On Topic Categorization of PubMed Query Results

Andreas Kanavos[1], Christos Makris[1] and Evangelos Theodoridis[1,2]

1.Computer Engineering and Informatics Department
University of Patras
Rio, Patras, Greece, 26504
Email: {kanavos, makri, theodori}@ceid.upatras.gr

2.Computer Technology Institute
Rio, Patras, Greece, 26504
Email: theodori@cti.gr

Abstract. Nowadays, people frequently use search engines in order to find the information they need on the Web. Especially Web search constitutes a basic tool used by million researchers in their everyday work. A very popular indexing engine, concerning life sciences and biomedical research is PubMed. PubMed is a free database accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. The present search engines usually return search results in a global ranking making it difficult to the users to browse in different topics or subtopics that they query. Because of this mixing of results belonging to different topics, the average users spend a lot of time to find Web pages, best matching their query. In this paper, we propose a novel system to address this problem. We present and evaluate a methodology that exploits semantic text clustering techniques in order to group biomedical document collections in homogeneous topics. In order to provide more accurate clustering results, we utilize various biomedical ontologies, like MeSH and GeneOntology. Finally, we embed the proposed methodology in an online system that post-processes the PubMed online database in order to provide to users the retrieved results according to well formed topics.

Keywords: semantic topic clustering, pubMed search, MeSH ontology

1 Introduction

Nowadays, people frequently use the Web in order to find the information they need [2]. Especially Web search and online indexing databases are very popular tools used by million researchers in their everyday work. Search engines are an inestimable tool for retrieving Web information. Users place queries by inserting appropriate keywords, and then the search engines return a list of results ranked in order of relevance to these queries. However, an inherent weakness of this information seeking activity is the lack in satisfying ambiguous queries,

and current techniques are generally not able to distinguish between different meanings of a query, and the results characterizing the different meanings will be mixed together in the finally produced list. As a consequence, the user has to access firstly a large number of unwanted Web pages in order to come across with those that interest him. An effective solution to this Web information retrieval problem is appropriately clustering search results, which consists of grouping the results returned by a search engine into topic categories.

Such systems usually utilize text mining techniques in order to extract the topic structure that the results in the document set capture. Text mining refers to the discovery of previously unknown knowledge that can be found in text collections. In recent years, the text mining field has received great attention due to the abundance of textual data especially coming from Web. Document clustering has proven to be a challenging computer science problem having a large number of application domains. As problem it becomes even more interesting and demanding with the development of the World Wide Web [10].

Most of the text mining techniques are based on word and/or phrase analysis of the text. The statistical analysis of a term (word or phrase) frequency captures the importance of the term within a document. However, to achieve a more accurate analysis, the underlying mining technique should indicate terms that capture the semantics of the text from which the importance of a term in a sentence and in the document can be derived [12, 8]. Incorporating semantic features from the WordNet lexical database is one of many approaches that have been tried to improve the accuracy of text clustering techniques. In these works the proposed models analyze terms and their corresponding synonyms and/or hypernyms on the sentence and document levels. In particular if two documents contain different words and these words are semantically related, then the proposed model based on this similarity can measure the semantic-based similarity between the two documents. The similarity between documents relies on appropriate semantic-based similarity measures, that are applied to the matching concepts between documents.

The vast amount of documents available in the biomedical literature makes the manual handling, analysis and interpretation of textual information a daunting task. Automated methods that help users to search through this unstructured set of valuable archives are becoming increasingly important in scientific research. There is a large amount of knowledge recorded in the biomedical literature, as a significant number of articles published each year increases dramatically, following the advances in computational methods and high-throughput experimentation. The PubMed database, which is considered one of the most complete repositories of biomedical articles, contains more than 11 million abstracts and receives more than 70 million queries each month. A very popular indexing engine, concerning the life sciences and biomedical research is PubMed[1] created at 1996. PubMed is a free database accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics.

