Detecting Research Topics via the Correlation between Graphs and Texts

Yookyung Jo
Department of Computer Science, Cornell University
Ithaca, NY, 14850
ykjo@cs.cornell.edu

Carl Lagoze
Computing and Information Science, Cornell University
Ithaca, NY, 14850
lagoze@cs.cornell.edu

C. Lee Giles
Information Sciences and Technology, The Pennsylvania State University
University Park, PA
giles@ist.psu.edu

ABSTRACT

In this paper we address the problem of detecting topics in large-scale linked document collections. Recently, topic detection has become a very active area of research due to its utility for information navigation, trend analysis, and high-level description of data. We present a unique approach that uses the correlation between the distribution of a term that represents a topic and the link distribution in the citation graph where the nodes are limited to the documents containing the term. This tight coupling between term and graph analysis is distinguished from other approaches such as those that focus on language models. We develop a topic score measure for each term, using the likelihood ratio of binary hypotheses based on a probabilistic description of graph connectivity. Our approach is based on the intuition that if a term is relevant to a topic, the documents containing the term have denser connectivity than a random selection of documents. We extend our algorithm to detect a topic represented by a set of terms, using the intuition that if the co-occurrence of terms represents a new topic, the citation pattern should exhibit the synergistic effect. We test our algorithm on two electronic research literature collections, arXiv and Citeseer. Our evaluation shows that the approach is effective and reveals some novel aspects of topic detection.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]

General Terms

Algorithms, Languages, Measurement

Keywords

topic detection, graph mining, probabilistic measure, citation graphs, correlation of text and links

KDD'07, August 12–15, 2007, San Jose, California, USA.

1. INTRODUCTION

The availability of large-scale linked document collections such as the Web and specialized research literature archives[6, 3] presents new opportunities to mine deep knowledge about the community activities behind the document collections. Topic discovery is one example of such knowledge mining that has recently attracted considerable research interest [12, 14, 9, 22, 23, 24, 17, 4, 16, 7, 15]. Topics are semantic units that can function as basic building blocks of knowledge discovery. Once discovered they can be used in a number of ways including information navigation, trend analysis, and high-level descrption of data [22, 15].

In this paper, we present a unique approach to topic detection that uses the correlation between the distribution of terms representing a topic and the distribution of links in the citation graph among the documents containing these terms. This distinguishes our work from other approaches to topic detection that focus on textual data alone [9, 22, 23, 14] or which detect topics and communities by studying graph properties without considering text features [10, 19, 11, 20]. Our appoach is based on the intuition that documents related to a topic should be more densely connected in the citation graph than a random selection of documents are connected in the citation graph. We therefore extract topics from the corpus by examining the structure of the *term citation graph* for each term in the corpus. A term citation graph of a term A is a subgraph of the full citation graph restricted to the documents that contain the term A and the edges between these term-specific nodes. If the term citation graph of a term A shows denser connectivity than a random subgraph of the full citation graph, it is likely that the term A represents a topic.

An illustration of our approach to topic detection is as follows. Let's imagine that we have a set of all documents containing a term α: for example "sensor network" or "association rule mining". Intuitively, if α represents a topic, then the documents containing this term will be interconnected in a relatively dense citation network (Figure 1. a)). This contrasts with another term η, for example "practical examples" or "six months", that are non-topic terms (i.e., general terms) for which the citation links among containing documents will be relatively sparse (Figure 1. b)). The notions of "dense" and "sparse" connectivity are relative to the connectivity of a citation graph consisting of a random selection of documents and their citation edges from the full citation graph.

We develop topic score measures that are log odds ratios

of binary hypotheses based on a probabilistic description of graph connectivity. For each term in the corpus, we take a look at its term citation graph. Our topic score measure tells, with statistical confidence, whether the connectivity of the term citation graph is significantly denser than what is expected from the citation graph of a random selection of documents. As a first approximation, we assume that a topic can be represented by a single term. We then extend our algorithm to detect topics that are not represented by a single term, but by the relation of a set of terms.

We test our algorithms on two electronic research literature collections, arXiv and Citeseer. Our experiments produce a ranked list of terms that on examination by field experts and based on our observations match prevailing topics in the corpus. Our evaluation of the lists uncovers a number of interesting characteristics of the lists of terms, including the discovery of topics in varying scale, the prevalence and specificity of topics, and the time evolution of topics.

This paper is structured as follows. Section 2 and Section 3 present our algorithm to detect topics represented by a single term and by a set of terms, respectively. Section 4 shows the results obtained by applying the algorithm to arXiv and to Citeseer. Section 5 reviews the related work. Section 6 discusses a few issues raised in the work. Section 7 concludes.

2. DETECTING TOPICS IN LINKED TEXTUAL CORPUS

The problem statement of this paper is "How do we detect topics in a linked textual corpus, such as a collection of research papers?". We address this research problem by producing a ranked list of terms where terms are ordered according to how likely a term represents a topic and how significant the topic represented by a term is. To accomplish this goal, we look at the citation graph of the corpus at the resolution of an individual term level.

Definition 1. In this paper, a "term" is defined as an n-gram phrase that consists of any n consecutive words from a document, where n is any positive integer. For example, "network", "for the", "association rule mining" are all valid examples of a term.

Conventionally, the citation graph of a corpus is a directed graph with nodes being the documents or research papers in the corpus, and with edges being the hyperlinks or the citation links. In this paper, we only consider the undirected version of the citation graph. We denote the undirected citation graph of the entire corpus as G_{all}.

The term citation graph of a term A, G_A, refers to a subgraph of the entire citation graph G_{all} with nodes restricted to the documents that contain the term A and the links between these documents. Precisely,

Definition 2. G_A, the term citation graph of a term A, is defined by

$$V(G_A) = \{d | \text{document } d \text{ contains a term } A, d \in V(G_{all})\}$$
$$E(G_A) = \{e(d_i, d_j) | d_i, d_j \in V(G_A), e(d_i, d_j) \in E(G_{all})\}$$

where $V(G)$ denotes the set of vertices in G, $E(G)$ denotes the set of edges in G, and $e(d_i, d_j)$ is an edge between the nodes d_i and d_j.

