Investigation of Input Optimization problem for Topic Modeling

Ali Daud[1], Muhammad Akram Shaikh[2], and Yan'an Jin[3]

[1]Department of Computer Science & Technology, 1-308, FIT Building, Tsinghua University, Beijing, China
[2]Department of Computer Systems & Software Engineering, Mehran University of Engineering and Technology, Jamshoro, Pakistan
[3]College of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China
ali_msdb@hotmail.com, akramshaikh@hotmail.com

ABSTRACT
This paper investigates the importance of input optimization for topic modeling by illustrating authors' interest modeling example. Finding authors' interest is an important problem investigated for fulfilling different recommendation tasks in academics social networks. Previously state-of-the-art Author-Topic model used grouping of author on the basis his co-authors in a single document. Intuitively, an author has a set of co-authors in a single document (subgroup) and a set of co-authors in all documents (group, which contains subgroup as a subset). We tried three methods to find authors interests with respect to different inputs but with similar structure. Firstly, traditional Author-Topic model (AT), in which each author of a subgroup is responsible for generating latent topics of that subgroup. Secondly, Inverse-Author-Topic model (IAT), in which each word of a subgroup is responsible for generating latent topics of that subgroup and thirdly, Governing-Author-Topic model (GAT), in which each author of a group, is responsible for generating latent topics of that group. Experimental results on the corpus downloaded from DBLP show that purposed method (GAT) with group based input outperformed baseline methods in terms of entropy measure, which shows better clustering ability of unsupervised methods. Additionally to ensure dominance of GAT all methods are applied for collaborator recommendation task and found GAT method to be the best, because of its optimal input.
Keywords: Input Optimization, Subgroup, Group, Authors Interests Topic Modeling, Unsupervised Learning

1 INTRODUCTION
Latent topic layer based topic modeling methods has been of great value because of capturing the semantics-based structure of words and relationships presents between the documents. Many researchers have been focusing on proposing new structure of topic models for various problems. For example, Topics over Time [1], Dynamic Topic model [2], Multi-scale Topic Tomography [3], and continuous time Dynamic Topic model [4] have been proposed to show how topic changes over time by simultaneously modeling time information for different time periods. Unfortunately all previous research works conducted for various problems, such as aforementioned problem of topic evaluation were focused on just proposing new structure of topic models by using similar single document as an input, and ignored input optimization problem for similar structure topic models for specific problem. The main motivation of this work is based on a successful expert finding work done previously, in which input for similar structure topic model is optimized to model the influence of semantics-based structure of words and relationships present between venues [5].

In this paper, we come up with a novel problem of input optimization for topic modeling methods. An example of authors' interests' topic modeling is investigated with three topic models with similar structure but different inputs to show the significance of this problem. Proposed different inputs are inspired from three different kinds of thoughts; 1) an author writes words to produce a document and inspires his co-authors in that single document (subgroup) based on his likeliness of research interests, such as AT in figure 2(b), 2) words in a document are probably inspiring authors to select research interests, such as IAT in figure 2(c) and 3) an author writes words to produce a document and inspires his co-authors in all documents (group) based on his likeliness of research interests, such as GAT in figure 2(d). Table 1 pictorially shows subgroup and group for Saumya K. Debray, which has in total 7 documents in this corpus with only 2 unique co-authors. Subgroup is just one document co-authored by her, while her group is collection of all the documents co-authored by her. Only processed titles of documents are used here for simplicity of representation.

Empirical study on DBLP corpus demonstrated the importance of studying input optimization for topic models as group input based method GAT performed better in terms of entropy and as well as proved most effective for collaborator recommendation task. To the best of our knowledge, we are the first to formulize the input optimization

problem for similar structure topics models by proposing subgroup and group notions, in which group can show more promising results by capturing better semantic-based structure of words and relationships between authors.

Table 1. Group and Subgroup Example

Group of Author "Saumya K. Debray"	Subgroup (Single Document)
reverse engineering itanium executables overlay automatic compaction kernel code demand code loading **Saumya K. Debray**,Gregory R. Andrews,Haifeng He	reverse engineering itanium executables Saumya K. Debray,Gregory R. Andrews,

The contributions of our work described in this paper are the followings:

1) Formulization of input optimization for topic models with similar structure for specific problem handling
2) Proposal of group based input rather than subgroup (document) based input for modeling authors interests
3) experimental verification of the effectiveness of GAT on the real-world corpus

The rest of the paper is organized as follows. In section 2, we formulize the problem. Section 3 illustrates briefly topic modeling and Latent Dirichlet Allocation followed by AT, IAT and GAT with its parameter estimation details. Section 4 discusses corpus, parameter settings, baseline methods, performance measures, and authors' interests modeling with empirical studies. Section 5 concludes this paper.