[1] http://www.ncbi.nlm.nih.gov/pubmed/

For extracting knowledge from online document repositories usually systems rely on predefined vocabularies, ontologies and metadata. Most of the methods in the scientific literature exploit MeSH [2]. MeSH is a a controlled vocabulary thesaurus defined by the National Library of Medicine, and includes a set of description terms organized in a hierarchical structure, where, as usual, more general concepts appear at the top, and more specific concepts appear at the bottom. MeSH's release in 2007 has totally 24.357 main headings, more than 164.000 supplementary concepts and over 90.000 entry terms.

In this work we propose a system that extends previous works by exploiting semantic text clustering techniques in order to present PubMed's query results clustered into topics. In order to provide more accurate clustering results, we employ various biomedical ontologies, like MeSH[3], in conjunction with WordNet[4] and common term clustering methods. The structure of the paper is as follows: in section 2 we briefly present previous works in this line of research, in section 3 we present the proposed method, in section 4 there are the implementation details and the experimental evaluation results, while in section 5 we present the conclusions and future directions of this work.

2 Related Work

Recently, there exists a significant activation in the line of research of biomedical document clustering, either by proposing novel clustering methods or by presenting new Web applications on top of document repositories like PubMed. Most of the research efforts utilize natural language processing techniques and show that information contained in terminologies and ontologies is very helpful as background knowledge to improve the performance of document clustering.

In [15], they initially investigated if the biomedical ontology MeSH improves the clustering quality for MEDLINE articles. In their work they performed a comparison study of various document clustering approaches such as hierarchical clustering methods, bisecting k-means, k-means, and suffix tree clustering in terms of efficiency, effectiveness, and scalability. According to their results, the employment of MeSH significantly enhances clustering quality on biomedical documents. Also in [16] they propose a method that represents a set of documents as bipartite graphs using domain knowledge in ontology. In this representation, the concepts of the documents are classified according to their relationships with documents that are reflected on the bipartite graph. Using the concept groups, documents are clustered based on the concepts contribution to each document. Their experimental results on MEDLINE articles showed that this approach can compete some well known approached like BiSecting k-means and CLUTO.

Following a similar approach, in [13] they propose PuReD-MCL (PubMed Related Documents-MCL), which is based on the graph clustering algorithm MCL and relevant resources from PubMed. Their method avoids using natural

[2] http://www.nlm.nih.gov/mesh/
[3] http://www.nlm.nih.gov/mesh/
[4] http://wordnet.princeton.edu/

language processing (NLP) techniques directly; instead, it takes advantage of existing resources, available from PubMed. PuReD-MCL then clusters documents efficiently using the MCL graph clustering algorithm, which is based on graph flow simulation. They applied their methodology to different datasets that were associated with Escherichia coli and yeast, and Drosophila respectively.

In [1] they deal with Biomedical word sense disambiguation with ontologies and metadata. For many clustering algorithms word sense disambiguation is a fundamental step for mapping documents to different topics. Usually, ontology term labels can be ambiguous and have multiple senses. Three approaches to word sense disambiguation, which use ontologies and metadata, are employed. The first method assumes that the ontology defines multiple senses of the term. It computes the shortest path of co-occurring terms in the document to one of these senses. The other method defines a log-odds ratio for co-occurring terms including co-occurrences inferred from the ontology structure. Finally, the last method trains a classifier on metadata. The authors have made experiments on ambiguous terms from the Gene Ontology and MeSH and found out that over all conditions their technique achieves 80% success rate on average.

In [3] they investigated the accuracy of different similarity approaches for clustering over a large number of biomedical documents while in [17] they claimed that current approaches of using MeSH thesaurus have two serious limitations: Firstly, important semantic information may be lost when generating MeSH concept vectors, and secondly, the content information of the original text has been discarded. They propose a method for measuring the semantic similarity between two documents over the MeSH thesaurus. Secondly, they propose the combination of both semantic and content similarities to generate the integrated similarity matrix between documents.

Similarly in [9] they report a novel approach to facilitate MeSH indexing, assigning MeSH terms to MEDLINE citations for archiving and retrieval purposes. For each document they retrieve k neighbour documents, they obtain a list of MeSH main headings from neighbours, and they rank the MeSH main headings using a learning-to-rank algorithm. Experimental results show that the approach makes fairly accurate MeSH predictions.

