Exploiting Explicit Semantics-based Grouping for Author Interest Finding

Ali Daud

Department of Computer Science, International Islamic University,
Islamabad, Pakistan 44000
ali.daud@iiu.edu.pk

Abstract . This paper investigates the problem of finding author interest in co-author network through topic modeling with providing several performance evaluation measures. Intuitively, there are two types of explicit grouping exists in research papers (1) authors who have co-authored with author A in one document (subgroup) and (2) authors who have co-authored with author A in all documents (group). Traditional methods use graph-link structure by using keywords based matching and ignored semantics-based information, while topic modeling considered semantics-based information but ignored both types of explicit grouping e.g. State-of-the-art Author-Topic model used only one kind of explicit grouping single document (subgroup) for finding author interest. In this paper, we introduce Group-Author-Topic (GAT) modeling which exploits both types of grouping simultaneously. We compare four different topic modeling methods for same task on large DBLP dataset. We provide three performance measures for method evaluation from different domains which are; perplexity, entropy, and prediction ranking accuracy. We show the trade of between these performance evaluation measures. Experimental results demonstrate that our proposed method significantly outperformed the baselines in finding author interest. The trade of between used evaluation measures shows that they are equally useful for evaluating topic modeling methods.

Keywords: Author Interest, Subgroup and Group, Co-author Network, Performance Measures, Topic Modeling

1 Introduction

Social network analysis has been an active area of research with the proliferation of social applications in different social networks, e.g. Academic social networks such as DBLP and CiteSeer, tagging networks such as Bibsonomy and Delicious, video sharing networks such as Flicker and YouTube, blogging networks such as Blogger and WordPress. Academic social network or Co-author network have several knowledge discovery problems which are useful for fulfilling different suggestion or recommendation tasks. Author interest finding is one of the interesting problems useful for suggesting reviewers for papers, finding collaborators for projects, finding supervisors, finding program committee members for conferences etc.

Co-author network provide the basis for exploiting author interest. Intuitively, there is two type of natural grouping exists in co-author networks (1) Authors who have co-authored with author A in a document (subgroup) e.g. Fig. 1 shows a single document or subgroup which consists of paper title words and co-authors of that paper and (2) authors who have co-authored with author A in all documents (group) which consists of all papers title words and all the co-authors who have written papers with Author A in those papers. For example Fig. 1 show a group of 4 papers title words and co-authors for author A.

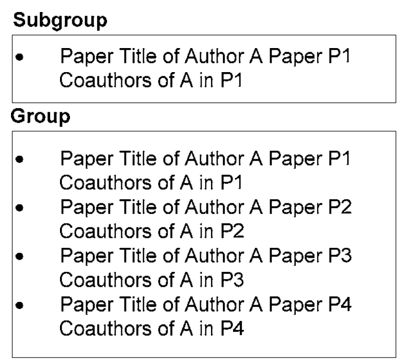

Figure 1: An illustration of Group and Subgroup.

Previously, three major frameworks used to identify the author interest are (1) stylistic features (such as sentence length), author attribution and forensic linguistics based methods to identify what author wrote a given piece of text [7,9] (2) graph-link structure based methods by using keywords as a basis for representation and analysis for relationships among authors [12,16] and (3) topic modeling based methods for capturing semantics-based intrinsic structure of words presented between subgroups [11,14,15]. Above mentioned frameworks based on writing styles and network connectivity ignored the semantics-based intrinsic structure of words, while semantics-based topic modeling methods exploited grouping at only subgroup and ignored grouping at group level.

In this paper, we investigate the problem of author interest finding by proposing GAT which models the author interest and relationships by considering both type of explicit grouping. Experimental results and discussions elaborate the significance of problem and usefulness of our method. We should mention that exploitation of author interest (writing habits without considering his research level) and expert finding (writing habits with considering his research level) are notably two different knowledge discovery problems [4].

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

(1) formulization of author interest finding problem from subgroup to group level

(2) demonstrate that perplexity and entropy (for train and test data) is equally useful for evaluating topic models performance with the fact that entropy provides more lucid results

To the best of our knowledge, we are the first to deal with modeling author interest finding problem by proposing group level method and experimentally showing the relationship between perplexity and entropy.

