Learning Influence Propagation of Personal Blogs with Content and Network Analyses

Il-Chul Moon
Dept. of Electrical Engineering
KAIST
Daejeon, Republic of Korea
icmoon@smslab.kaist.ac.kr

Dongwoo Kim, Yohan Jo, Alice H. Oh
Dept. of Computer Science
KAIST
Daejeon, Republic of Korea
{dw.kim, yohan.jo, alice.oh}@kaist.ac.kr

Abstract— **Weblogs (blogs) serve as a gateway to a large blog reader population, so blog authors can potentially influence a large reader population by expressing their thoughts and expertise in their blog posts. An important and complex problem, then, is figuring out why and how influence propagates through the blogosphere. While a number of previous research has looked at the network characteristics of blogs to analyze influence propagation through the blogspace, we hypothesize that a blog's influence depends on its contents as well as its network positions. Thus, in this paper, we explore two different influence propagation metrics showing different influence characteristics: Digg score and comment counts. Then, we present the results of our experiments to predict the level of influence propagation of a blog by applying machine learning algorithms to its contents and network positions. We observed over 70,000 blog posts, pruned from over 20,000,000 posts, and we found that the prediction accuracy using the content and the network features simultaneously shows the best F-score in various measures. We expect that this research result will contribute to understanding the problem of influence propagation through the blogosphere, and to developing applications for recommending influential blogs to social web users.**

Keywords- Personal Blog, Blog Influence Prediction, Network Text Analysis

I. INTRODUCTION

Blogs, a type of social web media, have become a major communication medium for the general public as well as governments and corporations. There is an increasing number of personal blogs in Blogspace, Blogspot, LiveJournal, and others. Also, the New York Times, Microsoft and even the U.S. Army maintain their own blogs to communicate with the public. Whether a blog is owned by an individual or an organization, a blog is often used to share the author's information, expertise and opinion through the blog posts. Then, our question to follow is how far the shared information, expertise and opinion can propagate through the blogosphere to influence other blogs and readers [11]. We further ask whether we can predict a blog's potential influence by using machine learning algorithms with only limited knowledge about the blog's content and network.

Unlike most of the previous studies [2, 3] that have approached this question of influence either by looking at just the network properties or just the contents of blogs, we approach this question by looking at both the content and the network properties of the blogs. Specifically, we apply a machine learning model to the blog influence propagation prediction problem.

Conceptually, two reasons motivated us to predict the blog influence, particularly personal blogs, by content and network analyses. Firstly, the content in the blogs often contain more personal and informal stories, and the influence of the personal blogs tends to follow different propagation patterns compared to that of the traditional Web contents. For instance, previous works reveal that blogs containing personal stories are more easily accepted by readers than the traditional media stories [6]. However, little has been known about how much the contents of the personal blogs contribute to the influence propagation compared to the network positions of the blogs. This point is especially true when we consider the personal blogs compared to the blogs from a well-known organizations, such as news outlets, governments, and corporations. Personal blog's influence is much harder to predict because its network position alone is not as distinct as that of blogs of large organizations. Hence, we need additional inputs to see the influence, and the additional information may come from the content analysis.

Second, blogs are often networked with other blogs by blog readers, not by content producers as in the traditional Web contents. The social Web enables the content consumers to create such hyperlinks to form document networks, which is a very distributed and emergent network construction, not authoritative and planned. This new paradigm of network emergence yields different influence propagation compared to the old mass media style influence. Therefore, we need to comprehend the impact of the consumer-created document network to the influence propagation.

The goal of the research presented in this paper is to understand influence propagation in the blogosphere. Particularly, we aim to predict a personal blog's potential influence level based on the blog's content and network position. The value of learning comes from predicting a global influence measure based on a blog's local features. This learning model may reveal the importance of the combined analyses of network and contents. Also, we setup multiple influence measures, so we may find out which features are important in predicting which influence measure.

978-0-7695-4211-9/10 $26.00 © 2010 IEEE
DOI 10.1109/SocialCom.2010.103

We hypothesized that this learning is feasible because we assumed that both the content and the network properties of blogs are key factors in determining a blog's potential to influence others. Therefore, we identified the topics and the network positions of blogs and used them as features to feed into a machine learning algorithm for predicting influence propagation. After running this machine learning experiment with over 70,000 blog posts, reduced from over 20,000,000 posts, the results show that this topic and network combination nearly always improves the accuracy and the precision of identifying the highly influential blogs, and the combination generally improves the F1 score of the highly influential blogs. The contribution of the results can be summarized by the bullet points below.