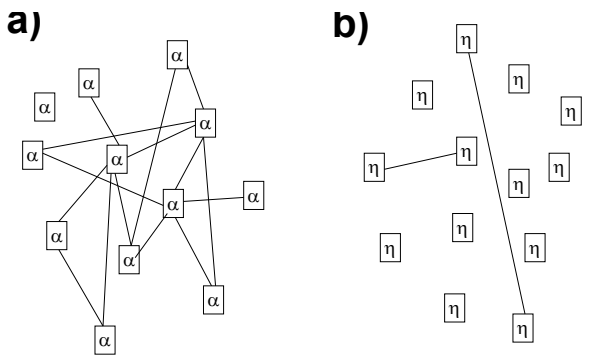

Figure 1: The term citation graphs. a) α : a term representing a topic. (e.g. "sensor network", "association rule"), b) η : a term not representing a topic. (e.g. "practical examples", "six months")

Given a term, we want to make a binary decision with statistical confidence about whether the term is relevant to a topic or not. We use the following intuition. If a term represents a topic, then the document nodes in its term citation graph will be well-connected by citations. On the other hand, if a term does not represent a topic, the documents in its term citation graph are not related to each other, thus their distribution is random with respect to citation patterns. Figure 1 shows this intuition. Figure 1 a) is the term citation graph for a topic term α showing dense connectivity. Figure 1 b) is the term citation graph of a non-topic term η showing sparse connectivity comparable to that of a random selection of documents.

We formalize this notion by setting up two hypotheses. Given a term A, hypothesis H1 says that A is relevant to a topic, and hypothesis H0 says A is not. We make an observation $O(G_A)$ about the connectivity of the term citation graph of A, G_A. We compute the loglikelihood of the observation $O(G_A)$ under hypothesis H1 and the loglikelihood of $O(G_A)$ under hypothesis H0. The difference of the two loglikelihoods becomes the topic score for the term A.

$$\begin{aligned}&\text{TopicScore}(A)\\ &= \log(P(O(G_A)|H1)) - \log(P(O(G_A)|H0))\\ &= \log\left(\frac{P(O(G_A)|H1)}{P(O(G_A)|H0)}\right)\end{aligned} \quad (1)$$

The topic score represents how well hypothesis H1 explains the connectivity observation, compared to hypothesis H0.

We take the observation $O(G_A)$ to indicate, for each node in G_A, whether the node has at least one link to the rest of the graph or not. Under hypothesis H1, it is very likely that a node in the graph is connected to the rest of the graph by at least one link. The document either cites or is cited by another document that shares the topic. We use the parameter p_c, with a value close to 1, to denote this probability of a node in G_A having at least one link to any other node in G_A. We present the result with p_c set to 0.9 in Section 4. [1] Then, the loglikelihood of $O(G_A)$ with

[1]Our experiment with several values of p_c shows that the result is not sensitive to a particular choice of values for p_c, as long as the value is close to 1.

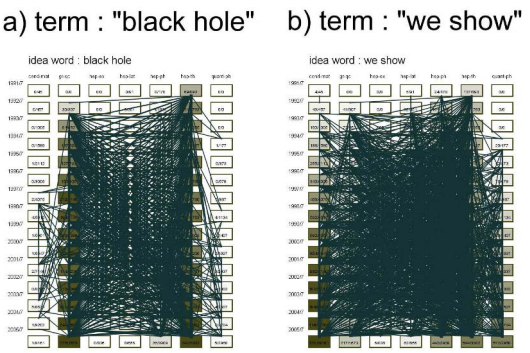

Figure 2: The term citation graphs from arXiv. a) for a topic term "black hole", b) for a stop phrase "we show"

hypothesis H1 is given as follows.

$$\begin{aligned}
&\log\left(P\left(O\left(G_A\right)|H1\right)\right)\\
&=\log\left(\prod_i P\left(O_i\left(G_A\right)|H1\right)\right)\\
&=\sum_i \log\left(P\left(O_i\left(G_A\right)|H1\right)\right)\\
&=n_{c,A}\log\left(p_c\right)+\left(n_A-n_{c,A}\right)\log\left(1-p_c\right) \qquad (2)
\end{aligned}$$

where n_A is the number of nodes in G_A, and $n_{c,A}$ is the number of nodes in G_A that have at least one link that points to another node within G_A, and $O_i\left(G_A\right)$ is the per-node observation for node i.

The loglikelihood of $O\left(G_A\right)$ with hypothesis H0 is more interesting. Under the null hypothesis H0 that a term A is not relevant to a topic, the documents in G_A are not related to each other. Thus, given a node i in G_A and one of its citation links, the probability that the other end of this link points to any node within G_A is $\frac{n_A-1}{N-1}$, where n_A is the number of nodes in G_A and N is the number of nodes in the entire corpus. That is, determining which node a citation link of a node i connects to can be considered as a random process with respect to G_A, where any node in the entire corpus is equally likely to be the destination of the link. Then, the probability that a node i in G_A is connected to any other nodes in G_A by at least one link is given as, $1-\left(1-\frac{n_A-1}{N-1}\right)^{l_i}$, where l_i is the number of all links of a node i.

The loglikelihood of $O\left(G_A\right)$ with hypothesis H0 is given as follows.

$$\begin{aligned}
&\log\left(P\left(O\left(G_A\right)|H0\right)\right)\\
&=\sum_i \log\left(P\left(O_i\left(G_A\right)|H0\right)\right)\\
&=\sum_{i\in V_c(G_A)} \log\left(1-\left(1-\frac{n_A-1}{N-1}\right)^{l_i}\right)\\
&+\sum_{i\in(V(G_A)-V_c(G_A))} l_i\cdot\log\left(1-\frac{n_A-1}{N-1}\right) \qquad (3)
\end{aligned}$$

where $V_c\left(G_A\right)$ denotes the set of nodes in G_A that has at least one link to any other node in G_A.

It should be noted that our null hypothesis H0 is based on the randomness of the citation connectivity, not on the absolute sparseness of the connectivity. This enables our topic score to effectively filter out high-frequency common phrases as non-topic terms. This is illustrated in Figure 2. [2] Figure 2 a) shows the term citation graph derived from arXiv for a prevalent topic term "black hole" and Figure 2 b) for a stop phrase "we show". As shown, it is not easy to discern from the graph visualization that the topic relevance of "black hole" is much greater than that of "we show". However, as will be seen in Section 4, our topic score measure assigns the highest score to "black hole", and the lowest score to "we show". This is because, for the term "we show", the random connectivity assumption of the null hypothesis H0 defaults to the dense connectivity as shown in Figure 2 b), while the hypothesis H1 assumes even denser connnectivity.

If we generate the topic scores in Eq.1 for all possible terms in the corpus and order them, we get a ranked list of terms, where terms are ranked according to how likely they represent the topics of the corpus. The terms at the top ranks are the terms representing the topics prevalent in large scale. This is because the term citation graphs of the topics prevalent in large scale have many instances of per-node observations that support the hypothesis H1 over H0.

