

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Enhanced Content Development Headed for Knowledge Discovery from Citation Networks

K. Subhashni¹, Dr. K.Subramanian M.Sc., M.Phil., Ph.D.,²

Master of Philosophy, Department of Computer Science, J.J. College of Arts and Science, Pudukkottai,
Tamil Nadu, India¹

Assistant Professor, Department of Computer Science, J.J. College of Arts and Science, Pudukkottai,
Tamil Nadu, India.²

ABSTRACT: The main objective of this system is to develop a system that serves both purposes, leading readers to the key literature in the areas of their interest, while also providing suggestions that may be as valuable as they are non-obvious. The recommendation system (based on a seed paper) needs to determine: (a) what papers are relevant to the seed paper, (b) of these, which are the most important and (c) for different user types. The scholarly literature is expanding at a rate that necessitates intelligent algorithms for search and navigation. For the most part, the problem of delivering scholarly articles has been solved. If one knows the title of an article, locating it requires little effort and, paywalls permitting, acquiring a digital copy has become trivial. However, the navigational aspect of scientific search – finding relevant, influential articles that one does not know exist – is in its early development. In this paper, we introduce Eigenfactor Recommends – a citation-based method for improving scholarly navigation. The algorithm uses the hierarchical structure of scientific knowledge, making possible multiple scales of relevance for different users. We implement the method dynamically and generate many recommendations from user articles from various bibliographic collections using server side scripting nature.

KEYWORDS: Citation Network, Reference Evaluation, Document Level Citation, Knowledge Discovery.

I. INTRODUCTION

Back in the twentieth century, when libraries were predominantly physical institutions, the process of document acquisition facilitated serendipitous discoveries of related work and the formation of unexpected intellectual connections. To read a journal article, a scholar had to physically acquire the issue in which it was printed; in the process she would often stumble fortuitously across other relevant articles in the same issue. To read a book, a researcher would first have to navigate the library stacks and scan across dozens of titles—often finding even more relevant volumes in the process.

No longer. Today, digital document delivery approaches a single-click level of ease. A scholar types a few words of a title, or the surnames of a few authors into a search engine, portal, or document repository, and can then proceed immediately to the required document without any of the search and browsing time that we routinely had to invest only ten or fifteen years ago. For the most part, this represents an enormous increase in efficiency, but something has been lost as well: readers are no longer exposed to related material as part of the document acquisition process.

Fortunately, this loss is not a necessary consequence of the digital transition in scholarship. We can deliberately engineer tools to replicate the benefits of these previous serendipitous processes, but with hugely greater time efficiency and perhaps with better-than-chance ability to direct researchers to relevant studies.

Balancing serendipity with relevancy, however, is challenging given the exponential growth of scientific knowledge. The volume of scholarly literature roughly doubles every decade. More than a million articles are added to this corpus on an annual basis. We are long past the days in which a scholar could keep up with the literature simply by reading through each issue of each journal in her field as it was published.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Exhaustive search no longer scales, and we face a desperate need for informetric tools that facilitate document discovery. Of course, tools for targeted search—designed to help researchers find needed papers when they know approximately what they are looking for will play an important role, as will systems for more efficiently navigating the topography of academic scholarship from papers the researcher does not know exist. But there is a role of chance discovery as well, the sort of chance discovery that used to be a regular miracle among the stacks of our university libraries.

This is our aim in the present system: to develop a system that serves both purposes, leading readers to the key literature in the areas of their interest, while also providing suggestions that may be as valuable as they are non-obvious.

We propose a new citation-based approach to this problem called Eigenfactor Recommends (EFrec) that meets these three objectives. We couple the hierarchical structure of the citation network - which reflects the natural hierarchical structure of scientific domains, fields, subfields, and so forth - with importance scoring based upon a network centrality measure. In this way we use hierarchical clustering to determine relevance and then recommend papers based upon their importance within these clusters.