- This work demonstrates the importance of inter-disciplinary work, particularly between the content and the network analyses, in the social web study. We support this importance by observing the enhanced results in the blog influence prediction.
- This work illustrates the effects of using multiple machine learning algorithms through a research process, in our case, the author-topic model, a variant of the LDA model, for identifying topic distribution, and the SVM multi-class classifier using the topic distribution as input features.

II. Previous Research

While the content-only [3, 11, 12] and the network-only [1, 2] analyses on blogs are popular, the hybrid of the two analyses has emerged only very recently. Yoo et al [4] merged the content and the network analyses to estimate the importance of email messages. They computed the bag-of-words features for each email, and they calculated the network metrics and clusters for email senders and receivers. After that, they merged the two numeric vectors as a feature set and ran a support vector machine (SVM) to predict the priority and the importance. Their results showed the combination of the content and network features led to smaller errors than using either content or network analysis alone. This is a successful example of utilizing two aspects to solve a social web media problem. Our investigation follows this approach in a very similar way. However, we use a more sophisticated version of the content analysis by using LDA topic analysis, rather than creating a simple bag-of-words. Furthermore, we investigated the feature importance by examining the feature weights from a trained SVM.

Also, researchers tried to build a generative model for a document and its hyperlinks, i.e. Nallpati et al. [10]. This model was designed to address two problems simultaneously: (1) discovering topics and (2) modeling topic specific influence of blogs. The authors used a combination of pLSA [13] and Link LDA [14], in which pLSA represents the linked document and Link LDA the linking document. In their experiments, the log-likelihood and link prediction showed higher performance than the pure Link LDA. While this work is intended to predict a future link by considering topics, ours is a measure prediction in the overall network.

Another important work is designing an influence measure by using blog network and contents. For example, Agarwal et al. [8] identified influential bloggers by defining influential posts in four different network aspects: (1) inlinks, (2) outlinks, (3) the number of comments, and (4) the length of the post. A post's influence is defined basically as a weighted sum of the first three network features, scaled by a weight proportional to the fourth feature. The authors computed the influence scores of bloggers from a community blog (The Unofficial Apple Weblog), and compared the scores with data collected from Digg, which is a social networking site where people bookmark and tag interesting websites. This past work is utilizing very simple features such as degree counts, comment counts and post lengths to come up with a influence measure. We develop the measure with more sophisticated topic and network analyses and compare the measure results to Digg score.

III. Dataset of Personal Blogs

We used the TREC Blogs08[1] data collection from NIST. This collection is compiled by the University of Glasgow to be used as a standard test collection for research. Table 1 shows the organization of Blogs08 collection. It consists of over 20 million blog posts, and about half of them are collected from Blogspot.com and Livejournal.com. We made our analysis dataset using posts from Blogspot.com and Livejournal.com in 2008 because we chose to limit our focus on personal blogs. To exclude either too short posts or spam blogs, which can produce unexpected biases in our analysis, we filtered the posts in Blogspot.com and Livejournal.com by applying the conditions below.

1) A post should contain more than 50 words.
2) A post should provide meaningful strings from the body of the HTML.
3) A post should have at least one reference to another post in the dataset.
4) A post should be written between 2008-01-01 and 2008-12-31
5) A post should be written in English.

Pruning the posts that do not meet the above criteria, we were left with 72,143 posts from the collection. The descriptive statistics of the filtered dataset are listed in Table 2 and visualized in Figure 1. After selecting the posts of interest, we applied the Porter Stemmer[2], and we removed the ten most frequently occurring words. This was done because the top ten words had too high frequencies compared to the rest of the words, and the LDA variant model would form a meaningless topic composed of those high frequency words.

