Phrase Based Topic Modeling for Semantic Information Processing in Biomedicine

Zhiguo Yu
Division of Biomedical Informatics
Department of Biostatistics
University of Kentucky
Lexington, Kentucky 40506
zhiguo_yu@uky.edu

Todd R Johnson
Division of Biomedical Informatics
Department of Biostatistics
University of Kentucky
Lexington, Kentucky 40506
todd.r.johnson@uky.edu

Ramakanth Kavuluru*
Division of Biomedical Informatics
Depts. of Biostatistics & Computer Science
University of Kentucky
Lexington, Kentucky 40506
ramakanth.kavuluru@uky.edu

***Abstract*—Given that unstructured data is increasing exponentially everyday, extracting and understanding the information, themes, and relationships from large collections of documents is increasingly important to researchers in many disciplines including biomedicine. Latent Dirichlet Allocation (LDA) is an unsupervised topic modeling technique based on the "bag-of-words" assumption that has been applied extensively to unveil hidden semantic themes within large sets of textual documents. Recently, it was extended using the "bag-of-n-grams" paradigm to account for word order. In this paper, we present an alternative phrase based LDA model to move from a bag of words or n-grams paradigm to a "bag-of-key-phrases" setting by applying a key phrase extraction technique, the C-value method, to further explore latent themes. We evaluate our approach by using a phrase intrusion user study and demonstrate that our model can help LDA generate better and more interpretable topics than those generated using the bag-of-n-grams approach. Given topic models essentially are statistical tools, an important problem in topic modeling is that of visualizing and interacting with the models to understand and extract new information from a collection. To evaluate our phrase based modeling approach in this context, we incorporate it in an open source interactive topic browser. Qualitative evaluations of this browser with biomedical experts demonstrate that our approach can aid biomedical researchers gain better and faster understanding of their document collections.**

I. Introduction

Knowledge discovery is a fundamental and important activity in biomedical research. Extracting and understanding the information, themes and relationships from large collections of documents are important tasks for biomedical researchers. Hence, an efficient and convenient way to discover knowledge from large sets of documents is needed for biomedical researchers, especially for those who are not familiar with the computer science or informatics techniques. Latent Dirichlet Allocation (LDA) [1], is a popular topic modeling method developed to automatically extract a set of semantic themes from large collections of documents. LDA is an unsupervised machine learning approach and can be viewed as a three-level hierarchical Bayesian model. It has already been applied in the context of biomedical research, for example, in the psychology domain for predicting behavior codes arising from couple therapy transcripts [2] and for risk stratification in ICU patients [3] using nursing text from the first 24-hours of patients' ICU stays. Thus further study of topic models is important in the context of understanding large unstructured datasets that arise in biomedical and clinical domains.

Unlike clustering approaches where documents are grouped into mutually exclusive clusters based on document based features, topic models represent each document as a mixture of different topics and each topic as a distribution of unique words. Finally the topics are represented as bags of words where only top m words (for some m) are shown for each topic. Since the words within each topic are ranked according to the conditional probabilities $P(w|t)$ learned when training the model where w is a word and t is a topic, the top few words of each topic provide insights into the subject of the topic. These top words are usually displayed to the user to give her or him a sense of what the topic is about. Thus, this original LDA model was developed based on the popular assumption of "bag-of-words", in which the word order is ignored. In many data mining applications, results of LDA were found be contain ambiguous lists of words as representatives of the topics because of the inherent polysemy and homonymy of words. Thus this original model was found to be difficult for researchers to comprehend the topics based on these top ranked word sets.

In general, single words convey less information than phrases. Some verbs or prepositions are even meaningless without related words. For example, the meaning of 'magnetic resonance imaging' cannot be completely determined from any one of these three words in isolation, 'magnetic', 'resonance' or 'imaging'. Thus the bag-of-words assumption cannot meet the needs of extracting salient themes from large sets of documents and in 2006, Wallach developed a bigram topic model [4] based on the original LDA (or just LDA), in which she incorporates bigram statistics into the latent topic variables to add the dependencies between consecutive words. In 2007, Wang et al. presented another topic model, called the *topical n-gram model* [5], based on Wallach's bigram model, that can form longer n-grams. Although the

*Corresponding author

topical n-gram model approach enriches the generated topics by longer sequences of words, the topic generation process is still based on individual words with the word context providing evidence to form a longer n-gram. We call this approach the "bag-of-n-grams" method.