As hinted above, the bottommost ranked terms have clear intuitive interpretation as well. These terms are the stop words or common phrases, as their term citation graphs exhibit the large scale statistical evidence that can be better explained by H0 than by H1.

3. DETECTING TOPICS REPRESENTED BY A SET OF TERMS

Some topics are not detectable by a single term but by the appearance of a set of terms. This may occur, for example, when a new term is not coined for a topic, but the topic is represented by the relation between a few general terms. For example, let's consider a topic M represented by the co-occurrence of two terms "quantum computer" and "quantum dot". This is a research topic in physics about using "quantum dot" as a hardware device for "quantum computer". However, each individual term "quantum dot" or "quantum computer" represents a much broader research topic than the given topic M. The term "quantum computer" represents any topic related to quantum computing: examples are quantum computer algorithms, fault tolerant quantum computing, and many kinds of hardware devices for quantum computer. "quantum dot" is a nano-scale semiconductor material. The term "quantum dot" represents a broad research topic including material property study, and using quantum dot to make applications such as laser, quantum computer logic gate, etc. Thus, looking at a single term is not going to reveal the topic M.

The problem of detecting a topic represented by a set of

[2] To aid the visualization, the term citation graphs from arXiv are illustrated in the following ways. The vertical axis is a time scale where time follows downward. The horizontal axis spans 7 research fields of arXiv. A paper at a particular time and a field is placed in the small rectangle at the corresponding position. The darkness of a rectangle represents the number of papers contained in the rectangle. The links between rectangles are the citation links between the papers in the rectangles.

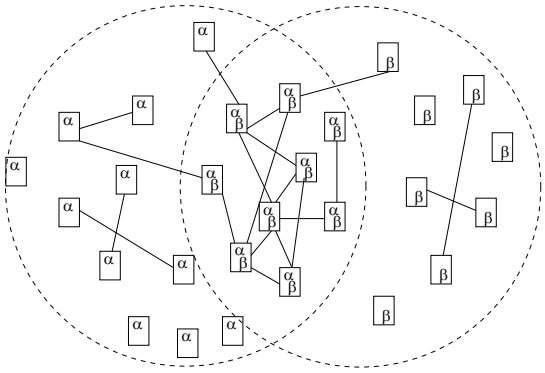

Figure 3: The term citation graphs of terms α and β, and their intersection

terms but not by an individual term is different from finding the co-occurrence counts of terms. The mere high count of co-occurrence is not what we want. The co-occurrence count of two stop words might be high, but it does not carry topic information. Also, the normalized co-occurrence, defined as the co-occurrence count divided by the occurrence count of a single term, is not what we want either. At the extreme, we may think of terms A and B that always occur together with high frequency. But this topic is detectable by looking at a single term A or B by the method explained in Section 2. Finally, it should be noted that our goal is different from association rule mining [1]: the above example of terms A and B co-occurring with high frequency qualifies for an association rule, but not for detecting a topic represented by a set of terms. Then, how do we detect topics represented by a set of terms?

We again look at the term citation graphs and use the following intuition that is illustrated in Figure 3. In Figure 3, a small rectangle containing α is a document containing a term α. Similarly, a small rectangle containing β is a document containing a term β. A small rectangle containing both α and β is a document containing both terms. A link connecting two documents is a citation link. The left big circle encloses the term citation graph of the term α, which is G_α. The right big circle encloses G_β. The documents and links within the intersection of the two circles constitute the citation graph for the documents containing both terms, which we denote as $G_{\alpha\cap\beta}$. Figure 3 shows that the documents containing both terms α and β are significantly more densely connected than G_α or G_β. This indicates that there is a nontrivial topic represented by the co-occurrence of α and β, but not by one of them. On the other hand, if there is no significant topic represented by the marriage of α and β then the occurrence of the term β within G_α or the occurrence of the term α within G_β will not be correlated to the citation pattern. In this case, $G_{\alpha\cap\beta}$ should have the link connectivity comparable to that of the same size random subgraph of G_α or G_β.

We formalize this notion as follows. Given a term A and a term B, we want to detect whether the connectivity of $G_{A\cap B}$ is significantly higher than what we could normally expect from the connectivity of G_A or G_B. To account for the connectivity of any term citation graph G, we use an observation [3] that considers, for each citation link of each node in G, whether the link ends with a node *within* G or *outside* G. If for each link of a node in G the probability that it ends with another node within G is p, then the loglikelihood of the connectivity observation on G is,

$$\ln\left(P\left(O\left(G\right)|p\right)\right) = \sum_i \left(c_i\left(G\right)\ln\left(p\right) + \left(l_i - c_i\left(G\right)\right)\ln\left(1-p\right)\right) \quad (4)$$

where l_i is the total number of citation links of a node i, and $c_i\left(G\right)$ is the number of citation links of a node i that fall within G. Let $p^*\left(G\right)$ be the value of p that maximizes Eq.4. With the number of nodes in G fixed, $p^*\left(G\right)$ tends to increase, as the connectivity of G gets denser.

Let's consider G_A and $G_{A\cap B}$ under the hypothesis that the co-occurrence of terms A and B does not represent a new topic. Under this null hypothesis, the generative process of determining which document in G_A contains the term B is an independent random process with respect to the distribution of the citation links of G_A. Thus, if we let p_{0A} be our guess for $p^*\left(G_{A\cap B}\right)$ under the null hypothesis, our best guess for p_{0A} is the probability that maximizes the average loglikelihood of the following subgraphs of G_A. The subgraphs we consider are any subgraphs of G_A that have the same number of nodes as that of $G_{A\cap B}$, and the citation links between them. Formally, p_{0A} is given as follows.

$$p_{0A} = \arg\max_p \frac{1}{{}_nC_k}\sum_{G_\sigma}\ln\left(P\left(O\left(G_\sigma\right)|p\right)\right) \quad (5)$$

where n is the number of nodes in G_A, k is the number of nodes in $G_{A\cap B}$, and the summation over G_σ runs over any graph G_σ that satisfies the following

$$\begin{aligned} &V\left(G_\sigma\right) \subset V\left(G_A\right), \\ &|V\left(G_\sigma\right)| = |V\left(G_{A\cap B}\right)|, \\ &E\left(G_\sigma\right) = \{e\left(v_i,v_j\right)|e\left(v_i,v_j\right) \in E\left(G_A\right), \\ &\qquad\qquad v_i,v_j \in V\left(G_\sigma\right)\} \end{aligned} \quad (6)$$

p_{0A} can be analytically obtained to be

$$p_{0A} = \frac{\left(k-1\right)\sum_{i\in V(G_A)} c_i\left(G_A\right)}{\left(n-1\right)\sum_{i\in V(G_A)} l_i} \quad (7)$$

Now, we think of the alternative hypothesis that says the co-occurrence of terms A and B represents a new topic. Under this hypothesis, our guess for $p^*\left(G_{A\cap B}\right)$, which we denote as p_{1A}, should be significantly higher than p_{0A}. We set it as $p_{1A} = m \cdot p_{0A}$, where m is a mulplicative parameter greater than 1.