The rest of the paper is organized as follows. Section 2 illustrates our proposed method and related methods for finding author interest. Section 3 discusses corpus, parameter settings, performance measures, baseline methods, and results and discussions and section 4 concludes this paper.

2 Author Interest Topic Modeling

In this section, before describing our group author interest topic modeling, we first briefly introduce topic modeling idea followed by author-topic model, inverse-author-topic model and conditionally-independent-author-topic model.

2.1 Topic Modeling

Topic modeling brought new notion to the unsupervised learning methods by providing soft clusters of data. Instead of using just a keyword as a measure of relationship for collection of documents in traditional language models fundamental topic modeling assumes that there is a hidden topic layer $Z = \{\mathbf{z_1}, \mathbf{z_2}, \mathbf{z_3}, \ldots, \mathbf{z_t}\}$ between the word tokens and the documents, where $\mathbf{z_i}$ 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 W. 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 [2,10]. The generating probability of word w for a document d for the state-of-the-art topic model Latent Dirichlet Allocation is given in Eq. 1.

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

2.2 Author-Topic Model (AT)

AT [15] is a two way stochastic process which is based on the idea that author thinks about a topic and starts writing a paper on that topic with the help of co-authors. In AT a randomly chosen author from a subgroup is responsible for generating words of a document. In AT, each author (from set of A authors) of a document d is associated with a multinomial distribution θ_a over topics which is sampled from Dirichlet α and each topic is associated with a multinomial distribution Φ_z which is 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. AT has successfully discovered topically related authors but did not consider explicit group information.

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

2.3 Inverse-Author-Topic Model (IAT)

IAT is a two way stochastic process which is based on the idea that a randomly chosen word from a subgroup is responsible for generating authors of a document. This idea is opposite to the basic idea of AT. In IAT, each word (from set of W words) of a document d is associated with a multinomial distribution θ_w over topics which is sampled from Dirichlet α and each topic is associated with a multinomial distribution Φ_z which is 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. IAT did not consider explicit group information.

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

2.4 Conditionally-Independent-Author-Topic Model (CIAT)

CIAT is based on the idea that words and authors are independently generated from a subgroup which is a variation of GM-LDA used for image annotation [3]. AT and IAT assumes that randomly chosen author or word generates a topic, respectively. On the contrary CIAT assumes that authors and words are independently generated by the topic. In CIAT, topics are sampled from multinomial distribution θ with Dirichlet α over words and authors and each word and author is associated with a multinomial distribution Φ_z which is sampled from Dirichlet β over words and a multinomial distribution ψ_z which is sampled from Dirichlet μ over authors of a document, respectively. The generating probability of word w and author r of a document d is given in Eq. 4. CIAT did not consider explicit group information.

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

2.5 Group-Author-Topic Modeling (GAT)

GAT is a two way stochastic process which is based on the idea that author thinks about a topic and his thinking is influenced by all co-authors of his papers. In GAT a randomly chosen author from a group is responsible for generating words of a group. In the proposed approach, we viewed a group as a composition of authors all documents (subgroups). Symbolically, for a group G 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 d_i is a subgroup of a group and a_{di} are the author (s) of subgroup d_i.

Subgroup based methods considers that an author is responsible for generating latent topics of the document, while, group based method considers that an author is responsible for generating latent topics of the group (please see Fig. 1 and 3). In GAT, each author (from set of A authors) of a group g is associated with a multinomial distribution θ_a over topics which is sampled from Dirichlet α and each topic is associated with a multinomial distribution Φ_z which is sampled from Dirichlet β over words of a group for that topic. The generating probability of word w for author r of a group g is given in Eq. 5.

$$P(w|r,g,\emptyset,\theta) = \sum_{z=1}^{T} P(w|z,\emptyset_z)P(z|r,\theta_r) \tag{5}$$