2 PROBLEM FORMULIZATION

Authors' interests finding addresses the task of discovering the right person related to a specific knowledge domain. The question can be like "Who is interested in writing on topics *Z*?" In general topic layer based authors interests finding process, main task is to probabilistically rank discovered authors for given number of topics. Latent topic layer based correlations between the authors is an appropriate way to find authors interests.

Formally for finding specific area authors, we need to calculate the probability $P(z|r)$ and $P(w|z)$ where z is a latent topic, r is author and w is the words. To find aforementioned probabilities; two generative scenarios are 1) author is responsible for generating latent topics which then generates words (real world situation) and 2) word is responsible for generating latent topics which then generates authors and; two grouping scenarios are 1) authors and words of a single document (subgroup) are responsible for generating latent topics and 2) authors and words of all documents (group) are responsible for generating latent topics. Based on the two generative and grouping scenarios, we define three different methods AT, IAT and GAT, which are similar in terms of structure but with different inputs.

1) Symbolically, for a AT, in which each author of a subgroup is responsible for generating latent topics of that subgroup, we can write it as: $D = (\mathbf{w}_1,\mathbf{a}_{d1})$, where $\mathbf{w_i}$ is word vector of a document (subgroup) part of a group and $\mathbf{a_{di}}$ is author vector of document $\mathbf{w_i}$.
2) Symbolically, for a IAT, in which each word of a subgroup is responsible for generating latent topics of that subgroup, we can write it as: $D = (\mathbf{a}_{d1},\mathbf{w}_1)$, where $\mathbf{w_i}$ is word vector of a document (subgroup) part of a group and $\mathbf{a_{di}}$ is author vector of document $\mathbf{w_i}$.
3) Symbolically, for a GAT, in which each author of a group is responsible for generating latent topics of that group, we can write it as: $G = \{(\mathbf{w}_1,\mathbf{a}_{d1})+ (\mathbf{w}_2,\mathbf{a}_{d2})+(\mathbf{w}_3,\mathbf{a}_{d3})+ \ldots + (\mathbf{w}_i,\mathbf{a}_{di})\}$, where G is a group, $\mathbf{w_i}$ is word vector of a subgroup and $\mathbf{a_{di}}$ is author vector of subgroup $\mathbf{w_i}$.

Here two things need to be clearly explained. First, from words of document means only the title words of the paper (instead of using whole paper or abstract) which are usually real representative of the document and contains most important words to explain the main theme of the paper. Some preliminary/practical experiments show that there is no significant performance difference if one uses only title words, while on the other hand time complexity for model learning is significantly decreased. Second, we have not used any algorithm like k-means, principle component analysis or any others to first select best features as input for the similar structure topic models. Therefore, we can say that we are not doing input optimization by feature selection using complex algorithms. Nevertheless, one can focus on this problem by finding better matched algorithms of feature selection with specific topic models.

3 AUTHORS INTERESTS TOPIC MODELING

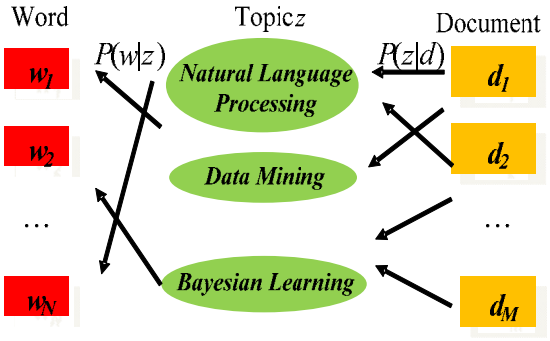
Before explaining our proposed Governing-Author-Topic (GAT) Model for modeling authors interests with optimal input, we first briefly introduce topic modeling and state-of-the-art topic modeling approach Latent Dirichlet Allocation (LDA) [6] for modeling text information of documents. Later, Author-Topic Model (AT) [7] and Inverse-Author-Topic Model (IAT) are explained, which are extensions of LDA for modeling both text and authors information simultaneously for finding authors interests.

