Summarizing topical contents from PubMed documents using a thematic analysis

Sun Kim, Lana Yeganova and W. John Wilbur
National Center for Biotechnology Information
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894, USA
{sun.kim,yeganova,wilbur}@ncbi.nlm.nih.gov

Abstract

Improving the search and browsing experience in PubMed® is a key component in helping users detect information of interest. In particular, when exploring a novel field, it is important to provide a comprehensive view for a specific subject. One solution for providing this panoramic picture is to find sub-topics from a set of documents. We propose a method that finds sub-topics that we refer to as themes and computes representative titles based on a set of documents in each theme. The method combines a thematic clustering algorithm and the Pool Adjacent Violators algorithm to induce significant themes. Then, for each theme, a title is computed using PubMed document titles and theme-dependent term scores. We tested our system on five disease sets from OMIM® and evaluated the results based on normalized point-wise mutual information and MeSH® terms. For both performance measures, the proposed approach outperformed LDA. The quality of theme titles were also evaluated by comparing them with manually created titles.

1 Introduction

PubMed[1], currently a collection of about 25 million bibliographic records, has grown exponentially in size. With the abundance and diversity of information in PubMed many queries retrieve thousands of documents making it difficult for users to browse the results and identify the information most relevant to their topic of interest. The query 'cystic fibrosis', for example, retrieves papers that discuss different aspects of the disease, including its clinical features, treatment options, diagnosis, etc. A possible solution to this problem is to automatically group the retrieved documents into meaningful thematic clusters or themes (these terms are used interchangeably). However, clustering alone does not solve the problem entirely, as a significant amount of human post-processing is required to infer the topic of the cluster.

There exists a vast collection of probabilistic clustering methods. One common problem among most of them is that different results are obtained depending on the cluster initialization, suggesting that some clusters are unstable or weak. However, there is no obvious way to effectively and efficiently evaluate the quality of clusters. In this paper, we combine EM-based thematic clustering (Kim and Wilbur, 2012) with the Pool Adjacent Violators (PAV) algorithm (Ayer et al., 1955; Wilbur et al., 2005). PAV is an isotonic regression algorithm which we use as a method for converting a score into a probability. Here, we show how PAV can be applied to evaluate the quality of clusters.

Another issue that motivated this research is that most existing algorithms produce clusters that are not self-descriptive. Presenting meaningful titles can significantly improve the user perception of clustering results. To that end, we utilize PubMed document titles and cluster-related term scores to automatically obtain a title for each theme. The method results in thematic clusters of documents with cluster titles.

Studies similar to our approach are ASI (Adaptive Subspace Iteration) (Li et al., 2004) and SKWIC (Simultaneous Keyword Identification and Clustering of text documents) (Frigui and Nasraoui, 2004). Both perform document clustering and cluster-dependent keyword identification simultaneously. SKWIC can only produce hard clustering, while ASI is computationally very expensive as it heavily depends on matrix operations. A study by Hammouda et al. (2005) sug-

[1] http://pubmed.gov

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 805–810,
Lisbon, Portugal, 17-21 September 2015.

gests automatic keyphrase extraction from a cluster of documents as a surrogate to providing a cluster title, but they treat document clustering and cluster-dependent keyword extraction as separate problems.

Topic modeling (Hofmann, 1999; Blei et al., 2003; Blei and Lafferty, 2005) is the most popular and an alternative approach that has a similar underlying goal of discovering hidden thematic structure of a document collection and organizing the collection according to the discovered topics. Topic models are based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words (Steyvers and Griffiths, 2007). However, topic modeling is not a document clustering scheme in nature. Although a list of keywords that represent a topic is available, the title of the cluster may not be evident.

2 Methods

We here describe the EM-based clustering algorithm, and show how PAV is incorporated with it to yield the PAV-EM thematic clustering technique. We further present a cluster summarization method to induce theme titles.