The following score $T_A\left(A,B\right)$ is our confidence about how likely the co-occurrence of terms A and B represents a new topic, with respect to a term A.

$$T_A\left(A,B\right) = \ln\left(\frac{P\left(O\left(G_{A\cap B}\right)|p_{1A}\right)}{P\left(O\left(G_{A\cap B}\right)|p_{0A}\right)}\right) \quad (8)$$

Note that our guess for p_{1A} need not be exactly $p^*\left(G_{A\cap B}\right)$ nor even close to it. The actual value of $p^*\left(G_{A\cap B}\right)$ only needs to be relatively closer to p_{1A} than to p_{0A} to make $T_A\left(A,B\right)$ positive. In particular, if the actual $p^*\left(G_{A\cap B}\right)$ is significantly larger than p_{0A}, $T_A\left(A,B\right)$ will be positive for

[3]Note that this observation $O(G)$ is different from the observation $O(G)$ for a single term topic detection in Section 2. The choice is made so that the new observation $O(G)$ can account for graph connectivity in a continuous spectrum.

a wide range of m. Thus, with large m, we could filter false positives, while we may only lose false negatives with weak confidence. We experimented on several values for m in the range of $[2, 10]$. While the result does not sensitively change over a wide range of m, the choice of $m = 6$ seems to provide a good balance between false positives and false negatives. In Section 4, we present the result with $m = 6$.

We then get $T_B(A, B)$ in the similar way by looking at $G_{A\cap B}$ and G_B. Our final score for judging whether the co-occurrence of terms A and B represents a new topic or not is given by taking the minimum of $T_A(A, B)$ and $T_B(A, B)$, reflecting our belief that the link density of $G_{A\cap B}$ should show a significant departure from that of both G_A and G_B.

$$\text{TopicScore}(A, B) = \min(T_A(A, B), T_B(A, B)) \quad (9)$$

4. EVALUATION

We use arXiv and Citeseer for evaluation.
We restrict the terms we consider to all possible bigrams in the corpus. We choose bigram as our term unit, because bigrams typically convey more concrete ideas than unigrams, yet higher grams might suffer from the explosion of the number of terms and sparseness of data for each term. But, it is only a choice of convenience and our algorithm can be applied to any n-grams. We further restrict the terms by pruning out low frequency terms that appear in less than 5 documents in the corpus and by pruning out 35 stop words.

4.1 Evaluation on arXiv Data

arXiv is an actively maintained online repository of research papers in physics. We take papers from year 1991 to year 2006 that span 7 major arXiv areas. This is in total 214,546 papers and 2,165,170 citation links between them, which amounts to 10.09 per-document citations. For each paper, we use its abstract as its document.

We perform the following experiments. First, for all possible terms appearing in the corpus, we compute the single term topic score measure of Eq.1, and get a ranked list of topics. Second, for all possible term pairs in the corpus, we compute the topic score of two terms as in Eq.9, and get a ranked list of topics.

The running time is reasonable. We used a pentium IV PC with 2GB memory. It took 45 minutes to generate the term citation graphs for all terms and their inverted index. It took 4 minutes to compute the single term topic scores for all terms. It took about 10 hours to compute the topic scores of two terms for all possible term pairs.

To consider all pairs of terms for two term topic scores could be prohibitive as there are a huge number of terms. But, we need to consider a pair of terms only when the two terms appear together in at least one paper. This co-occurrence matrix is pretty sparse when a document is an abstract. Thus we achieve a reasonable running time for the two term topic score experiment.

4.1.1 Detecting Topics Represented by a Single Term

Computing the topic scores for each term in the corpus according to Eq.1 gives a ranked list of topics. The ranked list of terms has 137,098 entries (terms), where top entries constitute topic terms and bottom entries constitute non-topic terms. Table 1 shows the top 15 entries from the ranked list. The first 2 columns represent the rank and the topic term respectively. The third column labeled as $< n, n_c, |E| >$ is an information about the term citation graph of the topic term: n is the number of nodes in the citation graph of a topic term, n_c is the number of nodes that has at least one link connecting to any other node within the graph, $|E|$ is the number of edges in the graph.

An objective and quantitative evaluation of the result is difficult due to the lack of standard formal measures for topic detection tasks. However, when the results were examined by the domain experts, they recognized the topics presented in Table 1. Also, we have an informal evidence that these top ranked terms do represent highly prevalent topics in the physics literature. When we typed in each topic term of the top 20 ranks as a search query to www.google.com, 19 of them returned Wikipedia entries within the top 5 of the google search results. The inspection of the Wikipedia articles reveals that most of them have serious physics research oriented contents. The one topic term that did not return the Wikipedia entry was "heavy quark". But, the second rank entry of its google search result is "The 5th international workshop of heavy quark physics", indicating that it also is a prevalent research topic in physics.

The topic terms at the top ranks are topics in large scale, as we can see from the term citation graph information of $< n, n_c, |E| >$ column. The topic term entries down to a few thousand'th level of the ranked list still present meaningful topics. Table 2 shows a few entries of topic terms around 100'th, 500'th, 1000'th, 2000'th ranks. There is an apparent trend of topic scale getting smaller as we go down to lower ranked topic terms, as seen from $< n, n_c, |E| >$ column. Topics discovered at these levels could be more interesting as they tend to represent more specific ideas than the more generic and prevalent top ranked topic terms. Figure 4 shows the term citation graphs of the topic terms at 100'th, 990'th, 1971'th ranks, respectively (Refer to Footnote 2 for how to read the graphs). We see the scale difference of the topic terms at different ranks. Figure 4 c) suggests that even at 1971'th rank, there is still a meaningful topic that binds the papers in the term citation graph.

As explained in section 2, the bottommost entries of the ranked list are stop words or common phrases, whose term citation graphs are much better explained by hypothesis H0 than by hypothesis H1. Table 3 shows the bottommost 15 terms of the ranked list.

It should be noted that the topics discovered by our algorithm have a varying degree of prevalence and specificity, that are natural in the given corpus. This is because we do not assume a predefined number of topics to discover, as language model approaches or graph-based clustering approaches do. Fixing the number of topics to discover has the effect of determining the scale of topics in advance.