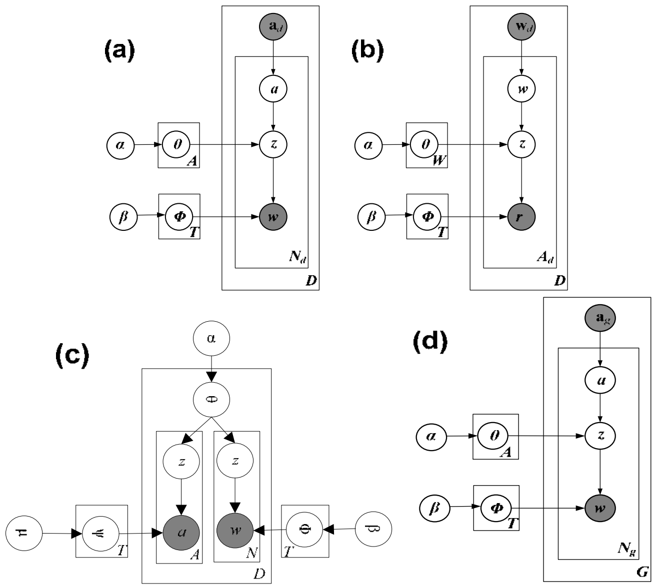

Figure 2. GAT (d) is shown with three related Models (a) Author-Topic (AT) Model, (b) Inverse-Author-Topic (IAT) Model, and (c) Conditionally-Independent-Author-Topic (CIAT) Model.

The generative process of GAT is as follows:
For each author $r = 1,\ldots, K$ of 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] to solve all related methods and our proposed method which has two latent variables z and r; the conditional posterior distribution for z and r is given by:

$$P(z_i = j, r_i = k \middle| w_{\boldsymbol{i}} = m, \mathbf{z}_{-i}, \boldsymbol{r}_{-\boldsymbol{i}}, \mathbf{a}_{\boldsymbol{g}}) \propto \frac{n_{-i,j}^{(wi)}+\beta}{n_{-i,j}^{(\cdot)}+W\beta} \frac{n_{-i,j}^{(ri)}+\alpha}{n_{-i,\cdot}^{(ri)}+A\alpha} \tag{6}$$

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 *A* 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.

For parameter estimation model needs to keep track of *W* x *Z* (word by topic) and *Z* x *A* (topic by author) count matrices for group. From these count matrices, topic-word distribution *Φ* and author-topic distribution *θ* can be calculated as given in Eq. 7 and Eq. 8, 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*.

$$\emptyset_{zw} = \frac{n_{-i,j}^{(wi)} + \beta}{n_{-i,j}^{(.)} + W\beta} \quad (7)$$

$$\theta_{rz} = \frac{n_{-i,j}^{(ri)} + \alpha}{n_{-i,.}^{(ri)} + A\alpha} \quad (8)$$

3 Experiments

3.1 Corpus

We downloaded five years paper corpus of conferences from DBLP database [6], by only considering conferences for which data was available for the years 2003-2007. In total, we extracted 112,317 authors and 90,124 papers. We then processed corpus by (a) removing stop-words, punctuations and numbers (b) down-casing the obtained words of papers, 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.

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 (e.g. L Ding or J Smith). This is a known as limitation of working with this type of data (please see [13] 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 name disambiguation problems.

3.2 Parameter Settings

Estimation of hyper-parameters *α* and *β* is done by using Gibbs sampling algorithm [8]. 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 experiments, for different number of topics Z the hyper-parameters α, β and μ were set at 50/Z, 0.01 and 0.01.

3.3 Performance Measures

We used three performance measures for evaluating the performance of methods from different domains (1) perplexity; a standard performance measure for evaluating topic modeling (soft clustering) by examining the generative power of trained model on unseen dataset. Lower values of perplexity indicate better generalization power of model on the words of test documents by the trained topics. For a test set of D documents the perplexity is given in Eq. 9, (2) entropy (under root of perplexity) for training set and testing set to measure the quality of discovered topics, which reveals the purity of topics is given in Eq. 10, less intra-topic entropy is better; a performance measure for evaluating traditional clustering (hard clustering), and (3) Prediction ranking accuracy; a performance measure for evaluating recommendation is given in Eq. 11. We employ the top-k recommendations, that is, each ranking algorithm needs to recommend the top k objects (words and authors) for documents by ranking randomly withhold objects from the original set mixed with objects not from the original set.

$$perplexity\ (D_{test}) = \exp\ \{-\frac{\sum_{d=1}^{D} \log p(\mathbf{w}_d)}{\sum_{d=1}^{D} N_d}\} \tag{9}$$

$$Entropy\ of\ (Topic) = -\sum_z P(z) log_2[P(z)] \tag{10}$$

$$Prediction\ Ranking\ Accuracy = \frac{k\ value + 1 - rank\ of\ object}{k\ value} \tag{11}$$

3.4 Baseline Methods

We compare our proposed method GAT with AT [15], which considers that words and authors of a document are dependent on each other and authors are responsible for generating words of a document, IAT, which considers that words and authors of a document are dependent on each other and words are responsible for generating authors of a document, and CIAT, which considers that words and authors of a document are independent of each other and topics are responsible for generating words and authors of a document.