2.1 Theme definition

Let D be a document set and let T be the set of terms in D. Let R denote the relation between elements of T and D. tRd means $t \in d$. We define a theme as a subject that is described by non-empty sets $U \subseteq T$ and $V \subseteq D$, where all the elements of U have a high probability of occurring in all the element of V. An EM framework is used to extract subject terms for a theme (Wilbur, 2002). In addition to the observed data R, a theme is defined by the latent indicator variables z_d, $\{z_d\}_{d \in D}$. The parameters are

$$\Theta = U(\|U\| = n_U), \{p_t, q_t\}_{t \in U}, \{r_t\}_{t \in T}, \quad (1)$$

where n_U is the size of the set U. For any $t \in U$, p_t is the probability that for any $d \in V$, tRd. q_t is the probability that for any $d \in D - V$, tRd. For any $t \in T$, r_t is the probability that for any $d \in D$, tRd. Assuming all relations tRd are independent of each other, the goal is to obtain the highest probabilities

$$p(R, \{z_d\}|\Theta) = p(R|\{z_d\}, \Theta)p(\{z_d\}|\Theta). \quad (2)$$

E-step (expectation step) evaluates the expectation of the logarithm of Eqn. 2. M-step (maximization step) maximizes this expectation over the parameters Θ. For each term, t, we define a quantity α_t which is the difference between the contribution coming from t depending on whether $u_t = 1$ or $u_t = 0$. The maximization is completed by choosing the n_U largest α_t's and setting $u_t = 1$ for each of them and $u_t = 0$ for all others. Details of this theme extraction scheme can be found in Wilbur (2002).

Algorithm 1 PAV-EM algorithm

Let D be the dataset, where $d \in D$.
Give a value for the parameter q.

Set $X = \emptyset$.
for $i \leftarrow 1, n$ **do**
  Create q random clusters.
  Run the theme clustering algorithm.
  For each cluster C and d with pz_d^C,
    $X \leftarrow X \cup \{< pz_d^C, 1, 1 >\}$[2] if $d \in C$,
    $X \leftarrow X \cup \{< pz_d^C, 1, 0 >\}$ if $d \notin C$.
Obtain the PAV function, $PAV(pz_d^C)$, over X.

Set $S = \emptyset$, where S is the output cluster set.
repeat
  Create q random clusters for $\{d|d \notin \cup S\}$.
  Run the theme clustering algorithm.
  Select any cluster C, where
    $C' = \{d|d \in C, PAV(pz_d^C) > 0.9\}$
    satisfies $|C'| > 10$.
    $S \leftarrow S \cup \{C'\}$.
until no more changes in S.

2.2 PAV-EM thematic clustering

In thematic clustering, a document is assigned to a theme that has the highest probability to the document (Kim and Wilbur, 2012). Although this approach shows a reasonable performance for theme-based document clustering, the dynamic nature of random initialization and multiple subjects described in a document may create many weak themes. Moreover, there is no clear guideline to distinguish strong and weak themes. Thus, we here propose a method that extracts strong themes more effectively. In the EM-based theme extraction scheme, the log odds score pz_d^C indicates the extent to which a document d is coupled with a specific theme C. If a cluster in-

[2]The second and the third arguments in the bracket are the weight and the probability estimate of the data, respectively.

cludes a reasonable number of documents that have high pz_d^Cs, it indicates that the cluster represents a strong theme. Therefore, we can obtain strong themes by collecting these clusters.

Let the probability $p(score)$ be a monotonically non-decreasing function of $score$. The PAV algorithm (Ayer et al., 1955; Wilbur et al., 2005) is a regression method to derive from the data that monotonically non-decreasing estimate of $p(score)$ which assigns maximal likelihood to the data. For our approach, $score = pz_d^C$.

