associated to a document. Zheng et al. [11] and Jahiruddin et al. [3] 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 use LSA to identify the latent semantics within a corpus, and borrow the term topic to refer to the underlying theme(s) for a given LSA dimension. However, this should not be confused with the meaning of this term within (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. We apply our method 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 identifying and ranking the relevant documents within each topic and obtaining a list of representative MeSH terms for each topic. These steps are described in the next sections.

3.1 Corpus Processing and LSA

Before applying LSA, the corpus was processed in order to identify terms from a fixed vocabulary. This vocabulary contains terms from the biomedical domain, and was created based on the UMLS Metathesaurus [9]. The documents are therefore represented by the set of domain terms occurring in them, instead of through the common bag-of-words approach, and the term-document matrix used for calculating the LSA is constructed from this representation. The Gensim framework [7] was used for calculating LSA.

3.2 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. Next, we proceed to ranking the relevant documents within each of the selected topics. For this, and given the selected dimensions, we want to consider two aspects: the similarity between the query and the document, and the association of the document to each given topic. To reflect these two aspects, we propose the following score:

$$Score(D_k, T_j) = |Sim(D_k, Q) \times V(D_k, T_j)|, \tag{1}$$

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.

3.3 Identifying Representative Terms

In order to represent each identified topic and facilitate the exploration of results by users we make use of the MeSH terms, which represent the major concepts in each Medline article, selected by expert annotators. The important aspect to consider is that we expect that each cluster of results represents 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. In order 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 calculate an association score for each MeSH term in each topic, creating a vector representation that we then use to compare the topics. For each MeSH terms and each topic, this score is calculated as the sum of the coefficients of the documents containing the MeSH term, normalized by the corresponding rank of that document in the topic, as shown in Eq. 2.

$$Score(M_i, T_j) = \sum_{k=1}^{N_j} \frac{V(D_k, T_j)}{Rank(D_k, T_j)}, \quad M_i \in D_k,\ D_k \in T_j, \tag{2}$$

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 in the rat model of Huntington's disease.
10829080	Severe deficiencies in dopamine signaling in presymptomatic Huntington's disease mice.
17065224	Dopamine enhances motor and neuropathological consequences of polyglutamine expanded huntingtin.
topic 25	
9620058	Polymorphisms of dopamine receptor and transporter genes and Parkinson's disease.
17290452	Higher nigrostriatal dopamine neuron loss in early than late onset Parkinson's disease?
8464534	Brain dopamine receptors: 20 years of progress.
8848171	Involvement of ventrolateral striatal dopamine in movement initiation and execution.
20188048	Recent discoveries on the function and plasticity of central dopamine pathways.

where N_j is the number of documents assigned to dimension T_j, $V(D_k, T_j)$ is the LSA coefficient for document D_k in dimension T_j, and $Rank(D_k, T_j)$ is the ranking position of document D_k in topic T_j.

4 Results

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. 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.

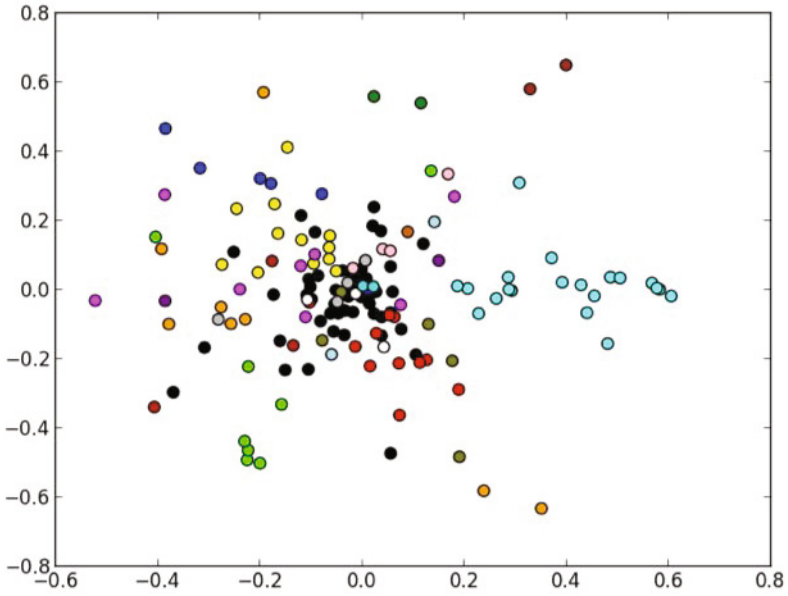

Fig. 1 Distribution of documents in the geometrical space created by MDS. The top 20 documents for each topic relevant for the query "Dopamine" are shown. The colour of the circle indicates the topic to which the document is assigned; black circles indicate documents assigned to more than one topic.

4.1 Multidimensional Scaling

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 [10]. 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. Figure 1 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. 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.

Fig. 2 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/)

4.2 Word Clouds

Another way of visualizing the results of LSA is by representing the topics by word clouds. 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. Figure 2 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.

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.

Acknowledgement. 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.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
2. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science 41, 391–407 (1990)
3. Jahiruddin, Abulaish, M., Dey, L.: A concept-driven biomedical knowledge extraction and visualization framework for conceptualization of text corpora. Journal of Biomedical Informatics 43, 1020–1035 (2010)
4. Kim, J.J., Rebholz-Schuhmann, D.: Categorization of services for seeking information in biomedical literature: a typology for improvement of practice. Briefings in Bioinformatics 9(6), 452–465 (2008)

5. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse Processes 25(2-3), 259–284 (1998)
6. Lu, Z.: PubMed and beyond: a survey of web tools for searching biomedical literature. Database 2011, baq036 (2011)
7. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of LREC 2010 Workshop New Challenges for NLP Frameworks, pp. 46–50. LREC (2010)
8. Shatkay, H.: Hairpins in bookstacks: information retrieval from biomedical text. Briefings in Bioinformatics 6(3), 222–238 (2005)
9. UMLS Metathesaurus Fact Sheet, `http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html`
10. Van Deun, K., Heiser, W.J., Delbeke, L.: Multidimensional unfolding by nonmetric multi-dimensional scaling of Spearman distances in the extended permutation polytope. Multivariate Behavioral Research 42(1), 103–132 (2007)
11. Zheng, H.-T., Borchert, C., Jiang, Y.: A knowledge-driven approach to biomedical document conceptualization. Artificial Intelligence in Medicine 49, 67–78 (2010)