In this paper, we propose a new LDA based model called the *Phrase LDA* where the topics are generated based on 'important' noun phrases instead of words or n-grams, thus our approach can be termed as using the "bag-of-key-phrases" approach. We use the C-value method for extracting the key phrases and build the LDA model based on the key phrases that have a C-value score (more on this later) that is above a threshold. We conduct two different evaluations of topic models extracted using our phrase LDA approach. The first, evaluates whether the "bag-of-key-phrases" approach is better at identifying semantic themes than the standard "bag-of-n-grams" approach. A user study with 11 participants using the "word intrusion" test [6] for topic model evaluation demonstrates that the Phrase LDA approach provides 7% improvement over the topical n-gram model. 8 out of 11 participants also answered that it was easier to do the evaluation for the Phrase LDA models. As topic models are high-level tools to summarize document sets, the outputs of topic models are not easy to understand by users unfamiliar with these models and the associated numerical distributions [7]. So, an efficient, effective, and convenient way is needed to interact and visualize the topics, documents, and corpus. In 2012, Chaney and Blei developed a visualization tool, termed a *topic browser* [8], to summarize a document collection and reveal inter-topic and document-topic relations. To evaluate our approach in the context of this visual processing task, as the second evaluation experiment, we incorporated the phrase LDA models into the topic browser implementation and conducted qualitative studies with biomedical researchers to explore different sets of documents and discover hidden semantic themes and connections. Initial results suggest that the phrase LDA approach is also very effective in visual exploration and semantic processing of document collections.

II. Background

In this section, we provide brief background on the original LDA, topical n-gram model, and the C-value method for key phrase extraction.

A. LDA and Topical n-gram Model

In the LDA [1] model, a document is represented as a mixture of latent topics and each topic is represented as a distribution of unique words. In the *generative modeling* perspective, LDA represents a large corpus of documents at three levels: the corpus level, the document level, and the word level as follows: 1) At the corpus level, LDA generates a topic-words distribution ϕ_z for each topic z from the topic-words Dirichlet prior β; 2) At the document level, LDA generates a document-topics distribution θ_d for each document d from the document-topics Dirichlet prior α; 3) At the word level, LDA generates the topic assignment z_n from the document-topics distribution θ_d first and then generates a word assignment w_n from the topic-words distribution ϕ_{z_n} for each word w_n in document d.

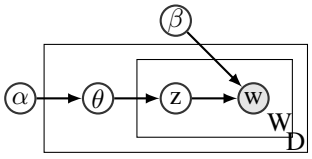

Figure 1: Graphic Model of LDA

Figure 1 is a graphical model [9] representation of LDA. The α and β are distributions as explained in the list above at the corpus level. D and W plates in the figure consist of distributions at the document level and word level respectively. This original LDA approach is based on the bag-of-words approach, where the words w are conditionally independent given their assigned topic z. However, as discussed in Section I, the word grouping topics are not often informative leading to the development of the topical n-gram model (TNG) [5], where two more dependencies are introduced at the word level. The first is the dependency between two consecutive words, the other being the dependency on the bigram status, which determines whether a bigram needs to be formed for the same consecutive word tokens depending on their nearby context. This model can also be expressed at three levels:

1) At the corpus level: (a) LDA generates a topic-words distribution ϕ_z for each topic z from the topic-words Dirichlet prior β; (b) LDA generates the bigram status Bernoulli distribution $\psi_{z,w}$ for each topic z and each word w from the Beta prior γ; (c) LDA generates the bigram distribution $\sigma_{z,w}$ for each topic z and each word w from the Dirichlet prior δ;
2) At the document level, LDA generates a document-topics distribution θ_d for each document d in the corpus from the document-topics Dirichlet prior α;
3) At the word level: (a) LDA generates a topic assignment z_n from the document-topics multinomial distribution θ_d; (b) LDA generates a bigram status x_n for each word w_n in document d from the Bernoulli distribution $\psi_{z_{n-1},w_{n-1}}$; (c) If the bigram status $x_n = 1$, LDA generates the word assignment w_n from the bigram status Bernoulli distribution $\sigma_{z_n,w_{n-1}}$, else LDA generates the word assignment w_n from the topic-words distribution ϕ_{z_n}.