In this line of research in [4] they move beyond online resources, processing usually titles and abstracts, and tackle the problems that appear when dealing with full documents. In their work they present two interesting findings: i) dealing with full documents has increased demands in processing time and memory and ii) there is higher risk for false positives as in a full paper there is content (eg. references, tables etc.) not belonging to the core topic of the document. They propose a method that trades-off between using only parts of the full text identified as useful and using the abstract. Also they focus on the design of summarization strategies using MeSH terms. Very recently in [7] they proposed a FNeTD (Frequent Nearer Terms of the Domain) method for PubMed abstracts clustering. Their method consists of a two-step process: i) identifying frequent words or phrases in the abstracts through a frequent multi-word extraction algo-

rithm and ii) identifying nearer terms of the domain from the extracted frequent phrases using the nearest neighbours search.

The goal of our work is to design a system that will cluster and annotate biomedical documents originated from the Web having a reasonable trade-off between response time and cluster quality. The novel approach in the proposed system is the exploitation of MESH in conjunction with WordNet and common term clustering methods. Finally, we embed the proposed methodology in an online system that post-process PubMed online database in order to provide to users the retrieved results grouped in topics.

3 Proposed Method

The overall architecture of the proposed system is depicted in Figure 1 while the proposed method is modulated in the following steps:

1. Initially, we query an online Web document repository/PubMed, in order to process the returned Web documents.
2. In the following step, we produce the document representations and our system enriches the information by annotating texts using information from utilized vocabularies and ontologies. Each document is processed, removing stop words and stemming the remaining terms. Then each document is represented as a tf/idf vector [2], and some terms of the document are annotated and mapped on senses identified in from WordNet and MeSH.
3. The clustering of the documents takes place by employing k-means; it could be possible to test and evaluate our technique on different types of clustering algorithms but this will be attempted in future work.
4. In the final step, the system assigns labels to the extracted clusters in order to facilitate users to select the desired one. In the simplest possible way, the label is recovered from the clusters' feature vector and consists of a few unordered terms, based on their tf/idf measure. On the other hand, the system uses during clustering the identified MeSH terms and hence these identified terms can also serve as potential candidates for cluster labelling as shown in [6].

3.1 Retrieving PubMed Results

For retrieving search results from the PubMed we have used its standard API[5] for document retrieval. The results are retrieved in XML formatted file that later on is parsed and author names, titles, abstracts, conference/journal name, and publication date are identified.

3.2 Document Representation

After the results are retrieved in the initial step, each result item consists of five different items: title, author names, abstract content, conference/journal name

[5] http://www.ncbi.nlm.nih.gov/books/NBK25500/

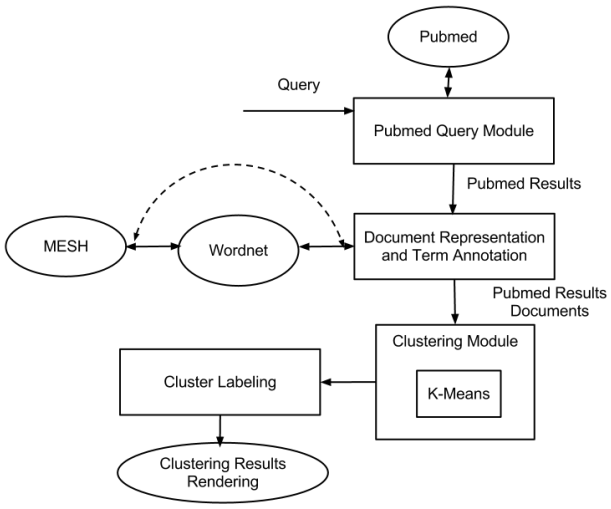

Fig. 1. Overall Architecture of the Proposed Method

and publication date. Each item is processed by removing, html tags, the stop words and then stemming each remaining term with the Porter stemmer. The aim of this phase is to prune from the input all characters and terms that can possibly affect the quality of group descriptions. For the time being in the clustering procedure the title and abstract are utilized however Conference/Journal name and publication dates could also be used, since they capture rich knowledge (eg. research works in the same topic usually published in same or similar journals etc.).

