Enhancing the Literature Review Using Author-Topic Profiling

Alisa Kongthon, Choochart Haruechaiyasak, and Santipong Thaiprayoon

Human Language Technology Laboratory (HLT),
National Electronics and Computer Technology Center (NECTEC)
Thailand Science Park, Klong Luang, Pathumthani 12120, Thailand
{alisa.kon,choochart.har,santipong.tha}@nectec.or.th

Abstract. In this paper, we utilize bibliographic data for identifying author-topic relations which can be used to enhance the traditional literature review. When writing a research paper, researchers often cite on the order of tens of references which do not provide the complete coverage of the research context especially when the targeted research is multidisciplinary. Author-topic profiling can help researchers discover a broader picture of their topic of interest including topical relationships and research community. We apply the Latent Dirichlet Allocation (LDA) to generate multinomial distributions over words and topics to discover author-topic relations from text collections. As an illustration, we apply the methodology to bibliographic abstracts related to Emerging Infectious Diseases (EIDs) research topic.

Keywords: Bibliographic data, text mining, Latent Dirichlet Allocation (LDA), author-topic profiling, literature review.

1 Introduction

The conventional literature review process usually starts by identifying a few state-of-the art papers. The information on authors and citations of these papers will be used extensively to identify other related literature. Researchers will then digest content of these articles on a one-by-one basis. Hence, such a process can often limit researchers' perspective to only particular pieces of literature.

With the current advancement in information technology, digital libraries can now be used to make documents more easily accessible. Bibliographic databases are becoming widely known as a starting point for the unconventional literature review process. These databases usually provide tools for researchers to search for their articles of interest. In addition to a typical basic or advanced search capability, modern bibliographic databases provide features to list and rank search results by fields such as author, affiliation, subject area, and publication year. Researchers can comprehend the data set by observing the list of a particular field. For instance, the list of leading subject areas represents the prominent subtopics within that research topic. However, subject areas are usually controlled and indexed by database providers. Using such subjects may not fully reflect the

G. Buchanan, M. Masoodian, S.J. Cunningham (Eds.): ICADL 2008, LNCS 5362, pp. 335–338, 2008.

real intention of the authors due to bias and error introduced by human editors. Hence, an approach to automatically extract topics from content of documents is needed.

There have been many studies on discovery latent topics from text collections [1]. Latent Semantic Analysis (LSA) uses singular value decomposition (SVD) to map high-dimensional term-by-document matrix to a lower dimensional representation called latent semantic space [2]. However SVD is actually designed for normally-distributed data. Such a distribution is inappropriate for count data which is what a term-by-document matrix consists of. As an alternative to standard LSA, Probabilistic Latent Semantic Analysis (pLSA) assumes each word in a document as a sample from a mixture model, where the mixture decompositions are multinomial random variables that can be viewed as representations of topics [3]. Hence each word is generated from a single topic, and different words in a document may be generated from different topics. However, the pLSA model encounters overfitting problem because the number of parameters grows linearly with the number of documents. Latent Dirichlet Allocation (LDA) is then introduced to correct such problem. LDA is a generative probabilistic model for a set of documents [4]. The basic idea behind this approach is that documents are represented as random mixtures over latent topics, where each topic is represented by a probability distribution over words.

Steyvers et. al. extended the LDA to include authorship information so that authors are linked to terms in documents via latent topics [5]. This model not only discovers what topics are expressed in a document, but also which authors are associated with each topic. Instead of associating each document with a distribution over topics, the author-topic model associates each author with a distribution over topics. However, one common problem associated with the topic model is how to effectively label the discovered topics. Typically, the topic is labeled numerically. Another common way to assign topic name is by appending the terms appear in that topic together. In this paper, we modify the author-topic model to capture the relationship between authors and topics from a set of bibliographic data. Each author is represented by a probability distribution over topics, and each topic is a probability distribution over words extracted from abstracts and controlled keywords prepared by the database provider. We utilize these two types of terms in our model so that we could have the most comprehensive information. For each derived topic, the controlled keyword is then used as a representative for all words within that topic. With such informative representation, researchers can better understand the concepts within their research of interest. Our approach promises to improve traditional literature review process by helping researchers depict the "forest" (broader patterns of research activity) before looking at the "trees" (important prior art).

2 The Author-Topic Profiling

In our author-topic profiling model, the input data consists of a set of m bibliographic documents denoted by $\mathcal{D} = \{D_0, \ldots, D_{m-1}\}$. Given a document

collection, the author-topic identification problem becomes the model fitting that finds the best estimate of the topic-word distributions and the author-topic distributions. Gibbs sampling is used to solve this model fitting problem. As a result, the LDA algorithm generates a set of n topics denoted by $\mathcal{T} = \{T_0, \ldots, T_{n-1}\}$. Each topic is a probability distribution over p words denoted by $T_i = [w_0^i, \ldots, w_{p-1}^i]$, where w_j^i is an estimated probability value of word j assigned to topic i. Based on this model, each author can be represented as a probability distribution over the topic set $\mathcal{T}$, i.e., $A_i = [t_0^i, \ldots, t_{n-1}^i]$, where t_j^i is an estimated probability value of topic j assigned to author i. For each generated topic, we also calculate the binomial Z-score, which measures the degree of independence of the term from the topic. The higher Z-score means that the term is more dependent on the topic. Hence we select a controlled keyword with the highest Z-score as the representative for each topic.