Figure 2 is a graphical model representation of the TNG model, where D is the document level, T is the topic level, and W is the token level. Comparing with Figure 1,

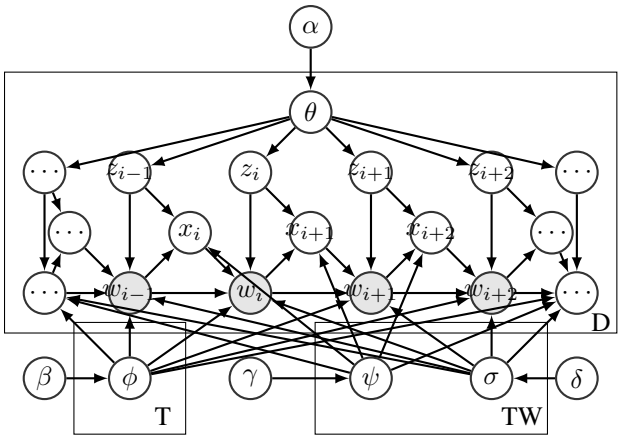

Figure 2: Graphic model of the topical n-gram Model

the bigram status Bernoulli distribution ψ and the bigram distribution of words σ are new in Figure 2. In the topical n-gram model, the last term of the n-gram is the word considered when generating the topic. That is, even though the topical n-gram model approach enriches the generated topics by longer sequences of words, the topic generation process is still based on individual words with the word context providing evidence to form a longer n-gram. As mentioned in Section I, constituent terms cannot capture the right meaning of the whole phrase. Besides, based on this approach, there is no way to remove those highly frequent n-grams that may not be important (e.g., "tend to show").

B. C-Value Method

Extractive text summarization is an approach where short summaries of a collection of documents are generated by selecting a few sentences or phrases from those documents that represent the gist of the collection in some way. The C-value [10] method is an extractive text summarization method that extracts key phrases that capture a summary of a collection of documents. It uses both linguistic information [11], [12] and the statistical information [13], [14] to identify key phrases. First the following three noun phrase regular expression filters are used to extract the candidate phrases.

1) `Noun* Noun`.
2) `(Adj|Noun)+ Noun`, and
3) `((Adj|Noun)+|((Adj|Noun)*(NounPrep)?)(Adj|Noun)*)Noun`

Here `Adj` stands for adjective and `NounPrep` stands for a noun followed by a preposition. Next for each candidate phrase, the C-value is computed based on its frequency and the frequencies of longer phrases that contain it in the given set of documents. The C-value formula can be written as

$$C(p) = \begin{cases} \log_2(len(p)) \cdot f(p) & \text{if } p \text{ is not nested} \\ \log_2(len(p)) \cdot \left(f(p) - \frac{1}{|T_p|}\sum_{q \in T_p} f(q)\right) & \text{if } p \text{ is nested} \end{cases}$$

where $C(p)$ is the C-value of phrase p, $len(p)$ is number of words in p, and T_p is the set of the longer noun phrases that contain p, and $f(p)$ is the frequency of p in the corpus. If p is not nested, it implies that it does not appear in longer phrases. When it is nested, we discount its C-value based on the number of its occurrences in longer phrases (the $\sum_{q \in T_p} f(q)$ part) and dampen this discount based on the number of unique longer phrases that contain it (the $\frac{1}{|T_p|}$ part). With this measure, the larger the C-value, the more important is the phrase relative to those with lower C-values.

III. Methods

In this section we describe the construction of our Phrase LDA model and the methodologies for the two different evaluations we conducted.

Phrase LDA Method with C-values

We use the traditional LDA method by reducing the contents of documents to noun phrases for which the C-value computed over the set of documents to be modeled is greater than 2. This threshold of 2 for the C-value is determined based on our experimental analysis. Note that phrases that occur multiples times in the same document are used as many times as they appear, that is, duplicates are retained. Next we described the two different evaluation strategies employed in our paper.

A. Evaluation Using "Phrase/n-Gram Intrusion"

Our first evaluation is a user study with 11 participants using the word (phrase) intrusion test [6] where we compare the Phrase LDA model with the topical n-gram model.

We obtained a corpus of 26,533 citations using the PubMed* query

```
public health[majr] AND united states[mh]
        AND "last 4 year"[dp]
```

to fetch the titles and abstracts of the matching articles from PubMed. This query fetches citations corresponding to articles that discuss public health as a major topic with US as a geographic location in the last four years. We chose this particular query as our participants are from the college of public health. We applied the time period constraint to limit the number of abstracts to a reasonable size. We treated each title and its corresponding abstract as a document. We first computed the C-value of the phrases from the corpus and retained only those phrases for each citation with the C-value larger than two. Based on this threshold, we chose 51,627 unique phrases out of the total 365,156 phrases. 84% of the chosen phrases have frequencies less than ten.

*PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) is a Web information system provided by the National Library of Medicine (NLM) to search more than 22 million biomedical scholarly papers/books and is the most popular scientific literature search system among biomedical researchers

Next the text for each citation is replaced with the C-value > 2 noun phrases (including duplicates) that appear in the citation (abstract and title) text. For the regular LDA , we chose 25,798 unique words out of 67,775 words based on the minimum frequency threshold of two. 87% of words thus selected have frequencies less than 100.

We built topic models for our corpus using the general LDA, topical n-gram model, and our phrase LDA model. For the original LDA and Phrase LDA we used the implementation LDA-C[†] with 50 topics. We used MALLET [15] toolbox for the implementation of topical n-gram model. We set 50 topics and 1,000 iterations for the total 26,533 documents. The comparison of sample topics generated by these three models is shown in Table I. As can be seen the n-gram models might not contain noun phrases and might just have frequent n-grams that are not necessarily informative or meaningful (e.g., "based medicine" of the first topic in Table I(b)).

Chang et al. [6] introduced an important intrinsic evaluation method called *word intrusion* for topic models that is independent of the application context. It involves human subjects who evaluate the intrinsic coherence of the topics generated. We extended this to "phrase/n-gram Intrusion" test to compare the quality of the topics generated by our model against the topical n-gram models.

We randomly chose 25 topics out of 50 topics generated by topical n-gram model and our phrase LDA model. For each selected topic, we then chose the top three phrases and randomly select one phrase out of the bottom five phrases as the *intruder phrase*. We mixed these four phrases together and present it to the user as a multiple choice question where the objective is to identify the intruder phrase. If the topics are semantically cohesive and meaningful, users should be able to easily identify the intruder phrase. If the topics are incohesive, users might find it difficult to identify the intruder phrase and may resort to guessing. We built an anonymous questionnaire[‡] based on this phrase intrusion approach through the online survey software program Qualtrics[§]. This questionnaire contains fifty questions and each question comes from one of the randomly selected 25 public health topics described earlier using the topical n-gram model and our phrase LDA model. To make sure that each questionnaire is endowed with a minimal level of user concentration and reading comprehension, we added several simple questions (e.g., a question with choices $\{Father,\ Mother,\ Brother,\ Cancer\}$) to the questionnaire. If a user got any one of these simple questions wrong, we exclude this response from our analysis.

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

[‡] https://uky.qualtrics.com/SE/?SID=SV_3qlfcTrBN6aNZt3

[§] http://www.qualtrics.com/

In [6], the model precision is defined as

$$MP_k^m = \frac{1}{S}\sum_s 1(i_{k,s}^m = w_k^m)$$

where MP_k^m is the precision of model m for topic k, $i_{k,s}^m$ is the intruder phrase selected by user s for the topic k and model m, w_k^m is the actual intruder phrase selected by us for model m for topic k, and S is the total number of the subjects. The function $1(<\ condition\ >)$ is a Boolean function that results in a 1 if $condition$ evaluates to TRUE and returns a 0 if $condition$ evaluates to FALSE. To compute the overall performance of a model, we calculate the average model precision as follows

$$AMP^m = \frac{1}{T}\sum_{k\in T} MP_k^m$$

where T is the total number of selected topics in model m.

B. Evaluation Using Topic Browser

Visualization tools are difficult to evaluate because they are primarily tools for supporting a creative process for developing insight and generating and then exploring hypotheses using open-ended discovery [16]. Thus a key measure of success of visualizations is whether they help biomedical researchers develop interesting new hypotheses and ask new questions, not to simply answer pre-existing questions. Given topic models are developed to automatically summarize, organize, and understand large document sets, evaluation should also focus on whether phrase LDA based tools help biomedical researchers fulfill these goals.

To evaluate our modified topic browser, we use the qualitative evaluation methodology developed by Saraiya et al. [17] for evaluating how well microarray visualization tools enabled biological insight. Three biomedical subject experts used our phrase LDA incorporated topic browser based on their respective areas of research interest. The browser was instantiated with the models built using a specific query tailored to their interests and supplied by each of them separately. The first subject's query was "*prescription drug abuse* that resulted in a total of 2649 records (titles and abstracts) when searched in PubMed. The second subject is a Biomechanics expert and supplied the related PubMed Boolean query "*(back OR trunk OR spine OR lumbar OR vertebral column) AND (biomechanic* OR mechanic* OR load* OR stability)*" to be searched in the title and abstract of articles resulting in 21,041 records. The last subject, an expert on the disease Myositis, provided the PubMed query "*myositis AND ("skeletal muscle" OR macrophages OR inflammation OR regeneration) AND (Dermatomyositis OR "idiopathic inflammatory myopathy" OR polymyositis OR "inclusion body myositis" OR "cancer associated myositis")*" resulting in 1549 records.