Consecutively, each search result is converted to a vector of terms using the tf/idf weighting scheme for each one of them[2]. As an alternative representation we use terms annotated to senses identified to WordNet as proposed in [12]. Consequently each one of the documents is mapped onto a set of WordNet senses, and from these set of senses we keep the top-k (usually k is between 5 to 10 senses) most expressive senses by using the Wu-Palmer sense similarity metric [14] on the WordNet graph. Furthermore, we utilize these extracted senses in order to identify MeSH terms. This is done by selecting for each sense the primary description and the appropriate terms and search for these set of terms the MeSH hierarchy. From the extracted MeSH terms again we select the more representative terms. For each pair of terms we count their distance in the MeSH hierarchy and then we select the top-k (usually k is 5-10 terms) with the minimum average distance to the other terms. Moreover, we produce a document representation vector by extracting MeSH terms by searching directly with the initial document terms. These multiple document representations are used in the following by the clustering algorithm. For the extracted MeSH term vectors we use again a similar vector space weighting scheme.

The set of search results along with their document vectors, extracted in this step, are given as input to the clustering algorithm, which is responsible for building the clusters and then assign proper labels to them.

3.3 Clustering and Cluster Labelling

For grouping the result document set we apply k-means as a common clustering technique [10], utilizing the different document vector representations of each document and adapting k-means at each case properly. For most of the cases we have used the cosine similarity distance upon vectors.

The four different clustering methods evaluated are:

1. **Tf/Idf_based:** In this case we use only the tf/idf document vectors and cosine similarity metric in the k-means method. We have used this setting as a point of reference to the performance of the other clustering method variations.
2. **WordNet_based:** In this case we use the vectors of extracted WordNet senses and cosine similarity metric in the k-means method. In the case of failure in extracting any WordNet sense vector we use its tf/idf vector when the k-means calculates its distance to the other documents. Furthermore, in a second step the initially produced clustering is used as bootstrapping for the k-means algorithm that now in this phase uses the MeSH document vectors that were extracted from the WordNet senses.
3. **MeSH_based:** In this third case we use directly the vectors of extracted MeSH terms, as they were extracted by the most significant terms (using tf/idf). In the case of failure in extracting any MeSH term vector we use its tf/idf vector when k-means calculates its distance to the other documents.
4. **Hybrid:** In this last case we modify k-means to use a complex distance based at the same time on the tf/idf and the MeSH vectors in the spirit of [5], which combines different levels of term and semantic representations. k-means calculates document distances using the cosine distance of term vectors and of the MeSH vectors normalized at 50% respectively. So term distance and semantic distance have the same weight in the final distance metric.

In each cluster that our system produces, we assign a label with various term/senses that define each topic. In the simpler approach, we assume that the label is recovered from the clusters feature vector and consists of a few unordered terms. The selection is based on the set of the most frequently occurring keywords in the clusters documents and we can measure this frequency by the tf/idf scheme which we have already mentioned in the document representation. In particular, we have used the intersection of the most frequently occurring keywords of all the documents that occur in a specific cluster. An alternative is to calculate the centroid and use the most expressive dimensions of this vector.

The second approach takes advantage of semantic MeSH terms. From these ones we use the most weighted ones by ranking them with the probability to

be selected as annotations in a new document. This probability is calculated by counting the number of documents where the term was already selected as a keyword divided by the total number of documents where the term appeared.

4 Implementation and Evaluation

We implemented our system as a Web server using Java EE and Web user interface using Java Server Pages (JSP). We have utilized the JWNL[6] library, in order to interoperate with WordNet 2.1 [7].

For the processing of the documents, we used the LingPipe[8] library. It is a Java tool kit for processing text using computational linguistics. Stemming, stopword methods, vector space representation, tf/idf weighting, query resolution schemes and clustering algorithms, were implemented by using this library.