3.1 Topic Modeling

Fundamental topic modeling assumes that there is a hidden topic layer $Z = \{z_1, z_2, z_3, \ldots, z_t\}$ between the word tokens and the documents, where z_i denotes a latent topic and each document d is a vector of N_d words $\mathbf{w_d}$. A collection of D documents is defined by $D = \{\mathbf{w_1}, \mathbf{w_2}, \mathbf{w_3}, \ldots, \mathbf{w_d}\}$ and each word w_{id} is chosen from a vocabulary of size V. For each document, a topic mixture distribution is sampled and a latent topic Z is chosen with the probability of topic given document for each word with word having generated probability of word given topic [6,8].

Figure 1 provides pictorial representation of topic modeling, in which topic layer is used between words and documents to match documents with the queries. We explain the role of topic layer with the help of an information retrieval example. Suppose a user enters a query natural language processing and following two papers can be retrieved. First paper title contains the query words natural language processing so found related to the query, while second paper title includes dependency parsing not included in the users query words even then it is found related to a query because of semantic similarity of natural language processing and dependency parsing words in a topic "Natural Language Processing" whose top ten words with their assigned probabilities are shown in figure 1.

- Paper1: A Maximum Entropy Approach to **Natural Language Processing**
- Paper2: A Pipeline Framework for **Dependency Parsing**



Latent Topic z	
Word	*Prob.*
Natural	0.121514
Language	0.059027
Parsing	0.0547
Word	0.052969
Processing	0.050199
Dependency	0.045179
Information	0.03237
Grammar	0.025447
Tagging	0.020773
Sense	0.020081

Figure 1. Topic Modeling

3.2 Latent Dirichlet Allocation (LDA)

LDA [6,9] is a state-of-the-art topic modeling approach which makes use of latent topic layer to capture semantic dependencies between the words. It is a three-level Bayesian network that generates a document using a mixture of topics. It generates a document in a three steps process. First, for each document d, a multinomial distribution θ_d over topics is randomly sampled from a Dirichlet distribution with parameter α. Second, for each word w, a topic z is chosen from this topic distribution. Finally, the word w is generated by randomly sampling from a topic-specific multinomial distribution Φ_z. So, the generating probability of word w from document D is given as:

$$P(w|d,\theta,\emptyset) = \sum_{z=1}^{T} P(w|z,\emptyset_z)P(z|d,\theta_d) \quad (1)$$

3.3 Author-Topic Model (AT)

Following topic modeling basic idea of modeling words and documents, words and authors are modeled by considering latent topics to discover the research interests of authors [7]. The main idea of AT is based on the assumption that authors are responsible for generating words of documents on the basis of their research interests by using latent topic layer. In this model, each author (from set of K authors) of a document d is associated with a multinomial distribution θ_a over topics is sampled from Dirichlet α and each topic is associated with a multinomial distribution Φ_z sampled from Dirichlet β over words of a document for that topic. The generating probability of

word *w* for author *r* of a document *d* is given in Eq. 2. It has successfully discovered authors' interests and semantic relationships between them.

$$P(w|r,d,\emptyset,\theta) = \sum_{z=1}^{T} P(w|z,\emptyset_z)P(z|r,\theta_r) \tag{2}$$

3.4 Inverse-Author-Topic Model (IAT)

The reverse of the basic idea of AT model is that words are responsible for generating authors of documents on the basis of their semantic similarities by using latent topic layer, which is main idea of IAT. In this model, each word (from set of *W* words) of a document *d* is associated with a multinomial distribution θ_w over topics is sampled from Dirichlet α and each topic is associated with a multinomial distribution Φ_z sampled from Dirichlet β over authors of a document for that topic. The generating probability of author *r* for word *w* of a document *d* is given in Eq. 3. It can be used to discover authors' interests and semantic relationships between them.

$$P(r|w,d,\emptyset,\theta) = \sum_{z=1}^{T} P(r|z,\emptyset_z)P(z|w,\theta_w) \tag{3}$$

3.5 Governing-Author-Topic Model (GAT)

The main idea of GAT is based on the assumption that authors group is responsible for generating words of groups on the basis of their research interests by using latent topic layer. The intuition is based on the fact that usually all the co-authors have an influence on a single authors interests which makes an explicit group of people with similar interest. We believe that this kind of grouping in authors' social network provides realistic view of authors' interests and relationships.