All subjects were given 15 minutes of instruction and demonstration on how to use the browser along with a list

Topic 1	Topic 2	Topic 3
quality	clinical	cost
medical	trials	costs
electronic	trial	per
data	studies	life
records	randomized	economic

(a) LDA

Topic 1	Topic 2	Topic 3
health care	clinical modification	birth defects
public policy	diagnostic mammography	birth defect
health policy	distraction index	adverse pregnancy
public health policy	screening parameters	congenital heart defects
based medicine	negative rate	multiple births

(b) Topical n-gram Model

Topic 1	Topic 2	Topic 3
newborn screening	health disparities	air pollution
clinical research	high rates	data sets
clinical studies	health care professionals	air pollutants
association study	social determinants	measurement error
health programs	health interventions	regression models

(c) Phrase-Based LDA

Table I: Three sample topics generated by LDA, TNG, and phrase LDA. Top 6 words/phrases are listed per topic.

of the kinds of questions that could be explored with it. This was designed to replicate the natural process whereby researchers learn to use new tools recommended by other colleagues. Subjects were then requested to answer questions that are pertinent to identify high level comprehension such as "*Can you get a brief idea of what these documents are talking about?*" and "*Do these topics make sense to you?*" and so on. After this, they were instructed to continue to use the browser to explore the models until they felt that they would not gain further insight.

IV. Results

In this section we describe the results of both evaluations methods outlined in Section III-A and III-B.

A. Phrase/n-Gram Intrusion Evaluation Results

All the 11 users completed the intruder phrase recognition questionnaire (generated as described in Section III-A) using an online interface. All these users are graduate students who work on public health topics at the University of Kentucky. Five of them are in the age group 22-25, four in age group 26-29, and the remaining three are at least 30 years old. The average time that the subjects spent on this questionnaire is about 20 minutes. The model precision for topical n-gram model for the 25 topics was 48% and for the phrase LDA was 55%. So our phrase LDA has achieved a 7% improvement over the topical n-gram model in this intrinsic evaluation. Furthermore, 8 of the 11 users chose the topics generated by our phrase LDA model as easier to understand than those generated by the topical n-gram model. The results show that our adaptation of original word based LDA to key phrase based LDA has resulted in better topic cohesion and also the overall comprehension of the topics generated compared to the topical n-gram model. Using the C-value method, we extract important phrases by filtering out the phrases with low C-values to improve the overall comprehension while still maintaining cohesion.

B. Topic Browser Evaluation Results

Three biomedical researchers completed the evaluation of the topic browsers we built for them based on their interests as discussed in Section III-B. They spent an average of 35 minutes for this task.

The first expert found a few 'target' documents quickly using the topic browser with 60 topics generated for the "prescription drug abuse" collection[¶]. The target documents are those that the expert found as relevant in his prior searches using PubMed. After quickly reviewing other documents, topics, and their connections, the expert confirmed his suspicion that little is published in the area of interest – effect of drug screening programs on drug abuse.

The second research explored the 70 topics for the theme "back pain and biomechanics". The subject was able to quickly determine the field of research behind each topic generated. Here are several examples as provided by the him (indicated by top few phrases of the topic): "low back pain, risk factors, work load" suggests studies conducted in the area of occupational biomechanics, ergonomics, and epidemiology that have an emphasis on prevention; "back pain, low back pain, chronic back pain, pain patient, mechanical low back pain" suggests studies conducted by people in the area of health science like physical therapy, with an emphasis on rehabilitation; "muscle activity, muscle forces, lumbar

[¶]We uploaded the full topic browser for the drug abuse topic here: http://sweb.uky.edu/~zyu224/drug-abuse60/browse/topic-presence.html

spine, shear forces, trunk muscle" suggests an engineering approach to trunk biomechanics. This subject found 59 of 70 topics meaningful and was also able to identify synonymous phrases across different subfields of biomechanics – an important task that can help a translational researcher bridge two related disciplines.