In order to evaluate our method we used the TREC Genomics 2007 dataset[9]. The key features of this dataset is linkage and annotation. Linkage among resources allows the user to explore different types of knowledge across resources. Also it provides topic annotation of each one of the documents. Here we took from the whole datasets random samples with different number of documents ($n = 10, 20, 30, \cdots, 90$) and we clustered these documents with the evaluated methods. Finally we evaluated the produced clusters by using precision, recall and FMeasure scores.

As we can see in Figure 2, the proposed representation and clustering strategies achieve notable precision and recall for small and average number of processed documents. As the number of processed documents increases the performance of the methods seems to decrease. The Hybrid method seems in most of the cases to achieve better performance concerning the precision metric. On the other hand the WordNet_based method seems to achieve better recall in most of the cases. Observing the FMeasure, Hybrid and common Tf/Idf_based seems to perform better for large number of processed documents. For small number of documents Hybrid and WordNet_based seem to have better FMeasure scores. One of our findings is that postprocessing clustering with MeSH and/or WordNet features seems to produce almost better clustering results. If we assume that we have preprocessed feature vectors of documents the convergence time of the k-means is significant smaller.

5 Conclusion

Clustering online biomedical documents is a critical task for researchers. It allows better topic understanding and favours systematic exploration of search results and knowledge coverage. This paper describes various semantic representation of PubMed documents in order to be categorized into topics. Here in this work

[6] http://sourceforge.net/projects/jWordNet/
[7] http://wordnet.princeton.edu/
[8] http://alias-i.com/lingpipe/
[9] http://ir.ohsu.edu/genomics/

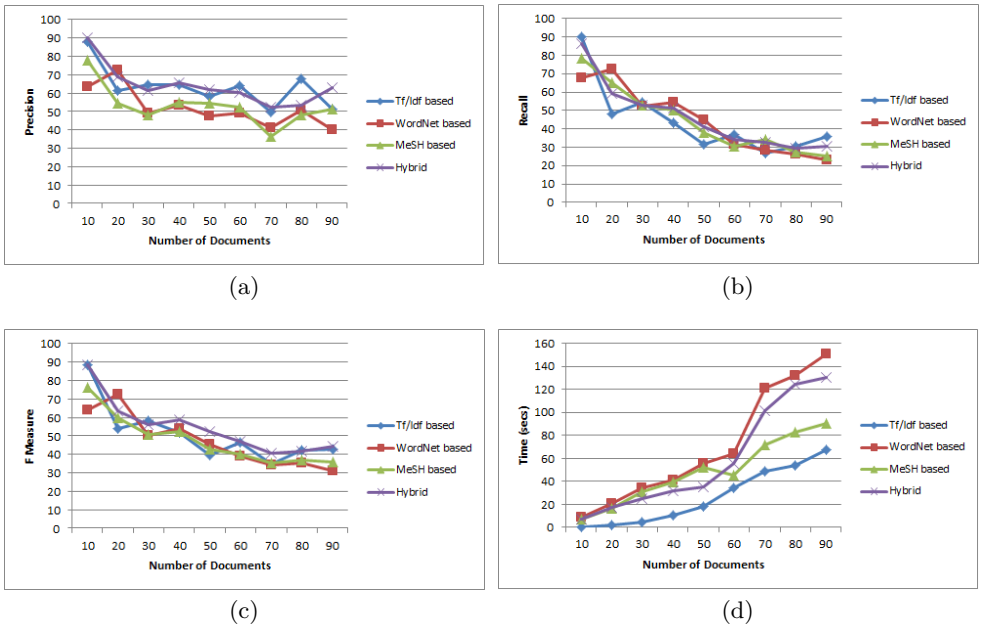

Fig. 2. (a) Precision , (b) Recall , (c) FMeasure and (d) Computation time for, different number of documents.

we covered various representation methods, and various clustering alternatives that show quite promising.