In this model, each author (from set of *K* authors) of a group is associated with a multinomial distribution θ_a over topics and each topic is associated with a multinomial distribution Φ_z over words of a venue for that topic. Both θ_a and Φ_z have symmetric Dirichlet prior with hyper parameters α and β. The generating probability of the word *w* for author *r* of a group *g* is given as:

$$P(w|r,g,\emptyset,\theta) = \sum_{z=1}^{T} P(w|z,\emptyset_z)P(z|r,\theta_r) \tag{4}$$

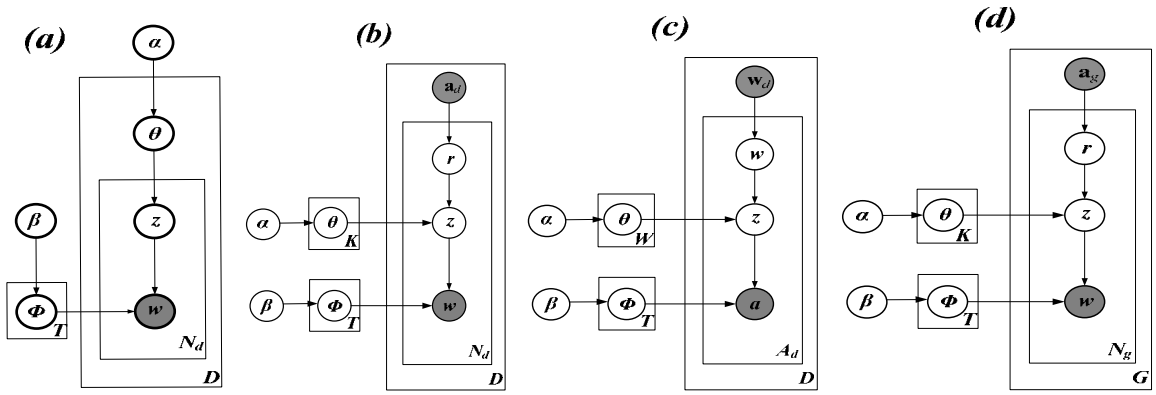


Figure 2. Authors Interests Topic Modeling a) Latent Dirichlet Allocation b) Author-Topic Model c) Inverse-Author-Topic Model and d) Governing- Author-Topic Model

The generative process of GAT is as follows:
For each author $r = 1,\ldots, K$ of a group *g*
Choose θ_r from Dirichlet (α)
For each topic $z = 1,\ldots, T$
Choose Φ_z from Dirichlet (β)
For each word $w = 1,\ldots, N_g$ of group *g*
Choose an author *r* uniformly from all authors $\mathbf{a}_g$
Choose a topic *z* from multinomial (θ_r) conditioned on *r*
Choose a word *w* from multinomial (Φ_z) conditioned on *z*

Gibbs sampling is utilized [1] for parameter estimation in our method which has two latent variables *z* and *r*; the conditional posterior distribution for *z* and *r* is given by:

$$P(z_i = j, r_i = k | w_i = m, \mathbf{z}_{-i}, \boldsymbol{r}_{-i}, \mathbf{a}_g) \propto \frac{n_{-i,j}^{(wi)} + \beta}{n_{-i,j}^{(.)} + V\beta} \frac{n_{-i,j}^{(ri)} + \alpha}{n_{-i,.}^{(ri)} + R\alpha} \quad (5)$$

where $z_i = j$ and $r_i = k$ represent the assignments of the i^{th} word in a group to a topic j and author k respectively, $w_i = m$ represents the observation that i^{th} word is the m^{th} word in the lexicon, and z_{-i} and r_{-i} represents all topic and author assignments not including the i^{th} word. Furthermore, $n_{-i,j}^{(wi)}$ is the total number of words associated with topic j, excluding the current instance, and $n_{-i,j}^{(ri)}$ is the number of times author k is assigned to topic j, excluding the current instance, W is the size of the lexicon and R is the number of authors. "." Indicates summing over the column where it occurs and $n_{-i,j}^{(.)}$ stands for number of all words that are assigned to topic z excluding the current instance.

During parameter estimation, the algorithm only needs to keep track of W x Z (word by topic) and Z x R (topic by author) count matrices. From these count matrices, topic-word distribution Φ and author-topic distribution θ can be calculated as:

$$\emptyset_{zw} = \frac{n_{-i,j}^{(wi)} + \beta}{n_{-i,j}^{(.)} + V\beta} \quad (6)$$

$$\theta_{rz} = \frac{n_{-i,j}^{(ri)} + \alpha}{n_{-i,.}^{(ri)} + R\alpha} \quad (7)$$

where, $\emptyset_{zw}$ is the probability of word w in topic z and θ_{rz} is the probability of topic z for author r. These values correspond to the predictive distributions over new words w and new topics z conditioned on w and z.

For better understanding of difference between proposed and related models, Table 2 provides the general description of models and problems handled by using these models.