To see the overall property of the entire ranked list, we present two plots, Figure 5 and Figure 6. Figure 5 is a plot of term rank vs. the log size of the term citation graph averaged over 100 consecutive terms. It shows that the term frequency gets higher, as the rank gets close to either the highest or the lowest ranks. This is because in a large-scale term citation graph one hypothesis is strongly preferred over the other due to many instances of per-node observations that support the hypothesis.

To show the connectivity of term citation graphs, we devise the following measure and use it in the plot of Figure 6. Given a term citation graph G_A, $c_i(G_A)$ denotes the number of links of a node i that falls within G_A, l_i denotes the

top rank	topic (term)	$< n, n_c, \|E\| >$
1	black hole	$< 4978, 4701, 38952 >$
2	quantum hall	$< 1863, 1493, 4862 >$
3	black holes	$< 3131, 2896, 22824 >$
4	higgs boson	$< 2079, 1896, 12607 >$
5	renormalization group	$< 3738, 2920, 8490 >$
6	quantum gravity	$< 2014, 1724, 9693 >$
7	standard model	$< 7848, 7145, 53829 >$
8	heavy quark	$< 1671, 1473, 6570 >$
9	cosmological constant	$< 2141, 1815, 7134 >$
10	quantum dot	$< 1366, 1031, 2926 >$
11	chiral perturbation	$< 1132, 1050, 5578 >$
12	form factors	$< 1578, 1354, 5616 >$
13	lattice qcd	$< 1425, 1265, 5240 >$
14	string theory	$< 3818, 3539, 26250 >$
15	hubbard model	$< 1702, 1167, 2678 >$
...	...	...

Table 1: The topic terms of top 15 ranks from arXiv

total number of links of a node i, n_A denotes the size of G_A, and N denotes the size of the full citation graph. We call $\sum_i c_i (G_A) / \sum_i l_i$ as edge containment. It reflects how clustered G_A is with respect to the rest of the full citation graph. We normalize the edge containment by the relative size of the term citation graph. We call the resulting quantity $\frac{\sum_i c_i(G_A)/\sum_i l_i}{n_A/N}$ as normalized edge containment. The normalized edge containment should default to 1 if the citation pattern of G_A is random. Figure 6 shows a plot of term rank vs. the normalized edge containment. As expected, the topic terms at high ranks show high normalized edge containment, while the non-topic terms at low ranks show low normalized edge containment. What is interesting to note is that the graph is not monotonically decreasing: up to the top few thousand ranks, the normalized edge containment keeps increasing. This agrees with our observation that the middle rank topics are more specific than the top rank topics.

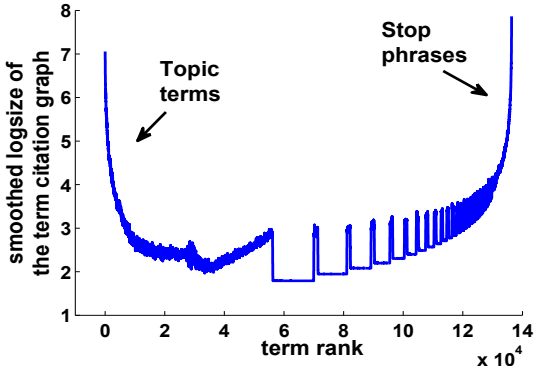

Figure 5: A plot of term rank vs. log(size of term citation graphs)

4.1.2 *Detecting Topics Represented by a Set of Terms*

By computing the term pair topic scores of Eq.9 for all possible pairs of terms in arXiv corpus, we get a ranked list

top rank	topic (term)	$< n, n_c, \|E\| >$
...	...	...
95	fractional quantum	$< 552, 381, 729 >$
96	qcd corrections	$< 597, 500, 1175 >$
97	mass matrix	$< 742, 606, 2627 >$
98	string field	$< 505, 465, 5708 >$
99	entangled states	$< 634, 472, 1014 >$
100	potts model	$< 426, 321, 718 >$
101	electroweak symmetry	$< 673, 559, 2052 >$
...	...	...
497	vacuum expectation	$< 713, 443, 696 >$
498	higgs doublets	$< 280, 205, 384 >$
499	boundary state	$< 168, 147, 529 >$
500	spin polarization	$< 494, 261, 406 >$
501	abelian gauge	$< 537, 319, 837 >$
...	...	...
989	matrix string	$< 76, 69, 222 >$
990	charmed baryons	$< 77, 61, 104 >$
991	geometric phases	$< 102, 67, 87 >$
992	kerr black	$< 189, 115, 229 >$
993	kp hierarchy	$< 90, 62, 95 >$
994	pseudoscalar mesons	$< 272, 164, 201 >$
...	...	...
1968	traversable wormholes	$< 42, 35, 94 >$
1969	b-meson decays	$< 90, 61, 71 >$
1970	penguin operators	$< 53, 44, 87 >$
1971	two-dimensional qcd	$< 42, 34, 43 >$
...	...	...

Table 2: The topic terms at various ranks from arXiv

bottom rank	topic (term)	$< n, n_c, \|E\| >$
1	we show	$< 26906, 19479, 53311 >$
2	has been	$< 9992, 4231, 5528 >$
3	we find	$< 21474, 15187, 42792 >$
4	we present	$< 16898, 10808, 24410 >$
5	we study	$< 19976, 14192, 37322 >$
6	we have	$< 8396, 3411, 3773 >$
7	we also	$< 15983, 11074, 33095 >$
8	have been	$< 6636, 2422, 2686 >$
9	we discuss	$< 12837, 8410, 18755 >$
10	we consider	$< 11551, 7079, 13647 >$
11	does not	$< 6155, 2488, 2814 >$
12	our results	$< 6224, 2815, 3144 >$
13	we investigate	$< 8437, 4585, 5788 >$
14	into account	$< 4910, 1952, 2521 >$
15	we propose	$< 6387, 3127, 4325 >$
...	...	...