Algorithm 1 shows the theme clustering process using the PAV algorithm. For the given dateset D and the initial number of clusters q, theme clustering is performed n times, and an isotonic regression function is learned by applying the PAV algorithm. Note that q is an initial guess for the number of clusters and it is not guaranteed to remain the same in the output set. For our experiments, we set $q = 50$ and $n = 100$. After the PAV algorithm is applied, theme clustering is performed. At each iteration, we select any cluster in which there are more than 10 documents with PAV scores higher than 0.9. Unselected documents are re-used for clustering in the next iteration. This procedure is repeated until there are no more changes in the selected cluster set S.

2.3 Theme summarization

After obtaining themes (document clusters and their subject terms), we summarize each theme by choosing a text segment from PubMed document titles. A title should cover as many subject terms as possible, but also it should be well-formed, i.e. be descriptive enough and humanly understandable. To achieve this goal, we first extract all possible candidates from document titles as follows:

(i) Extract all possible candidates as n-grams, where $n = 1, ..., 20$. Noun phrases are treated as units and must be totally inside or outside a candidate.

(ii) Check POS tags for starting and ending words in a candidate. Starting with a conjunction, verb, preposition and symbol is not allowed. Ending with a conjunction, verb, preposition, symbol, determiner, adjective or certain pronouns is not allowed.

(iii) Discard any candidates that start or end with '-' or '.'. The candidates including certain characters such as '/', ';', ':' are also removed.

(iv) Check grammatical dependency relations. We discard candidates for which the head word of a preposition does not appear in the same candidate as the proposition. Also, we validate the case, 'between A and B', so that A and B are not separated.

Next, for each candidate, a score is calculated by

$$score(cand_i) = \log \frac{\prod_{t \in U}(tf_t \alpha_t)}{\prod_{t \notin U} tf_t}, \qquad (3)$$

where tf_t is the term frequency of the term t. However, an ideal title should have enough words to be descriptive, hence we subtract $(len(cand_i) - 5)^2$ from $score(cand_i)$, where $len(cand_i)$ is the number of words in $cand_i$, and choose the top score as a title.

3 Experimental Results

We applied our method to the five disease sets, "cystic fibrosis", "deafness", "DiGeorge syndrome", "autism" and "hypertrophic cardiomyopathy" from OMIM[3]. These sets consist of 3000, 3000, 956, 2917 and 1997 PubMed documents, respectively, and are available at `http://www.ncbi.nlm.nih.gov/CBBresearch/Wilbur/IRET/PAVEM`.

For evaluating PAV-EM and comparing with the topic modeling method, latent Dirichlet allocation (LDA) (Blei et al., 2003), both approaches were performed 10 times for each disease set and scores were averaged over all runs. Mallet[4] was used to run LDA. The same tokenization was applied to LDA and PAV-EM. The number of topics given for LDA was 50 and the recommended optimization parameter was used for producing LDA topics.

Table 1 presents average runtimes[5] for LDA and PAV-EM. LDA and PAV-EM spent 15.2 and 13.3 seconds on average for processing the smallest set, "DiGeorge syndrome". However, in larger sets, e.g. "autism", it took 46.9 and 31.3 seconds for LDA and PAV-EM, respectively. We also ran another implementation[6] of LDA, which was 30 times slower than Mallet. While PAV-EM and

[3] `http://www.ncbi.nlm.nih.gov/omim`

[4] `http://mallet.cs.umass.edu`

[5] Both methods were tested on a single linux server. The processing times reported do not include the preprocessing stages done by Mallet and our implementation.

[6] `http://www.cs.princeton.edu/~blei/lda-c`

Dataset	LDA	PAV-EM
Set 1	25.7	18.4
Set 2	36.5	24.7
Set 3	15.2	13.3
Set 4	46.9	31.3
Set 5	30.3	19.2

Table 1: Average runtimes for LDA and PAV-EM in seconds. Sets 1, 2, 3, 4 and 5 are "cystic fibrosis", "deafness", "DiGeorge syndrome", "autism" and "hypertrophic cardiomyopathy", respectively.