Table 2. Generative summary of GAT and related Models

Model	**Summarized Generative Process**
AT	An author of a document is responsible for generating words for that document on the basis of latent topics.
IAT	A word in a document is responsible for generating authors for that document on the basis of latent topics.
GAT	A group of an author of documents is responsible for generating words for that group on the basis of latent topics.

4 EXPERIMENTS

This part discusses about corpus and related settings. Then, we report results and discussions by using two different performance evaluation measures, which are entropy and recommendation accuracy.

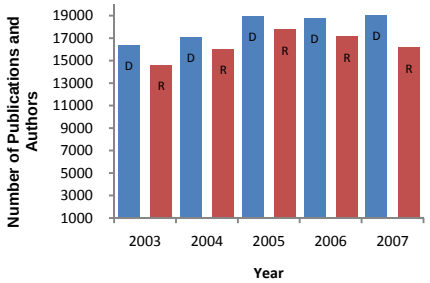


Figure 3. Histogram Illustrating Data Distribution

4.1 Corpus

We downloaded five years (2003-2007) research publications corpus of conferences from DBLP [10]. In total, we extracted 112,317 authors, 90,124 publications for 261 conferences. We then processed corpus by a) removing stop-words, punctuations and numbers b) down-casing the obtained words of publications, and c) removing words and

authors that appear less than three times in the corpus. This led to a vocabulary size of V=10,872, a total of 572,592 words and 26,078 authors in the corpus. Figure 3 shows fairly smooth yearly data distribution for number of publications (*D*) and authors (*R*) in the conferences.

There is certainly some noise in data of this form especially author names which were extracted automatically by DBLP from PDF, postscript or other document formats. For example, for some very common names there can be multiple authors (like L Ding or J Smith). This is a known as limitation of working with this type of data (please see [11] for details). There are algorithmic techniques for name disambiguation that could be used to automatically solve these kinds of problems; however, in this work we do not focus on this.

4.2 Parameter Settings

One can estimate the optimal values of hyper-parameters α and β (fig. 3 (b)) by using Expectation-Maximization [8] or Gibbs sampling algorithm [9,12]. EM algorithm is susceptible to local maxima and computationally inefficient [6], consequently Gibbs sampling algorithm is used. For some applications topic models are sensitive to the hyper parameters and need to be optimized. For application in this paper, we found that our topic model based methods are not sensitive to the hyper parameters. In our Gibbs sampling algorithm based experiments, for 150 topics Z the hyper-parameters α and β were set at $50/Z$ and 0.01 respectively, by following the values used in [7]. The number of topics Z was fixed at 150 on the basis of human judgment of meaningful topics and measured perplexity [13,14]. We ran 3 independent Gibbs sampling chains for 1000 iterations each. All experiments were carried out on a machine running Windows XP 2006 with Intel(R) Core(TM)2 Duo CPU T5670 (1.80 GHz) and 2 GB memory.

4.3 Baseline Methods

We compared proposed GAT with AT and IAT by using same number of topics for comparability. The number of Gibbs sampler iterations used for baseline methods are 1000 (with 3 independent Gibbs sampling chains for 1000 iterations each) and parameter values same as the values used for GAT.

4.4 Performance Measures

In our experiments, at first we used average entropy (under root of perplexity [13]) to measure the quality of discovered topics, which reveals the purity of topics. Entropy (please see eq. 8) is a measure of the disorder of system, less intra-topic entropy is usually better. Secondly, we evaluated proposed methods by showing their application for collaborator recommendation by using ranking accuracy performance measure. It shows the ability to produce an ordered list of objects that matches how a user would have listed the same objects [15,16]. Our ultimate goal is to measure the effectiveness of suggesting top-ranked objects (co-authors) for each research paper. That is, each method needs to recommend the top k authors. Firstly, for each paper d, we randomly withhold one joined author r from his original set of joined coauthors to form papers training dataset. Secondly, for each paper d, we select k-1 additional random authors that were not in paper d's original set; the withheld author r together with these k-1 other authors form paper d's evaluation set (of size k, which is 2,4,6,8 and 10 in this work). For paper d, all methods calculate the score for each of the k authors in the evaluation set. Lastly, for each paper d, we order the k authors in his evaluation set by their predicted score to obtain a corresponding rank between 1 and k for each. Our objective is to find the relative position of each paper d's withheld author r. There are k possible ranks for r, ranging from the best rank where no random author precedes r in the ordered list, to the worst rank where all of the random authors appear before r. The best result we can hope for is that author r will precede the k-1 random objects in our ranking. Similar kind of evaluation method is used for community recommendation using Latent Dirichlet Allocation by Chen et al. [17], in which they ranked randomly withhold communities.