Table 3: The terms with the lowest topic scores from arXiv

where each entry is a pair of terms that might represent a topic. Since we are looking at the intersection citation graph of two terms, we get a sparser graph to look at. In order to alleviate the sparseness, we stemmed our corpus. Table 4 shows the top 12 entries of the ranked list. These entries are the topics that are represented not by a single term, but by the relation involving a set of terms. For example, the rank

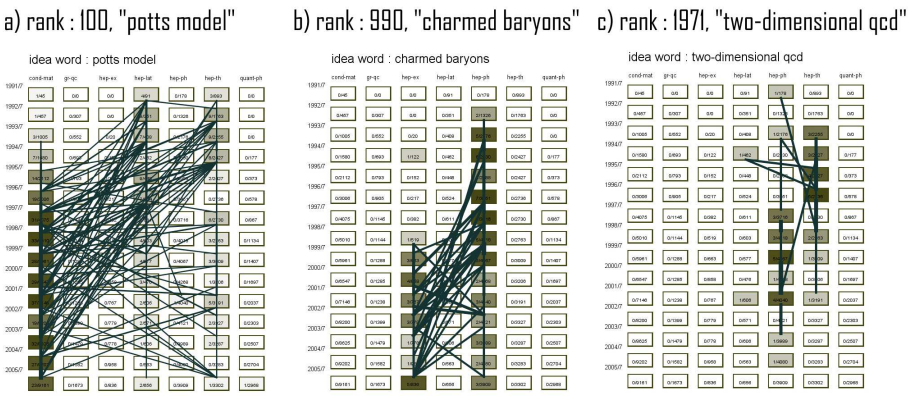

Figure 4: The term citation graphs of topic terms at various ranks from arXiv

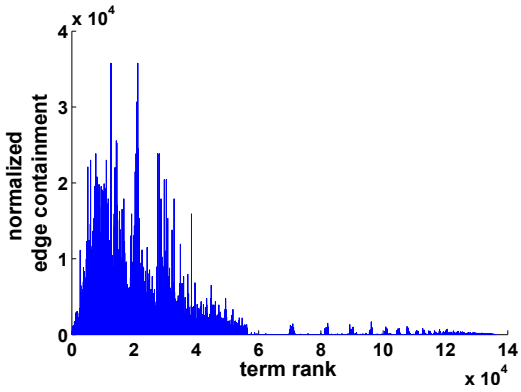

Figure 6: A plot of term rank vs. normalized edge containment

1 entry has "phase transit(ion)" and "standard model" as its topic terms. "phase transition" is a general term meaning a change in macroscopic state of a large-scale system. "standard model" is a prevalent theory of particle physics that describes the fundamental interactions of elementary particles. It turns out that the papers at the intersection of the two terms talk about the "phase transition" occuring in "standard model" or the "phase transition" occuring in minimal supersymmetric "standard model" which is an extension of "standard model". The individual term "phase transition" or "standard model" has a much broader research context than the topic identified. The rank 2 entry has "gauge theory" and "matrix model" as its topic terms. It turns out that there was a heavily cited paper that started the whole idea of analyzing "gauge theory" using the computational techniques from "matrix model", and the majority of papers in the intersection graph talk about the further development of this idea. As explained in the previous section 3, the papers of the rank 7 entry talk about using "quantum dot" as a hardware implementation of "quantum computer".

The last three columns of table 4 show the citation graph information $< n, n_c, |E| >$ for term A, term B, and their intersection, respectively. They show that the connectivity of the intersection graph $G_{A \cap B}$ exhibits significant departure from the same size random subgraph of G_A or G_B.

4.2 Evaluation on Citeseer Data

Our Citeseer data contains 716,771 papers, with 1,740,326 citations. This amounts to 2.43 citations per paper. For each paper, we use its title and abstract combined as its document. The number of bigrams in the corpus after pruning out the low-frequency bigrams and 35 stop words is 631,839. The majority of papers are from year 1994 to year 2004. We divided the documents into two different document sets. One set contains all the documents up to year 1999, and the other set contains all the documents since year 2000.

We performed the single term topic score measure of Eq.1 to each set. The top 25 topic entries of each set are shown in parallel in Table 5. We see that the top rank topics have changed significantly between the two time periods. We see that many top rank topics of the time frame since 2000 carry recent trends that were not significant before. Examples are "sensor networks", "(ad) hoc networks", "wireless sensor", "intrusion detection", "semantic web", "xml data", and "image retrieval". "support vector (machine)" was ranked 35th in the document set up to 1999, and it has risen to the rank 5 in the document set since 2000. "congestion control" more or less maintains its topic rank through the different time periods. We observe the fall of many top ranked topics of the document set up to 1999, in the time period since 2000. The ranked list of topic terms is quite instructive as well: initially, we did not recognize the 7th rank "interior point" of the time frame up to 1999 as a topic. But, it turns out "interior point" represents an important family of algorithms in linear programming.

As in the case of arXiv evaluation, we see that the ranked list of topic terms from Citeseer has meaningful topics even around a thousand'th level with the apparent trend of topic scale getting smaller as we go down the ranks. Due to the space limitation, however, we do not present the result.

In order to see the time evolution of topics more clearly, we performed the following experiment. We ran the single term topic score measure for the entire Citeseer document collection. Then, for each term in top 70, we generated a plot where x-axis is years spanning from 1994 to 2004 and y-axis is the number of documents of the term citation graph in a particular year normalized by the total number of documents in that year. Figure 7 shows the plots for 12 such topic terms out of top 70 terms. We see a sharp recent rise of

rank	term A	term B	$< n, n_c, \|E\| >$		
			for term A	for term B	for $A \cap B$
1	phase transit	standard model	$< 6862, 4693, 15535 >$	$< 7901, 7168, 54029 >$	$< 200, 159, 816 >$
2	gaug theori	matrix model	$< 5907, 5186, 35446 >$	$< 1332, 1217, 9187 >$	$< 168, 138, 1055 >$
3	form factor	sum rule	$< 2444, 2139, 10205 >$	$< 2120, 1702, 7775 >$	$< 285, 252, 1014 >$
4	dirac oper	random matrix	$< 618, 523, 3206 >$	$< 633, 450, 2475 >$	$< 88, 88, 714 >$
5	black hole	cross section	$< 6491, 6168, 64085 >$	$< 5188, 4411, 20358 >$	$< 84, 66, 280 >$
6	heavi quark	sum rule	$< 2047, 1817, 8556 >$	$< 2120, 1702, 7775 >$	$< 186, 151, 470 >$
7	quantum comput	quantum dot	$< 1975, 1768, 8652 >$	$< 2328, 1898, 7593 >$	$< 137, 118, 400 >$
8	gaug theori	spin chain	$< 5907, 5186, 35446 >$	$< 828, 591, 2299 >$	$< 56, 54, 330 >$
9	cross section	dark matter	$< 5188, 4411, 20358 >$	$< 1618, 1388, 8326 >$	$< 131, 120, 424 >$
10	black hole	planck scale	$< 6491, 6168, 64085 >$	$< 709, 554, 1523 >$	$< 92, 64, 310 >$
11	boundari condit	scalar field	$< 3300, 2113, 5510 >$	$< 4405, 3496, 10927 >$	$< 229, 134, 287 >$
12	cross section	standard model	$< 5188, 4411, 20358 >$	$< 7901, 7168, 54029 >$	$< 611, 483, 1232 >$
...	...	...	...	...	...