IV. Method

This paper aims to train a machine learning algorithm with the content and the network features to predict the various types of influence levels of blogs. To perform this blog influence level analysis with machine learning, four

[1]Open to public. More information is posted on a web page: http://trec.nist.gov/data.html
[2] http://tartarus.org/~martin/PorterStemmer

TABLE I. ORGANIZATION OF TREC BLOGS08 DATASET

All posts	Posts from BlogSpot	Posts from LiveJournal	Posts from other sites
28,517,411 (100%)	10,232,739 (36%)	5,933,478 (21%)	12,351,194 (43%)

TABLE II. DESCRIPTIVE STATISTICS OF FILTERED DATASET

	Number of Posts	Number of Unique Words	Number of Blogs of interest	Average # of Words per Post
Refined	72,143	53,257	4,165	225.24

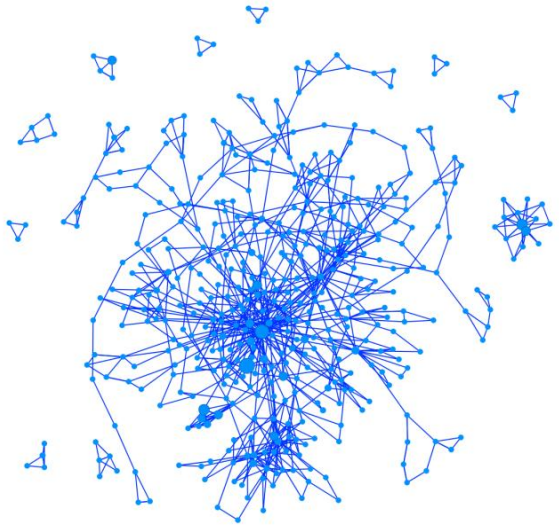

Figure 1. A visualization of the analyzed blog network, we recursively removed the isolate and the pendant blogs from the core to show the core topology in the visualization.

components must be integrated: 1) definition of influence metric, 2) content analysis for machine learning features, 3) network analysis for machine learning features, and 4) a machine learning algorithm. Below we describe the four components in detail.

A. Measuring Blog Influence Propagation

To perform the blog influence prediction study, we must first define what we mean by blog influence propagation and, more importantly, how to measure it. Conceptually, we define that a blog has an influence on its reader if it affects the reader's thoughts and invokes the reader to act for or against the blog author's opinion. For example, a blog reader may write comments on a blog where the reader felt necessary to show his opinion to the author. This shows the blog's influence to invoke the reader's reaction. Another example is a blog ranking system, such as Digg.com, where the readers recommend an influential blog to other bloggers.

1) Comment Count

We selected the comment counts of a blog as our first influence indication. Bloggers tend to write a comment when they want to show either agreement or disagreement, and moreover, bloggers write their additional information about the post. These are all indications that the blog drew the reader's attention to the content, and we view drawing blogger's attention is one indication of the influence of the blog.

2) Digg Score

Our second influence indication measure is the Digg score [9]. The Digg score is already utilized in the blog influence analysis in [8], and one of the reason that we chose this measure is the comparison of our results to the previous research. The Digg score is calculated by accumulating the responses from the Digg users. The responses are mainly promoting or demoting the influence score of a particular websites including blogs. The Digg is a third-party tool that is independent to Blogspot.com and Livejournal.com, so the responses will be more objective than the blog system dependent influence scores, such as Comment Count. It should be noted that we specifically calculated Digg scores in the year of 2008, when the blogs of interests are gathered.

B. Content Analysis of Blogs

We use the author-topic model to perform content analysis, which is an LDA based generative model for documents. Specifically, the author-topic model is used to analyze the related topics of each blog and the related words of each topic. This is possible when we map the authors in the author-topic model to the blogs; and the documents to the blog posts

The author-topic model is constructed in the following way. A post is a mixture of topics, and each topic has a probability distribution over words. Each blog has a topic distribution, which shows how much the blog is related to each topic. A blog generates a post in the following way. First, the blog, which is equivalent to the author (blogger) due to our assumption, is determined to write one word. Second, the topic of the word is chosen from the blog's topic distribution. Third, a real word is sampled according to the word distribution of the chosen topic. These three steps are repeated until this document is completed. These steps can be summarized to estimating the conditional probability distributions below.