$$Entropy\ of\ (Topic) = -\sum_z P(z) log_2[P(z)] \tag{8}$$

4.5 Results and Discussions

Authors Interests: We extracted and probabilistically ranked authors related to a specific area of research on the basis of latent topics. Table 3 shows authors' interests for different topics. It illustrates 4 topics out of 150, discovered from the 1000th iteration of the particular Gibbs sampler run. The words associated with each topic are quite intuitive and precise and depict a real picture of specific area of research. For example, topic # 19 "Semantic Web" shows quite specific and meaningful vocabulary (semantic, web, ontology, owl, rdf, annotation, semantics, and knowledge) when a user is searching for semantic web related documents or authors. Other topics, such as "Information Retrieval", "Image Retrieval" and "Web Security are quite descriptive that shows the ability of GAT to discover precise topics. The interested authors found associated with each topic are quite representative, as we have

analyzed and found that authors related to different topics are typically writing for that area of research. For example, in case of topic 19 “Semantic Web” top ranked authors web pages shows their interest in semantic web research topic and they are mostly publishing on this topic.

Table 3. Illustration of 4 topics with related authors, the titles are our interpretation of the topics

Topic 19 “Semantic Web”		**Topic 115 “Information Retrieval”**		**Topic 114 “Image retrieval”**		**Topic 73 “Web Security”**	
Word	**Prob.**	**Word**	**Prob.**	**Word**	**Prob.**	**Word**	**Prob.**
semantic	0.260961	retrieval	0.170177	image	0.217983	security	0.142105
web	0.138429	information	0.116340	retrieval	0.080310	attacks	0.062163
ontology	0.124851	query	0.051240	based	0.072752	secure	0.057901
ontologies	0.060605	feedback	0.041957	images	0.036579	protocols	0.051141
owl	0.033670	document	0.041586	segmentation	0.030640	protocol	0.043500
rdf	0.026936	relevance	0.038492	content	0.029695	analysis	0.021310
annotation	0.018105	search	0.028467	color	0.024162	integrity	0.021016
semantics	0.016670	evaluation	0.024754	region	0.019572	authentication	0.020281
approach	0.014352	expansion	0.023516	analysis	0.019437	attack	0.019693
knowledge	0.012365	term	0.021908	medical	0.019303	smart	0.018811
Author	**Prob.**	**Author**	**Prob.**	**Author**	**Prob.**	**Author**	**Prob.**
Carole A Goble	0.018056	W Bruce Croft	0.012987	Wei Ying Ma	0.036562	Angelos D Keromytis	0.014517
Robert Stevens	0.014153	Ryen W White	0.012304	Lei Zhang	0.023340	Wei Zhao	0.011956
Peter Haase	0.013177	Cheng Xiang Zhai	0.011621	Hong Jiang Zhang	0.020229	John C Mitchell	0.011103
Amit P Sheth	0.012201	David Carmel	0.011621	Bo Zhang	0.016341	Vitaly Shmatikov	0.010250
Steffen Staab	0.011714	Susan T Dumais	0.010938	Mingjing Li	0.015563	Trent Jaeger	0.009396
Phillip W Lord	0.011714	Mounia Lalmas	0.010254	Hanqing Lu	0.014785	Sushil Jajodia	0.008543
Luc Moreau	0.010738	Charles L A Clarke	0.010254	Xing Xie	0.012452	Klaus Rothbart	0.008543
Anupam Joshi	0.010250	Nick Craswell	0.010254	Xiaofei He	0.012452	Andrew S Tanenbaum	0.007689
Ian Horrocks	0.009762	Justin Zobel	0.010254	Xiaoou Tang	0.010896	Martin Abadi	0.007689
David DeRoure	0.009762	James P Callan	0.008888	Zhiwei Li	0.010896	Patrick Mc Daniel	0.006836

Proposed approach discovered several other topics such as data mining, neural networks, algorithms, graphs, XML databases and pattern recognition, also other topics that span the full range of areas encompassed in the dataset. In addition, by doing analysis of authors’ home pages and DBLP [10], we found that all authors assigned with higher probabilities have published many papers on their relevant topics. In the following we provide the links to the home pages of top five authors related to semantic web topic for confirmation.

http://www.cs.manchester.ac.uk/~carole/
http://www.cs.manchester.ac.uk/~stevensr/
http://semanticweb.org/wiki/Peter_Haase
http://knoesis.wright.edu/amit/
http://www.uni-koblenz.de/~staab/