Directions for further investigation in this line of research would be the achievement of a better trade-off between response time and clustering of online document results. Small response times are critical for user adoption of an online Web document repository not sacrificing quality of the results. It would be interesting to explore techniques on how to distribute computational effort in order to effectively preprocess offline frequent patterns of queries. Another interesting topic of research would be how it would be possible to transfer a portion of the computation that is performed in the client side at the user's Web browser taking in mind the user's preferences and previous browsing history. We would like also to use hierarchical clustering algorithms in order to identify the suitable number of formed clusters in order the avoid the intrinsic weakness of k-mean, and then run the k-means clustering for result refinement. Furthermore, we intend to embed better domain knowledge in the clustering procedure and apply better disambiguation strategies tailored for the MeSH ontology, using techniques similar to [1].

References

1. D. Alexopoulou, B. Andreopoulos, H. Dietze1, A. Doms, F. Gandon, J. Hakenberg, K. Khelif, M. Schroeder and T. Wachter, Biomedical word sense disambiguation with ontologies and metadata: automation meets accuracy. BMC Bioinformatics 10:28, 2009.
2. R. A. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval, Addison Wesley, 1999 (second edition, 2011, http://mir2ed.org/)
3. KW. Boyack, D. Newman, RJ. Duhon, R. Klavans, M. Patek, J. Biberstine, B. Schijvenaars, A. Skupin, N. Ma, K. Borner, Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches. PLoS ONE 6(3), 2011
4. S. Bhattacharya , V. Ha-Thuc, P. Srinivasan, MeSH: a window into full text for document summarization. Bioinformatics. Jul 1, 27(13):120-8, 2011.
5. A. Caputo, P. Basile, and G. Semeraro. SENSE: semantic N-levels search engine at CLEF2008 ad hoc robust-WSD track. In Proc. of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access (CLEF), 2008.
6. D. Carmel, H. Roitman, and N. Zwerdling. Enhancing cluster labeling using wikipedia. In Proc. of the 32nd international ACM SIGIR conference on Research and development in information retrieval. pp.139-146, 2009.
7. M. Rajathei David and S. Samuel. Clustering of PubMed abstracts using nearer terms of the domain. Bioinformation. 8(1): 20-25, 2012
8. R. Hemayati, W. Meng, and C. Yu. Semantic-based grouping of search engine results using WordNet. In Proc of 8th international conference on Web-age information management conference on Advances in data and Web management, 2007.
9. M. Huang M, A. Nvol A, Z. Lu, Recommending MeSH terms for annotating biomedical articles. J Am Med Inform Assoc; Sep-Oct;18(5):660-7, 2011
10. J. Kogan. Introduction to Clustering Large and High Dimensional Data, Cambridge University Press, 2007.
11. R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In Proc. of the sixteenth ACM conference on Conference on information and knowledge management. pp.233-242, 2007.
12. S. Shehata. A WordNet-Based Semantic Model for Enhancing Text Clustering. In Proceedings of the 2009 IEEE International Conference on Data Mining Workshops (ICDMW '09). IEEE Computer Society. pp. 477-482, 2009.
13. T. Theodosiou, N. Darzentas, L. Angelis and C. A. Ouzounis, PuReD-MCL: a graph-based PubMed document clustering methodology Bioinformatics: 24(17): 1935-1941.July 1, 2008
14. Z. Wu and M. Palmer. Verb semantics and lexical selection. In Proc. of the 32nd Annual Meeting of the Assoc. for Computational Linguistics, pp.133-138. 1994.
15. I. Yoo and X. Hu. Biomedical Ontology MeSH Improves Document Clustering Qualify on MEDLINE Articles: A Comparison Study. In Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS), 2006.
16. l. Yoo and X. Hu. Clustering large collection of biomedical literature based on ontology-enriched bipartite graph representation and mutual refinement strategy. In Proc. of the 10th Pacific-Asia conf. on Adv. in Knowledge Disc. and Data Mining, 2006.
17. S. Zhu, Jia Zeng, H. Mamitsuka. Enhancing MEDLINE document clustering by incorporating MeSH semantic similarity. Bioinformatics, Vol. 25, No.15(1944-1951), 2009.