$\mathbf{W}$: Word set, w: A word

$$\boldsymbol{\theta} = \theta_{TB} = p(\mathbf{T}|\mathbf{B}) = (\text{The prob. of topics in a blog})$$
$$= \phi_{wT} = p(\mathbf{W}|\mathbf{T}) = (\text{The prob. of words of a topic})$$

There are many techniques for inferring the conditional probability values, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. In this paper, we use Gibbs sampling. For each step, we assign a topic, T_j and a blog B_k to each word i according to

$$p(T_i = j, B_i = k | w_i = m, \mathbf{z}_{-i}, \mathbf{w}_{-i}, B_d) \propto \frac{C^{WT}_{mj} + \beta}{\sum_{m'} C^{WT}_{m'j} + V\beta} \frac{C^{AT}_{kj} + \alpha}{\sum_{j'} C^{AT}_{kj'} + T\alpha}$$

where m is a word index in the word set, $\mathbf{z}_{-i}$ represents the topic assignments to all words except word i, $\mathbf{w}_{-i}$ indicates that word i has no assignment, B_d is the author list, C^{WT} counts topic assignments to words except word i, C^{AT} counts topic

assignments to blogs except word$i$, V is the vocabulary size, T is the number of topics, and α and β are smoothing factors. In this way, we can obtain a blog's topic distribution gracefully by applying the author-topic model.

C. Network Analysis of Blogs

We construct the blog network by the below definition of nodes and links. From the network perspective, B_i is a blog network node, and the $InfluenceLink_{i,j}$ is a link of the blog network. We will come up with a set of machine learning features through the blog network analysis. The most common network analysis technique is applying network metrics to a constructed network. Each network metric has its own interpretations, and these specify the network position characters of the node in the network. We will use five most frequently used network metrics described below.

B: Blog set, B_i: A blog in the blog set

$InfluenceLink_{\mathrm{j,k}}$

$=$ (Discovered hyperlinks from a post in B_j to a post in B_k)

1) In-Degree, Out-Degree and Total-Degree Centrality

In the blog analysis, In-Degree Centrality means the number of blogs referencing a measured blog, and Out-Degree Centrality measures the number of blogs that a measured blog references. Total-Degree Centrality is the sum of the In-Degree and the Out-Degree Centralities. This type of metric is a local-metric which only concerns the direct neighbors of a specific blog. Hence, these metrics cannot measure the position characteristics from the global network perspective. We included these metrics because they are perhaps the simplest as well as the most effective measures to find influential posts.

$$|\mathbf{B}|: (The\ number\ of\ blogs\ in\ the\ set)$$

$$Indegree(B_i) = \frac{1}{|\mathbf{B}|}\sum_{j=1}^{|\mathbf{B}|} InfluenceLink_{ji}$$

$$Outdegree(B_i) = \frac{1}{|\mathbf{B}|}\sum_{j=1}^{|\mathbf{B}|} InfluenceLink_{ij}$$

$$Totaldegree(B_i) = \frac{1}{|\mathbf{B}|}\sum_{j=1}^{|\mathbf{B}|} \frac{InfluenceLink_{ji} + InfluenceLink_{ij}}{2}$$

2) Betweenness Centrality

Betweenness Centrality measures how often a blog is positioned on the shortest paths between any blog pair on the blog network. Because this measure deals with the shortest pathsin the network, this is a global metric unlike the degree centralities. Often, the shortest paths are regarded as the influence propagation route, and that is why we included this metric.

$$Betweenness(B_i) = \frac{1}{(n-1)(n-2)}\sum_{j=1, j\neq i}^{|\mathbf{B}|}\sum_{k=1, k\neq j, k\neq i}^{|\mathbf{B}|} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$$

σ_{jk} : $(number\ of\ shortest\ paths\ from\ B_j\ to B_k)$

$\sigma_{jk}(i)$: $(number\ of\ shortest\ paths\ from\ B_j\ to\ B_k\ through\ B_i)$

3) Clustering Coefficient

Clustering Coefficient calculates the connectivity among the direct neighbors of a specific blog. This is a local metric as the degree centralities, but clustering coefficient has a distinct aspect by observing the neighbor connections excluding the measured blog itself. This represents a blog's effect to its neighbors' organization. We included this metric because we hypothesized that the influence propagates farther if a blog is able to facilitate the links among its neighbors.

$$ClusteringCoefficient(B_i) = \frac{1}{Z}\sum_{i\in Neighbor(B_i)}\sum_{j\in Neighbor(B_i), j\neq i} InfluenceLink_{ij}$$

The five metrics above are calculated for each blog, and we gathered the five values as a numeric vector. We then utilize this vector as a machine learning feature from the network analysis side.