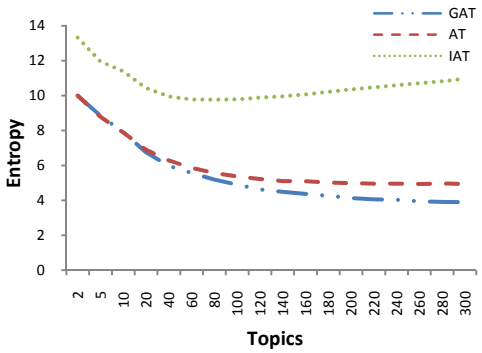


Figure 4. Average Entropy curve as a function of different number of topics, lower is better

Entropy based Comparison: We provide quantitative comparison between proposed and baseline approaches. Figure 4 shows the average entropy of topic-word distribution for all topics measured by using eq. 8. Lower entropy for different number of topics T= 2, 5, 10, …, 300 proves the effectiveness of GAT for obtaining better topics. The performance difference for different number of topics between GAT and IAT is pretty much even and clear, which corroborate that proposed approach superiority is not sensitive to the number of topics in comparison to the IAT. The performance of GAT and AT for topics 2, 5, 10, 60 is pretty much the same. But GAT performed better than

AT when number of topics increases from 60, which corroborate that GAT has less sparse topics as compared to AT. Less sparse topics can result in the better performance of topic modeling approach. The relationship between less sparse topics and the better performance of a topic model for specific tasks, is investigated in detail for expert finding [5] and conference mining [18] problems for large dataset in academic social networks.

Collaborator Recommendation Application based Comparison: We show the effectiveness of proposed approach for an important collaborator recommendation task in academics social networks. It is focused on suggesting collaborators to authors, usually on the basis of similar area of research or research projects they are interested in. We compared GAT with AT and IAT. Figure 5 shows topic wise accuracy of all approaches for number of topics from 20,40,…,200. GAT approach consistently outperformed AT and IAT due to exploiting group and subgroups structures together for the text and authors. While we can also see that GAT performance is very much stable as compared to the baselines for different number of topics. It shows that GATs superior performance is not affected by different number of topics. IAT and AT both have an impact on performance for different number of topics.

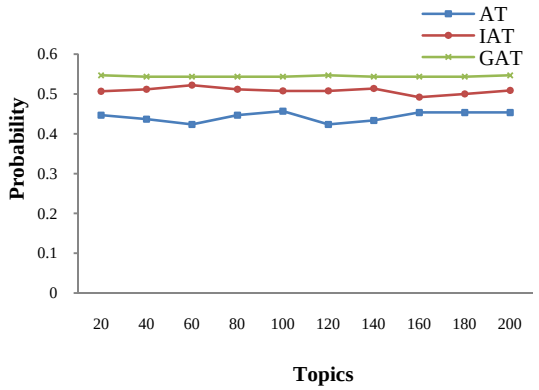


Figure 5. Collaborator recommendation (topic wise)

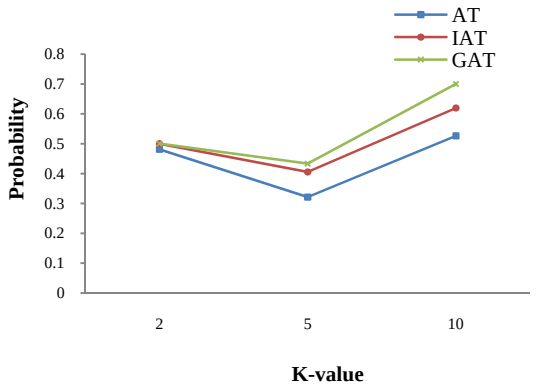


Figure 6. Collaborative recommendation (top k-recommendations)

Figure 6 shows Top K-Recommendation wise accuracy for collaborative recommendation task for AT, IAT and GAT for values of k = 2,5, and 10. Overall GAT approach performed better and has stable performance for different values of k as compared to baselines.

Table 4. Collaborator Recommendation

Average Accuracy	**AT**	**IAT**	**GAT**
Collaborator Recommendation	0.443	0.508	0.544

In table 4, we summarize the average accuracy results for collaborator recommendation task. One can see that group based input oriented approach GAT performed 10.1% and 3.6% better than AT and IAT approaches, respectively in terms of accuracy for collaborator recommendation task, which is significant. It is interesting that IAT performed better than AT for this task, which shows that words generating authors can be better choice when group structures are not exploited. As when group structures are exploited and authors generated words in the GAT the best performance for this task is obtained.