Method	Topic terms	
	Top 5	Top 10
LDA	2.8906	10.9760
PAV-EM	4.0322	14.6213

Table 2: NMPI scores for LDA and PAV-EM.

LDA can be implemented in parallel computation[7], this indicates that PAV-EM may be more efficient to obtain themes for a larger set of PubMed documents.

The PAV-EM algorithm automatically learns themes from unlabeled PubMed documents, hence the performance measures that are used in supervised learning cannot be applied to our setup. Recent studies have shown more interest in topic coherence measures (Chang et al., 2009; Newman et al., 2010; Mimno et al., 2011), which capture the semantic interpretability of topics based on subject terms. Table 2 shows the topic coherence scores measured by normalized point-wise mutual information (NPMI). For both top 5 and top 10 subject terms, PAV-EM achieves better NMPI scores than LDA. NPMI is known to be strongly correlated with human ratings (Aletras and Stevenson, 2013; Röder et al., 2015) and is defined by

$$\text{NPMI} = \sum_{i=2}^{N} \sum_{j=1}^{i-1} \frac{\log \frac{p(t_i,t_j)+\epsilon}{p(t_i)p(t_j)}}{-\log\left(p(t_i,t_j)+\epsilon\right)}, \qquad (4)$$

where $p(t_i, t_j)$ is the fraction of documents containing both terms t_i and t_j, and N indicates the number of top subject terms. $\epsilon = \frac{1}{D}$ is the smoothing factor, where D is the size of the dataset.

MeSH (Medical Subject Headings) is a controlled vocabulary for indexing and searching biomedical literature (Lowe and Barnett, 1994).

[7]A parallel implementation of LDA appears in Wang et al. (2009)

MeSH	Method	Prec.	Recall	F1
Top 1	LDA	0.4529	0.3827	0.4125
	PAV-EM	0.3842	0.5303	0.4427
Top 3	LDA	0.3935	0.3931	0.3925
	PAV-EM	0.3388	0.5239	0.4086

Table 3: Classification performance based on top significant MeSH terms appearing in themes.

MeSH terms assigned to an article are often used to indicate the topics of the article, thus these terms can be used to identify how well documents are grouped by topics. In each cluster, p-values of MeSH terms are calculated using the hypergeometric distribution (Kim and Wilbur, 2001), and the top N significant MeSH terms are used to calculate precision, recall and F1. Table 3 compares PAV-EM with LDA[8] for the MeSH term-based performance. In the table, PAV-EM provides higher recall and F1 for top 1 and top 3 MeSH terms. Higher recall has an advantage in our task because the theme summarization process uses a consensus among PubMed documents to reach a theme title.

The next experiment was performed to compare machine generated titles with manually labeled titles. Although human judgements are subjective, it is not uncommon to collect human judgements for evaluating topic modeling methods (Mei et al., 2007; Chang et al., 2009; Xie and Xing, 2013). To validate the performance of the theme summarization approach, we first chose 500 documents from each disease set, and produced themes and titles. For each topic, five strongest themes were chosen, and they were shown to three human annotators with extracted subject terms. Table 4 shows an example of the proposed approach and the manual annotation for the "hypertrophic cardiomyopathy" set. Among 25 themes, our approach correctly identified 21 theme titles. We assumed that a machine-generated title was correct if it included any of manually annotated titles.

4 Conclusion

This study was inspired by an EM-based thematic clustering approach. In this probabilistic framework, theme terms are iteratively selected and documents are assigned to a most likely theme. The number of themes is dynamically adjusted

[8]For LDA, each document was assigned to the highest scoring topic.