D. Machine Learning for Influence Level Prediction

The two analysis methods described above, content analysis and network analysis, are often applied separately in the social Web problems. An important motivation of this paper is observing the effect of combining the content and the network analyses. To combine the two analysis approaches, we identify the similarities of analysis results from the two approaches. One similarity is that the two results are often in the numeric vector formats. The result from the content analysis is a topic distribution over blogs. Similarly, the result from the network analysis is a vector of network metrics for blogs. Hence, we found that combining two numeric vectors to explain the quantified blog influence can be one way to fuse the content and the network analyses.

As a fusion of two analyses, we use machine learning algorithms with the combined vector feature sets. Specifically, we train a machine learning algorithm to predict the blog influence level with three different feature sets.

- Topic Model : Use the topic distribution probabilities from the author-topic model (50 features)
- Network Model : Use network metrics (5 features)
- Topic-Network Model : Use both the topic distributions as well as network metrics (55 features)

We particularly designed these three feature sets to experiment with the impact of having either content analysis alone, network analysis alone, or content and network analyses together. We used a linear SVM [3] for the influence level prediction. To evaluate the trained SVM, we used four evaluation metrics: accuracy, precision, recall, and F-measure.

[3] We used an open source machine learning package, WEKA. More information can be found from the below URL. http://www.cs.waikato.ac.nz/ml/weka/

TABLE III. DETAILED SVM CLASSIFIER PERFORMANCE OF THE TOPIC, THE NETWORK, AND THE TOPIC-NETWORK MODELS. THE SHADED CELLS CONTAIN THE BEST RESULT AMONG THE THREE MODELS IN A SPECIFIC MACHINE LEARNING PERFORMANCE FOR A CERTAIN INFLUENTIAL MEASURE. THE PRESENTED PERFORMANCES ARE RESULTED IN BY USING THE OPTIMAL WEIGHT FOR EACH MODEL. 30% OF BLOGS ARE USED FOR TRAINING, AND THE SAMPLINGS ARE DONE BY 30 TIMES.

	Comment Count				Digg Score			
Model Name	Accur.	Prec.	Recall	F1-Measure	Accur.	Prec.	Recall	F1-Measure
Topic	0.631	0.172	0.691	0.275	0.688	0.197	0.583	0.295
Network	0.857	0.307	0.309	0.308	0.851	0.283	0.207	0.239
Topic-Network	0.716	0.213	0.664	0.322	0.714	0.211	0.571	0.308

V. INFLUENCE PROPAGATION PREDICTION

In this section, we describe the machine learning analysis to predict the influence propagation of the blogs by the topic, the network, and the network-topic models. Before describing the results of the analysis, we explain how we handled the skewed distribution of the influential blogs, as there are 10% of influential blogs out of 4165 blogs. We used boosting, which, fundamentally, is used to improve accuracy by over-sampling the influential training instances that are critical in setting the decision boundary of the learner [5]. Our technical method to implement boosting is giving more weights to the influential blog training instances, so that the weights can function as the over-sampling factor of the training instances [7]. Without boosting, because of the skewed distribution, the recall rates for the upper tier blogs will be very low, as the learner will put the decision boundary to classify most of the blogs as belonging to the un-influential blog. That way, the false-negative error, or Type II error, will be minimized due to the small number of the influential blogs. Once we put more weight on the influential blog training instances, we make the learner to relax the classification boundary to include more instances in the true classification boundary.

With this compensation for the skewed distribution, we ran the machine learning experiments using the three models. Figure 4 shows the precision-recall curve, the accuracy, the precision, the recall, and the F_1 measure changes by the weight factor increases. We sampled 30% of the entire instances for training and the rest for testing. This sampling was done thirty times, and we report the averages. We initially varied the size of training set ranging from 10% to 50%, yet 30% was optimal with fewer under- and the over-fitting problems.

Through this machine learning experiment, we found three major results. First, the different influence measures are better predicted by using different input feature sets. For instance, considering the F1-measure, the influential blogs from the comment count perspective are better classified by the order of the Topic-Network model, the Network model, and the Topic model. On the other hand, the influential blogs from the Digg score are more accurately predicted by the order of the Topic-Network Model, the Topic Model, and the Network model. This means that the comment counts are better explained by the network features and the Diggs score are by the topic features. At the same time, the predictions are better off when they have both feature information.