5 CONCLUSIONS

This study investigates an input optimization problem for similar structure topic modeling methods when they are used for specific problem, such as authors' interest finding as an investigation point. Traditional Author-Topic model and its two variants in terms of different inputs have shown the significance of finding optimal input in topic modeling domain. Proposed group based input method GAT proved to be effective for finding topical authors interests in terms of entropy, which shows its ability to produce refined topics of probabilistically related words and authors relationships (please see details for impact of refined topics on the better performance of method in [8] for conference mining problem). It is evident from entropy based comparison that performance of similar structure topic models can also be increased by input optimization other than by just proposing new model structure. The effectiveness of method with optimized input is also confirmed for collaborator recommendation task. As a whole,

we conclude, that it is significant to optimize the input of topic models with similar structure to get the optimized results for solving different problems.

Acknowledgements. The work is supported by the Higher Education Commission (HEC), Pakistan. We are thankful to Jie Tang for sharing his topic modeling codes.

REFERENCES

[1] Wang, X., and McCallum, A. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, USA, August 20-23, 2006.

[2] Blei, D. M., and Lafferty, J. Dynamic Topic Models. In Proceedings of 23rd International Conference on Machine Learning (ICML), Pittsburgh, Pennsylvania, USA, June 25-29, 2006.

[3] Nallapati, R., Cohen, W., Ditmore, S., Lafferty, J., and Ung, K. Multiscale Topic Tomography. In Proceedings of 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007.

[4] Wang, C., Blei, M. D., and Heckerman, D. Continuous Time Dynamic Topic Models. In Proceedings of Conference on Uncertainty in Artificial Intelligence (UAI), Helsinki, Finland, July 9-12, 2008.

[5] Daud, A., Li, J., Zhu, L., and Muhammad, F. A Generalized Topic Modeling Approach for Maven Search. In Proceedings of International Asia-Pacific Web Conference and Web-Age Information Management (APWEB-WAIM), Q. Li et al. (Eds.): APWeb/WAIM 2009, LNCS 5446, pp. 138–149, 2009.

[6] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet Allocation. Journal of Machine Learning and Research (JMLR), vol. 3, pp. 993-1022, 2003.

[7] Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. The Author-Topic Model for Authors and Documents. In Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI), Banff, Canada, 2004.

[8] Hofmann T. Probabilistic Latent Semantic Analysis. In Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI), Stockholm, Sweden, 1999.

[9] Griffiths, T. L., and Steyvers, M. Finding Scientific Topics. In Proceedings of the National Academy of Sciences (NAS), USA, pp. 5228-5235, 2004.

[10] DBLP Bibliography Database. http://www.informatik.uni-trier.de/~ley/db/.

[11] Newman, M. E. J. Scientific Collaboration Networks: I. Network Construction and Fundamental Results. Physical Review E, 64, 016131, 2001.

[12] Andrieu ,C., Freitas, N. D., Doucet, A. and Jordan, M. An Introduction to MCMC for Machine Learning. Journal of Machine Learning (JMLR), vol. 50, pp. 5–43, 2003.

[13] Azzopardi, L., Girolami, M., Risjbergen, and K. van. Investigating the Relationship between Language Model Perplexity and IR Precision-Recall Measures. In Proceedings of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, July 28-August 1, 2003.

[14] Daud, A., Li, J., Zhou, L., and Muhammad, F. Knowledge Discovery through Parametric Directed Probabilistic Topic Models. a Survey. Journal of Frontiers of Computer Science in China (FCS), DOI:10.1007/s11704-009-0062-y, 2009.

[15] Jarvelin, K., and Kekalainen, J. Cumulated Gain-Based Evaluation of IR Techniques. ACM Transactions on Information Systems, vol. 20(4), pp. 422-446, 2002.

[16] Koren, Y. Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 426-434, 2008.

[17] Chen, W.Y., Chu, J.C., Luan, J., Bai, H., Wang, Y., and Chang, E. Y. Collaborative Filtering for Orkut Communities: Discovery of User Latent Behavior. In Proceedings of the 18th International conference on World Wide Web (WWW), April 20-24, Madrid, Spain, 2009.

[18] Daud, A., Li, J., Zhu, L., and Muhammad, F. Conference Mining via Generalized Topic Modeling. In Proceedings of International European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases (ECML PKDD), W. Buntine et al. (Eds.): ECML PKDD 2009, Part I, LNAI 5781, pp. 244–259, 2009.