Proposed approach	Annotator 1	Annotator 2	Annotator 3
cardiac myosin binding protein c	myosin binding protein c	cardiac myosin binding protein c	cardiac myosin binding protein c
ptpn11 mutations in leopard syndrome	ptpn11 mutations in leopard syndrome	ptpn11 mutations in leopard syndrome	ptpn11 mutations in leopard syndrome
cytochrome c oxidase	cytochrome c oxidase	mitochondrial cytochrome-c-oxidase deficiency	mitochondrial cytochrome c oxidase deficiency
friedreich ataxia and diabetes mellitus	friedreich ataxia	friedreich ataxia	friedreich ataxia
hepatitis c virus infection	hepatitis c virus	role of hepatitis c virus in cardiomyopathies	hepatitis c virus infection

Table 4: Comparison of the titles generated from the proposed approach and manual annotation for the "hypertrophic cardiomyopathy" set.

by probabilistic evidence from documents. The PAV algorithm is utilized to measure the quality of themes. After themes are identified, subject term weights and PubMed document titles are used to form humanly understandable titles. The experimental results show that our approach provides a useful overview of a set of documents. In addition, the method may allow for a new way of browsing by semantically clustered documents as well as searching with context-based query suggestions.

Acknowledgments

The authors would like to thank Donald C. Comeau and Rezarta Islamaj Doğan for their contribution to the manual evaluation. This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

N. Aletras and M. Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proc. International Conference on Computational Semantics (IWCS 2013)*, pages 13–22, March.

M. Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman. 1955. An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26(4):641–647.

D. Blei and J. Lafferty. 2005. Correlated topic models. In *Proc. Advances in Neural Information Processing Systems (NIPS 2005)*, pages 147–154, December.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proc. Advances in Neural Information Processing Systems (NIPS 2009)*, pages 288–296, December.

H. Frigui and O. Nasraoui. 2004. *Simultaneous Clustering and Dynamic Keyword Weighting for Text Documents*. Springer, New York, USA.

K. M. Hammouda, D. N. Matute, and M. S. Kamel. 2005. CorePhrase: keyphrase extraction for document clustering. In *Proc. International Conference on Machine Learning and Data Mining*, pages 265–274, July.

T. Hofmann. 1999. Probabilistic latent semantic indexing. In *Proc. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57, August.

W. Kim and W. J. Wilbur. 2001. Corpus-based statistical screening for content-bearing terms. *Journal of the American Society for Information Science and Technology*, 52(3):247–259.

S. Kim and W. John Wilbur. 2012. Thematic clustering of text documents using an EM-based approach. *Journal of Biomedical Semantics*, 3(Suppl 3):S6.

T. Li, S. Ma, and M. Ogihara. 2004. Document clustering via adaptive subspace iteration. In *Proc. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 218–225, July.

H. J. Lowe and G. O. Barnett. 1994. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *The Journal of the American Medical Association*, 271(14):1103–1108.

Q. Mei, X. Shen, and C. Zhai. 2007. Automatic labeling of multinomial topic models. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007)*, pages 490–499, August.

D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. 2011. Optimizing semantic coherence in topic models. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 262–272, July.

D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. 2010. Automatic evaluation of topic coherence.

In *Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2010)*, pages 100–108, June.

M. Röder, A. Both, and A. Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proc. ACM International Conference on Web Search and Data Mining (WSDM 2015)*, pages 399–408, February.

M. Steyvers and T. Griffiths. 2007. *Probabilistic Topic Models*. Erlbaum, Hillsdale, NJ, USA.

Y. Wang, H. Bai, M. Stanton, W.-Y. Chen, and E. Y. Chang. 2009. PLDA: Parallel latent Dirichlet allocation for large-scale applications. In *Proc. International Conference on Algorithmic Aspects in Information and Management (AAIM 2009)*, pages 301–314, June.

W. John Wilbur, L. Yeganova, and W. Kim. 2005. The synergy between PAV and AdaBoost. *Machine Learning*, 61(1-3):71–103.

W. John Wilbur. 2002. A thematic analysis of the AIDS literature. In *Proc. Pacific Symposium on Biocomputing*, pages 386–397, January.

P. Xie and E. Xing. 2013. Integrating document clustering and topic modeling. In *Proc. Conference on Uncertainty in Artificial Intelligence (UAI 2013)*, pages 694–703, July.