3 A Case Study of Emerging Infectious Diseases (EIDs)

We illustrate the application of our proposed method through a case study of Emerging Infectious Diseases (EIDs). This case study aims to explore the possibility of using converging technologies to combat EIDs [1]. The literature review is conducted by searching for related publications from *Compendex* database. To come up with an appropriate search terminology, we experimented with various Boolean search operators to cover "emerging infectious diseases."[2] The result data set contains 5,046 records. Obviously with the traditional literature review process, one cannot digest content by reading all of these papers. We apply the proposed author-topic profiling to enhance the conventional literature review process. Figure 1 shows six different derived topics (out of 50 topics). Each table in Figure 1 presents the top-10 words that are most likely to be generated when that topic is created, and the top-5 authors who are most likely to write a word if it has come from that topic. The topic name is associated with a controlled keyword with the highest Z-score.

Figure 1 shows quite representative results. Topics related to different research areas such as ecosystem, algorithms, biosensors, among others, can be derived from our data set. The words associated with each topic are also quite precise in a semantic sense of a particular area of research. The topic name also best describes words within that topic. Such analytical results can help researchers depict related topics to EIDs such as viruses and immunology. Moreover, previously unknown relevant topics such as ecosystem and biosensors can be discovered as well. Also if researchers need to focus on a biosensors topic, they can start searching for articles written by S.S. Iqbal or other leading authors.

[1] EID:Roadmapping Converging Technologies for Combat Emerging Infectious Diseases, http://www.apecforesight.org

[2] Search terminology - [(infectious disease) OR (infectious diseases) OR pandemic OR epidemic OR outbreak OR outbreaks OR flu OR influenza] NOT [(computer viruses) OR (computer worms) OR (network protocols)].

Topic: Viruses

Extracted terms		Top-5 Authors
virus	0.048	Hirschman L.
influenza	0.041	Jin M.
transmission	0.014	Subba V.
pandemic	0.011	Spiro D. J.
human	0.010	St. George K.
avian	0.010	
flu	0.009	
sars	0.007	
h5n1	0.007	
disease	0.006	

Topic: Ecosystem

Extracted terms		Top-5 Authors
species	0.015	Martikainen P.
forests	0.010	Siitonen J.
population	0.009	Punttila P.
outbreaks	0.008	Corbett D.
beetles	0.005	Erlandson J.M.
bark	0.004	
host	0.004	
insect	0.004	
climate	0.004	
moth	0.003	

Topic: Healthcare

Extracted terms		Top-5 Authors
health	0.020	Nichter L.S.
risk	0.010	Cox M.J.
public	0.008	Chou D.
disease	0.007	Liu B.
management	0.007	Weng L.
epidemic	0.006	
system	0.005	
outbreak	0.005	
cancer	0.004	
prevention	0.004	

Topic: Immunology

Extracted terms		Top-5 Authors
hiv-1	0.016	Ahuja S.K.
cells	0.015	Begum K.
infection	0.012	Jimenez F.
hiv	0.011	Telles V.
immune	0.010	Stahl-Hennig C.
virus	0.010	
cd8	0.008	
aids	0.007	
t-cell	0.006	
immunodeficiency	0.006	

Topic: Algorithms

Extracted terms		Top-5 Authors
data	0.017	Edmonds C.B.
system	0.012	Basham D.
information	0.011	Cronin A.
detection	0.007	Jagels K.
disease	0.006	Simmonds M.
algorithm	0.006	
network	0.006	
genome	0.005	
plague	0.005	
outbreaks	0.005	

Topic: Biosensors

Extracted terms		Top-5 Authors
detection	0.030	Iqbal S.S.
pcr	0.014	Bruno J.G.
dna	0.012	Batt C.A.
rapid	0.010	Mayo M.W.
system	0.009	Hsieh T.-M.
assay	0.009	
molecular	0.006	
real-time	0.005	
infectious	0.005	
sensitive	0.005	

Fig. 1. Sample topics related to EIDs. Each topic is illustrated with the top-10 words and the top-5 authors.

4 Conclusions

We proposed an approach called author-topic profiling to augment, not to replace, the conventional literature review. This proposed method is based on a probabilistic topic model which can automatically extract information about authors and topics from a large set of documents. In addition, our model can effectively label the discovered topics by utilizing controlled keywords. From the illustrative case study, our model was able to extract "hidden" information from our sample data set.

References

1. Steyvers, M., Griffiths, T.: Probabilistic topic models. In: Landauer, T., McNamara, D., Dennis, S., Kintsch, W. (eds.) Latent Semantic Analysis: A Road to Meaning. Lawrence Erlbaum, Mahwah (2006)
2. Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society of Information Science 41(6), 391–407 (1990)
3. Hofmann, T.: Probabilistic latent semantic indexing. In: Proc. of the 22nd annual international ACM SIGIR conference, pp. 50–57 (1999)
4. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet Allocation. Journal of Machine Learning Research, 993–1022 (2003)
5. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.: Probabilistic Author-Topic Models for Information Discovery. In: Proc. of the 10th ACM SIGKDD international conference on Knowledge Discovery and Data mining, pp. 306–315 (2004)