Table 4: The top 12 entries of two term topic scores from arXiv

"sensor networks" and "semantic web", a significant rise of "support vector" and "energy consumption", a rise of "xml data" in a smaller scale, the fall of "logic programs", "petri nets", "interior points". "congestion control", "association rules", and "genetic programming" show less dramatic dynamics.

5. RELATED WORK

Our work is distinguished from previous work on topic detection in two ways. First, we look at the correlation between the term distribution and the citation link distribution for a topic. Second, we use for a topic measure the log odds ratio of binary hypotheses based on a probabilistic description of graph connectivity.

rank	topic (term) up to 1999	topic (term) since 2000
1	logic programs	sensor networks
2	model checking	hoc networks
3	semidefinite programming	logic programs
4	inductive logic	image retrieval
5	petri nets	support vector
6	genetic programming	congestion control
7	interior point	model checking
8	kolmogorov complexity	decision diagrams
9	automatic differentiation	wireless sensor
10	complementarity problems	ad hoc
11	congestion control	instrusion detection
12	complementarity problem	vector machines
13	conservation laws	mobile ad
14	linear logic	binary decision
15	timed automata	sensor network
16	situation calculus	energy consumption
17	real-time database	content-based image
18	motion planning	semantic web
19	duration calculus	fading channels
20	volume rendering	xml data
21	chain monte	source separation
22	association rules	timed automata
23	term rewriting	signature scheme
24	posteriori error	volume rendering
25	active database	xml documents
...	...	...

Table 5: The top 25 topic terms of two different time periods from Citeseer

Previous work on topic detection can be largely divided into two groups. The majority of papers take a language model based approach. This approach tends to focus on text, but a few papers extend the model to incorporate links. Another group of work is based on studies of graph properties. Most of these papers address a problem related to topic detection: community detection. They tend to use the non-probablistic aspects of graph properties. There are also related papers that share some of the ideas used in this paper. Specifically, these ideas are examination of patterns at individual term level, usage of log odds ratio to detect patterns, and investigation of the notion of term informativeness.

The language modeling approach [9, 22, 23, 24, 7, 16] assumes a multi-stage generative process where semantically meaningful modalities such as topics or authors are chosen as an intermediate step, and then the words are drawn from the multinomial distribution conditioned on these modalities. These papers differ in the design choice for the generative process. Examples of design decisions are the choice of modalities or the final features that will be produced. [9] uses the document generation process conditioned on topic distributions. [22] uses authors as distribution over topics as additional modalities. [23] detects topics over time by letting the generative process produce the timestamps of words as well as the words themselves. A number of papers extend the model to incorporate links. [7] treats the reference list of a paper as another final feature to produce, in addition to the bag of words. [16, 24] apply the language model approach to social network analysis where documents are the communication links such as e-mail messages between people. [4] aims to overcome the inability of latent dirichlet allocation used in the papers above for describing the correlation of topics, by including the correlation matrix of topics in the generative process. [17] computes the themes of a document collection by the mixture model using the EM algorithm.

Graph properties are used to study community structures by [10, 8, 11, 20, 19, 13, 2]. As a distance metric [10] uses the similarity of citation patterns, [8, 11] use the notion that

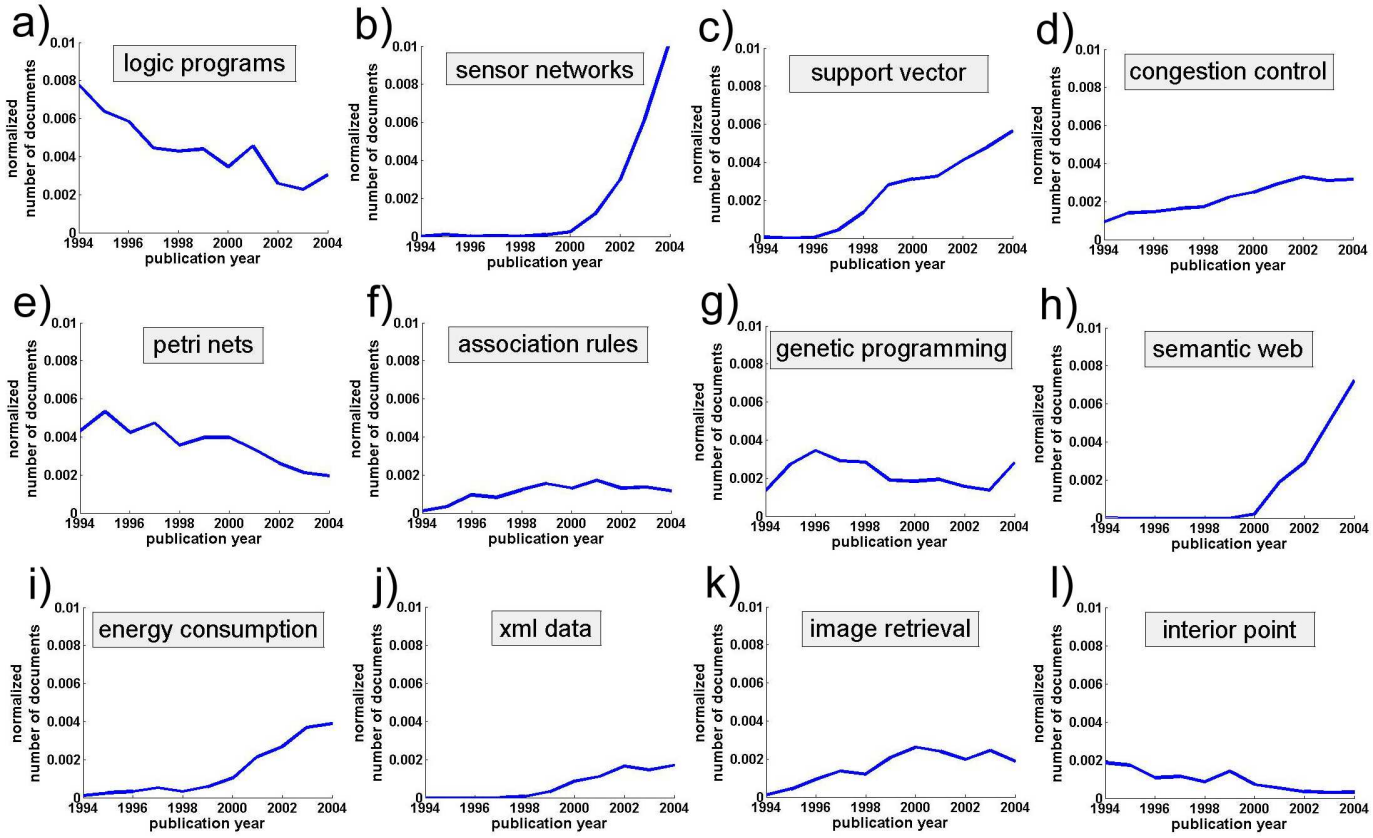

Figure 7: The topic evolution over time in Citeseer. a) "logic programs", b) "sensor networks", c) "support vector", d) "congestion control", e) "petri nets", f) "association rules", g) "genetic programming", h) "semantic web", i) "energy consumption", j) "xml data", k) "image retrieval", l) "interior point"

nodes have more links to the members of the same community than to other nodes, [20] introduces the concept of edge betweenness, and [19] uses the measures from bibliometry and graph theory. Some papers in this group combine the information from text as well. [13] extracts storylines for a query by identifying densely connected bipartites from the document-term graph of the search results. [2] improves the document categorization performance by starting from a text-based categorization result and then iteratively relabeling the documents to further satisfy the constraints imposed by the link proximity relation.