The third expert explored the 40 topics generated for the "myositis" subject. Given that this researcher had already done considerable research in this area, she was quite familiar with the documents and concluded that the browser captured most aspects of the "myositis" area. All three researchers made the following observations after using the phrase LDA based browser: (1) **Advantages**: The researchers found the topic browser interesting to explore. They also noted that it helped them save time reviewing the documents that they were interested in owing to the topic based presentation. Besides, it has been helpful in disambiguating similar topics being discussed by researchers in other related subfields. (2) **Improvements suggested**: The researchers felt that the browser needs a way to show the documents based on a combination of topics that they are interested in. They are also more interested in starting with phrases instead of reviewing the topics one by one. Hence the browser needs a way of helping users navigate from the phrases to the interesting topics. However, this is not necessarily a limitation of the phrase LDA approach but more of the browser implementation.

V. Conclusion

In this paper we presented a phrase-based LDA model and its application in semantic information processing in biomedicine. We conducted intrinsic model evaluation through a phrase intrusion detection user study, which resulted in 7% improvement in model precision. We also conducted a qualitative expert user study to evaluate the approach when used in a topic browser [8] to visually explore the topics. The result shows that this system helps users save search time and enables them to review and understand the documents by showing the semantic structure underlying these documents. We conclude with two future research directions: (1) Topic models built on increasing levels of abstraction might provide better ways of surfacing important and possibly new undiscovered themes when applied to document collections of research articles. We plan to explore the potential of named entity and relation based topic models in our future work. (2) For this paper we only conducted intrinsic evaluations based on user studies. For future work, we plan to conduct extrinsic application based evaluation by using phrase LDA model parameters as feature weights for biomedical text classification.

Acknowledgments

This publication was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, US National Institutes of Health (NIH), through Grant UL1TR000117.

References

[1] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[2] D. Atkins, T. Rubin, M. Steyvers, M. Doeden, B. Baucom, and A. Christensen, "Topic models: A novel method for modeling couple and family text data." *Journal of family psychology*, vol. 26, no. 5, pp. 816–827, 2012.

[3] L.-w. Lehman, M. Saeed, W. Long, J. Lee, and R. Mark, "Risk stratification of icu patients using topic models inferred from unstructured progress notes," in *AMIA Annual Symposium Proceedings*, 2012, p. 505.

[4] H. M. Wallach, "Topic modeling: beyond bag-of-words," in *Proceedings of the 23rd international conference on Machine learning*, ser. ICML '06. ACM, 2006, pp. 977–984.

[5] X. Wang, A. McCallum, and X. Wei, "Topical n-grams: Phrase and topic discovery, with an application to information retrieval," in *Proceedings of the 7th IEEE Intl. Conf. on Data Mining*, ser. ICDM '07, 2007, pp. 697–702.

[6] J. Chang, J. Boyd-Graber, C. Wang, S. Gerrish, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Neural Information Processing Systems, NIPS*, 2009.

[7] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[8] A. J.-B. Chaney and D. M. Blei, "Visualizing topic models," in *International Conference of Weblogs and Social Media*, ser. ICWSM '12, 2012.

[9] M. I. Jordan, "Graphical models," *Statistical Science*, pp. 140–155, 2004.

[10] K. T. Frantzi, S. Ananiadou, and J.-i. Tsujii, "The c-value/nc-value method of automatic recognition for multi-word terms," in *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, ser. ECDL '98, 1998, pp. 585–604.

[11] S. Ananiadou, "A methodology for automatic term recognition," in *Proceedings of the 15th conference on Computational linguistics-Volume 2*, 1994, pp. 1034–1038.

[12] D. Bourigault, "Surface grammatical analysis for the extraction of terminological noun phrases," in *Proceedings of the 14th conference on Computational linguistics-Volume 3*, ser. ACL '92, 1992, pp. 977–981.

[13] I. Dagan and K. Church, "Termight: Identifying and translating technical terminology," in *Proceedings of the 4th conf. on Applied natural language processing*, 1994, pp. 34–40.

[14] C. Enguehard and L. Pantera, "Automatic natural acquisition of a terminology*," *Journal of quantitative linguistics*, vol. 2, no. 1, pp. 27–32, 1995.

[15] A. K. McCallum, "Mallet: A machine learning for language toolkit," 2002, http://mallet.cs.umass.edu.

[16] B. Shneiderman and C. Plaisant, "Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies," in *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, 2006, pp. 1–7.

[17] P. Saraiya, C. North, and K. Duca, "An evaluation of microarray visualization tools for biological insight," in *INFOVIS*, 2004, pp. 1–8.