The second finding is the good recall performance of the Topic model in identifying the influential blogs from the comment count and the Digg score viewpoints. Recently, the influential blogs have been often identified by exploring the network structure alone. However, for a complete identification of influential blogs, we may have to look at the contents, as well.

The third finding is the better precision performance of the Network model. When a system needs to suggest a few influential blogs, which means that the system does not need to recover all the influential blogs, observing the network position alone may work better, which we suspect that that is the reason why many blog recommendation systems use the network position information.

VI. CONCLUSION

The social Web has become one of the most important trends of the Internet. Particularly, social media on the Internet, e.g., blogs, are now very prolific tools to communicate one's thoughts to others. Then, it is an interesting problem to understand the characteristics of blogs that can influence more people. This understanding will impact many applications such as blog searches, blog recommendations, and opinion mining. We approach this blog influence problem by analyzing blog influence levels with machine learning. We hypothesized that a blog's influence propagation is driven by two major factors: its content and network. Our machine learning experiments demonstrate that analyzing only the network or the content alone cannot provide the whole picture. Our combined analysis of both content and network proved to be the best way to learn the blog influence propagation. This blog analysis is the first work supporting a previous work of classification of importance email with the combination of content and network analyses. This time, we demonstrated a similar result in the blog analysis. Now, we conjecture that social web contents, such as email and blog, should be analyzed in both aspects to make the analysis complete. We expect that our findings will facilitate better social Web analysis by using machine learning with both content and network analysis.

ACKNOWLEDGMENT

This work was supported by Brain Korea 21 Project, the School of Information Technology, KAIST in 2010.

REFERENCES

[1] Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999) On power-law relationships of the internet topology, Computer Communications Review, Vol. 29, pp. 251-262

[2] Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N.,and Hurst, M. (2007) Cascading Behavior in Large Blog Graphs, Proceedings of SIAM International Conference on Data Mining (SDM), Minneapolis, MN, USA

[3] Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A. (2004) Information diffusion through blogspace, Proceedings of the 13th international conference on World Wide Web, New York, NY, USA, pp. 491-501

[4] Yoo, S., Yang, Y., Lin, F., and Moon, I. (2009) Mining social networks for personalized email prioritization, Proceedings of the Knowledge Discovery and Datamining 2009, Paris, France, pp. 967-976

[5] Ratsch, G., Onoda, T., and Muller, K.-R. (2001) Soft Margins for AdaBoost, Machine Learning, Vol. 42, Num. 3, pp. 287-320

[6] Johnson, T., and Kaye, B. (2004) Wag the blog: How reliance on traditional media and the Internet influence credibility perceptions of Weblogs among blog users, Journalism & Mass Communication Quarterly, Vol. 81, No. 3, pp. 622-642

[7] Batista, G., Prati, R. C., and Monard, M. C. (2004) A study of the behavior of several methods for balancing machine learning training data, ACM SIGKDD Explorations, Vol. 6 ,Iss. 1, Special issue on learning from imbalanced datasets, pp. 20 – 29

[8] Agarwal, N., Liu, H., Tang, L., and Yu, P. S. (2008) Identifying the influential bloggers in a community, Proceedings of the international conference on Web search and web data mining, February 11-12, Palo Alto, California, USA

[9] Digg (2010, retrieved) http://www.Digg.com

[10] Nallapati, R. and Cohen, W. (2008) Link-plsa-lda: A new unsupervised model for topics and influence of blogs, in Proceedings of the International Conference on Weblogs and Social Media (ICWSM). Association for the Advancement of Artificial Intelligence, 2008.

[11] Agarwal, N. and Liu, H. (2008) Blogosphere: Research Issues, Tools, and Applications. SIGKDD Explorations, 10(1): 18 - 31, July

[12] Yano, T., Cohen, W. W., and Smith, N. A. (2009) Predicting response to political blog posts with topic models. In Proc. Of NAACL-HLT.

[13] Hofmann, T. (1999) Probabilistic latent semantic indexing, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, p.50-57, August 15-19, Berkeley, California, United States

[14] Erosheva, E., Fienberg, S., and Lafferty, J. (2004) Mixed membership models of scientific publications. Proceedings of the National Academy of Sciences, 101(Suppl. 1)