3.5 Results and Discussions

We extracted authors related to a specific area of research on the basis of semantics-based similarity of topic so called latent topics. Table 1 shows authors' interests for different topics by using GAT. It illustrates 3 topics out of 150, discovered from the

1000th iteration of the particular Gibbs sampler run. The words and authors associated with each topic are quite 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 “Pattern Mining” and “Information Retrieval” are quite descriptive that shows the ability of GAT to discover precise topics. 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 74 “Semantic Web” top ranked authors web pages shows their interest in semantic web research topic and they are mostly publishing on this topic.

GAT also discovered several other topics such as image retrieval, neural networks, business process modeling, semi-supervised learning and XML databases. In addition, by doing analysis of authors’ home pages and DBLP [10], we have 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 authentication.

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/

Table 1. Illustration of 3 topics with related authors. Tiles are assigned to clustered words manually.

Topic 18 “Pattern Mining”		Topic 8 “Information Retrieval”		Topic 74 “Semantic Web”	
Word	**Prob.**	**Word**	**Prob.**	**Word**	**Prob.**
mining	0.242013	retrieval	0.160582	semantic	0.260961
patterns	0.101704	text	0.116067	web	0.138429
pattern	0.067878	document	0.074017	ontology	0.124851
frequent	0.047791	extraction	0.050939	ontologies	0.060605
privacy	0.035545	documents	0.045187	owl	0.033670
preserving	0.034873	information	0.031594	rdf	0.026936
discovery	0.034201	relevance	0.029353	annotation	0.018105
discovering	0.025315	categorization	0.026665	semantics	0.016670
databases	0.022626	topic	0.025694	approach	0.014352
sequential	0.020984	feedback	0.025619	knowledge	0.012365
Author	**Prob.**	**Author**	**Prob.**	**Author**	**Prob.**
Jian Pei	0.011381	Wei-Ying Ma	0.008950	Carole A Goble	0.018056
Jiawei Han	0.010317	ChengXiang Zhai	0.005378	Robert Stevens	0.014153
Wei Wang	0.007429	W. Bruce Croft	0.004808	Peter Haase	0.013177
Philip S. Yu	0.006061	Charles L. A. Clarke	0.003439	Amit P Sheth	0.012201
Hui Xiong	0.005263	Mounia Lalmas	0.003249	Steffen Staab	0.011714
Ada Wai-Chee Fu	0.005111	Weiguo Fan	0.003135	Phillip W Lord	0.011714
Srinivasan Parth.	0.004237	Tat-Seng Chua	0.002679	Luc Moreau	0.010738
Ke Wang	0.004161	David A. Grossman	0.002527	Anupam Joshi	0.010250
Kotagiri Rama.	0.003781	Xuanhui Wang	0.002527	Ian Horrocks	0.009762
Jianyong Wang	0.003667	James Allan	0.002451	David DeRoure	0.009762

3.5.1 Perplexity based Comparison. Perplexity is a standard measure for estimating the performance of probabilistic topic models. It shows generalization power of a topic model for the test dataset; with lower perplexity corroborate better performance.

Fig. 3 on the left side presents the perplexity for each method for different values of *Z*. GAT performs better than the subgroup based AT, CIAT and IAT. As the performance difference between the GAT, AT and ACIT is not very clear, we take the under root of perplexity which is shown in right side of Fig.4. It shows the performance difference between methods more clearly and proves the dominance of GAT over baseline methods. Fig. 3 may suggest that exploitation of both types of grouping for author interest finding task results in the better generalization power of model on the unseen dataset.