Our approach of looking at citation patterns at an individual term level and using the loglikelihood to explain the observation is inspired by [12]. [12] detects a topic as a burst of activities represented by the state transition in a markov chain. In an experiment on paper titles and presidential speeches the paper shows that topics can be effectively detected as time bursts in a single term level. The idea of anomaly detection by log odds ratio is used in a number of papers related to topic detection. [18] uses the log odds ratio of event frequencies to detect space-time clusters. [14] discovers a set of words as topic signature in a supervised learning setting by comparing the log odds ratio of word frequency in topic documents and non-topic documents. The ranked list of terms for topics produced by our algorithm shows a continuous spectrum of term informativeness in representing topics. The notion of term informativeness is explored in a number of related contexts. [5] detects the terms informative about the citation links to use them as features for document categorization. For this purpose, they use the expected entropy loss measure, which resembles the one used in the decision tree feature selection. [21] detects the informative terms for named entity detection, using the idea that informative terms are better modeled by a mixture of two unigram models while non-informative terms are better modeled by a single unigram model.

6. DISCUSSIONS

It is worthwhile to note that graph connectivity observations used in our algorithms are pluggable. One can plug in an observation that best suits one's need. As different observations may represent different aspects of graph connectivity, the choice of an observation affects the topic score result. For example, the observation used in Section 2 for a single term topic score concerns whether a node has a connection to the rest of the graph or not, but it does not distinguish how many connections a node has. Thus, the observation is generous on loosely connected topics. As a result, the highest topic scores are given to the large-scale prevalent topics even though these topics are not as tightly connected as the more specific smaller scale topics.

Another point to note is that because our algorithms dis-

cover topics without imposing any constraint on the relationship among topics such as restricting the number of topics to be discovered or assuming implicit mutual exclusion among topics, the topics discovered are suitable for expressing the complex relationship among topics. Specifically, with our topics, a single document can be involved in multiple topics, and topics could have hierarchical covering relations or non-hierarchical overlapping relations among themselves. Understanding the structure of the relationship between the topics is left as future work.

7. CONCLUSIONS

In this paper, we presented algorithms to detect topics from a linked textual corpus based on the unique approach of using the correlation between the term distribution and the link distribution for topics. Our algorithms produce a ranked list of terms for topics represented by a single term and for topics represented by a set of terms. Our evaluation on arXiv and Citeseer data show that the method is effective. Topics discovered by our algorithms reveal novel aspects of topic detection. The ranked list shows a continuous spectrum of topics of varying prevalence and specificity that are natural in the given corpus. The relations among the terms that represent topics are revealed by the two term topic score measure. As an interesting by-product, our algorithm can discern common phrases without the prior knowledge of stop word notion. The possibility of discovering complex topic relations and the pluggable characteristic of graph connectivity observations are discussed.

8. ACKNOWLEDGMENTS

We would like to thank Simeon Warner for providing arXiv data, and Isaac G. Councill for providing Citeseer data. We would like to thank Prof. John E. Hopcroft for his input on graph properties and his encouragement, and Prof. Thorsten Joachims for valuable discussions. This work was supported in part by National Science Foundation under award numbers IIS-0430906, 0227648, 0227888, and 0424671.

9. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *SIGMOD*, 1993.

[2] R. Angelova and G. Weikum. Graph-based text classification: Learn from your neighbors. In *Proceedings of SIGIR*, 2006.

[3] arXiv. *http://arxiv.org*.

[4] D. M. Blei and J. D. Lafferty. Correlated topic models. In *NIPS*, 2005.

[5] L. Bolelli, S. Ertekin, and C. L. Giles. Clustering scientific literature using sparse citation graph analysis. In *PKDD*, pages 30–41, 2006.

[6] Citeseer. *http://citeseer.ist.psu.edu*.

[7] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101, 2004.

[8] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *Proceedings of SIGKDD*, 2000.

[9] T. I. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, (5):5228–5235, 2004.

[10] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. Natural communities in large linked networks. In *Proceedings of SIGKDD*, 2003.

[11] H. Ino, M. Kudo, and A. Nakamura. Partitioning of web graphs by community topology. In *Proceedings of WWW*, 2005.

[12] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of SIGKDD*, 2002.

[13] R. Kumar, U. Mahadevan, and D. Sivakumar. A graph-theoretic approach to extract storylines from search results. In *Proceedings of SIGKDD*, 2004.

[14] C.-Y. Lin and E. Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the COLING Conference*, Strausbourg, France, 2002.

[15] G. S. Mann, D. Mimno, and A. McCallum. Bibliometric impact measures leveraging topic analysis. In *JCDL*, 2006.

[16] A. McCallum, A. Corrada-Emmanuel, and X. Wang. The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email. *Technical Report*, 2004.

[17] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text - an exploration of temporal text mining. In *Proceedings of SIGKDD*, 2005.

[18] D. B. Neill, A. W. Moore, M. Sabhnani, and K. Daniel. Detection of emerging space-time clusters. In *Proceedings of SIGKDD*, 2005.

[19] M. Newman. Scientific collaboration networks. i. network construction and fundamental results. *PHYSICAL REVIEW E*, 64, 2001.

[20] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *arXiv:cond-mat/0308217*, 2003.

[21] J. D. M. Rennie and T. Jaakkola. Using term informativeness for named entity detection. In *Proceedings of SIGIR*, 2005.

[22] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of SIGKDD*, 2004.

[23] X. Wang and A. McCallum. Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings SIGKDD*, 2006.

[24] D. Zhou, E. Manavoglu, J. Li, C. L. Giles, and H. Zha. Probabilistic models for discovering e-communities. In *Proceedings of WWW*, 2006.