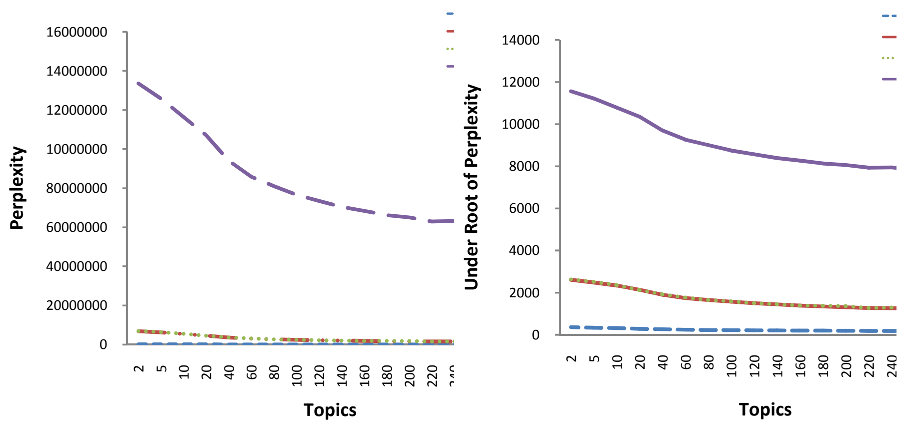

Figure 3. Perplexity for different number of topics.

3.5.2 Entropy based Comparison. Fig. 4 provides a quantitative comparison between proposed GAT, AT, CIAT, and IAT. Fig. 4 (left) shows the average entropy of topic-word distribution of training data for all topics calculated by using Eq. 10. Lower entropy for different number of topics T = 20,40, …300 proves the effectiveness of GAT for obtaining dense topics when compared to baselines. We see when number of topics are less than 40 the performance of GAT, AT and CIAT is same but when the number of topics increases one can see a clear performance difference of GAT with baselines. GAT exploits subgroup and group structures both so able to produce dense topics which results in better performance of method [5].

Fig. 4 (right) shows the average entropy of topic-word distribution of test data for all topics calculated by using Eq. 10. Similar results are obtained for test data except IAT entropy is lower than other methods for number of topics less than 40. We again see that when number of topics are less than 40 the performance of GAT, AT and CIAT is same but when the number of topics increases a clear performance difference of GAT with baselines is observed.

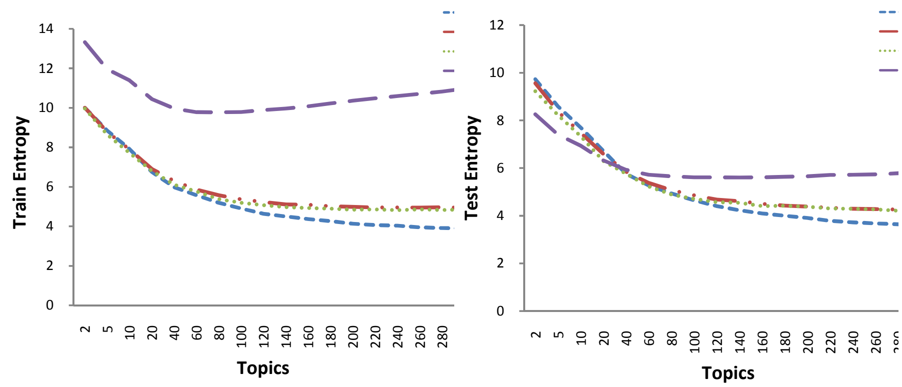

Figure 4. Average Entropy curve as a function of different number of topics for training and test dataset.

3.5.3 Prediction Accuracy based Comparison. We show quantitatively the effectiveness of our proposed method GAT for predicting words and authors of documents in Table 2. GAT performed better when compared with AT, CIAT and IAT for words ranking prediction with values of k= 2,5,10 and for number of topics varied from 2, 5, 10, 20, 40,…,300 shown in Table 1 are 0.56 for GAT, 0.49 for AT, 0.49 for CIAT and 0.43 for IAT. It shows that GAT performed 7% better than AT and CIAT and 13% better than IAT in terms of ranking accuracy which show the better performance of our proposed method. The average ranking accuracy results for author prediction is 0.54 for GAT, 0.44 for AT, 0.45 for CIAT and 0.50 for IAT which show that GAT performed 10% better than AT, 9% better than CIAT, and 4% better than IAT which is significant. Collectively one can say that exploiting subgroup and group level structure together not only increases the generative power of topic model but also helps to have increased ranking accuracy for predicting words and authors.

Table 2. Ranking accuracy for words and authors prediction.

Words Prediction				
Words	**K=2**	**K=5**	**K=10**	**Average**
GAT	0.504506	0.507626	0.689609	0.567247
AT	0.484855	0.426692	0.563054	0.491534
CIAT	0.501782	0.425927	0.556542	0.49475
IAT	0.474909	0.30546	0.511416	0.430595
Authors Prediction				
Authors	**K=2**	**K=5**	**K=10**	**Average**
GAT	0.5	0.433	0.7	0.544333
AT	0.481	0.321	0.526	0.442667
CIAT	0.514841	0.328181	0.527581	0.456867
IAT	0.49961	0.405338	0.619148	0.508032

4 Conclusions

This study deals with the problem of finding author interest. Two types of natural grouping existing in co-author networks is considered in this paper and found to be effective. GAT uses both type of explicit networks and performed better than baselines for several performance measures from different domains. We can say that both explicit grouping structures are important and should be considered simultaneously. We conclude that perplexity and entropy are equally useful for evaluating generative power of topics models with the fact that entropy based results are more understandable with the increasing number of topics. Exploitation of both types of explicit grouping structures also results in increased prediction ranking accuracy for words and authors.

Future work includes the formulization of grouping structure exits in other social networks and structures exploitation for finding their usefulness of social application on the Web by using novel methods.

Acknowledgements
The work is supported by the Higher Education Commission (HEC), Islamabad, Pakistan.

References

[1] Andrieu, C., de Freitas, N., Doucet A., and Jordan, M. I. An introduction to MCMC for Machine Learning. Machine Learning, vol. 50, pp: 5-43, 2003.
[2] Blei, D. M., Ng, A. Y., Jordan, M. I. Latent Dirichlet Allocation. Journal of Machine Learning Research, vol. 3, pp: 993-1022, 2003.
[3] Blei, D. and Jordan, and M. I. Modeling annotated data. In Proceedings of the Annual Conference on Research and Development in Information Retrieval (SIGIR), 2003.
[4] Daud, A., Li, J., Zhu, L., and Muhammad, F. Temporal Expert Finding through Generalized Time Topic Modeling. Knowledge-Based Systems (KBS), Vol. 23(6), pp: 615-625, July 2010.
[5] Daud, A., Li, J., Zhu, L., and Muhammad, F. Conference Mining via Generalized Topic Modeling. In Proceedings of 19th International European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases (ECML PKDD), W. Buntine et al. (Eds.): Part I, LNAI 5781, pp: 244–259, 2009.
[6] DBLP Bibliography Database. http://www.informatik.uni-trier.de/~ley/db/.
[7] Diederich, J., Kindermann, J., Leopold, E., and Paass, G. Authorship Attribution with Support Vector Machines. Applied Intelligence, vol. 19 (1), 2003.
[8] Griffiths, T. L., Steyvers, M. Finding scientific topics. In Proceedings of the National Academy of Sciences, pp: 5228-5235, 2004.
[9] Gray, A., Sallis, P., and MacDonell, S. Softwareforensics: Extending Authorship Analysis Techniques to Computer Programs. In Proceedings of the 3rd IAFL, Durham NC, 1997.
[10] Hofmann, T. Probabilistic Latent Semantic Analysis. In Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI), Stockholm, Sweden, July 30-August 1, 1999.
[11] Mimno, D. and McCallum, A. Expertise modeling for matching papers with reviewers. In Proceedings of KDD, pp: 500–509, 2007.

[12] Mutschke, P. Mining Networks and Central Entities in Digital Libraries: A Graph Theoretic Approach Applied to Co-author Networks. Intelligent Data Analysis, pp: 155-166, 2003.
[13] Newman, M. E. J. Scientific collaboration networks: I. Network construction and fundamental results, Physical Review E, 64, 016131, 2001.
[14] Kawamae, N. Author Interest Topic Model. In Proceedings of SIGIR, July 19–23, pp: 887-888, 2010.
[15] Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., and Steyvers, M. Learning Author-Topic Models from Text Corpora. ACM Transactions on Information Systems, pp. 1-38, March 2009.
[16] White, S., and Smyth, P. Algorithms for Estimating Relative Importance in Networks. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp: 266-275, 2003.