Thus, we are able to generate a spectrum (or scale) of recommendations for any given topic, paper, or set of key words. We can find papers that are very closely related but perhaps not yet very influential (Expert Recommendations). Alternatively, we can find papers that may be more distantly related but represent foundational contributions to the broader area of research (Classic Recommendations) for researchers new to a field. We find that this approach provides recommendations for a much larger set of papers than one can provide using a co-citation approach, and performs at least as well in AB testing. It also, distinctly, provides recommendations at different scales for different user types.

II. METHODOLOGIES

The objective of the EFrec algorithm is to find relevant papers, given a seed paper³. The data required to generate these recommendations is a citation graph. The output consists of a list of paper IDs and a set of recommended papers associated with each paper ID. The number of recommendations to be obtained can be set by the user, and range from one to hundreds of recommendations. Figure 1 illustrates the four-step process for generating paper recommendations.

(a) Assemble citation network

The first step requires assembling the citation graph of a relatively large corpus, where “relatively large” means at least a few hundred thousands papers and a few million citations – bigger the better. This method does not perform as well for small graphs. The graph is represented as a link list file where column 1 is the citing paper ID and column 2 is the cited paper ID. For the AMiner dataset, the network consisted of 2,092,356 nodes (papers) and 8,024,869 links (citations). Because each paper cites only a small number of other references, the networks underlying large corpora will be highly sparse, as described in the section Sparseness of Citation Graphs.

(b) Rank nodes

The second step consists of ranking the nodes. We rank the nodes according to the article-level Eigenfactor (ALEF⁴). The original Eigenfactor algorithm, which is closely related to original PageRank algorithm, performs well on journal-level citation graphs because these are cyclic, in that one can follow directed links (citations) from one journal out to other journals and then back to the originating journal. At the article level, however, citation graphs are acyclic or nearly so: a paper only cite those articles that preceded it in time. When standard PageRank approaches are applied to acyclic citation graphs, older papers are excessively weighted. ALEF is a modified version of PageRank (PR), tailored specifically for the time-directed acyclic networks associated with article-level citations. We show in a recent study that ALEF performs better than PageRank and degree centrality and have also demonstrated that ALEF better separates papers that contribute to a given theory. As mentioned above, the problem with using a PageRank approach on article-level citation graphs is that the “random walker” on the graph will move inexorably backwards in

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

time, and as a result will over-weight older papers. ALEF addresses these issues with the following two modifications: (1) it shortens the number of contiguous steps of the random walker before she teleports to another part of the network and (2) it teleports to links rather than nodes. These modifications help adjust for the over-weighting without losing the ability to exploit network structure.

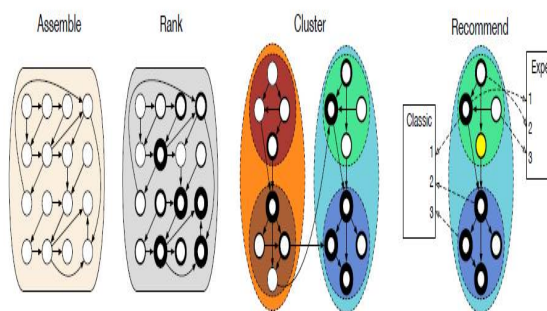


Fig. 1. Eigenfactor Recommends. The process for generating paper recommendations includes the following steps. (1) In the first step, the paper citation graph is assembled into an adjacency matrix using the citations (links) between papers (nodes). (2) The second step ranks the nodes using the article-level Eigenfactor algorithm (ALEF). This is a modified, time-directed version of the PageRank algorithm [15]. (3) The third step imports the citation graph and node rankings and then clusters the citation graph using the hierarchical MapEquation [23]. (4) The last step generates recommendations given a seed paper (highlighted in yellow). Expert recommendations are drawn from the lowest level of the hierarchical tree, while Classic recommendations include papers from one level up the hierarchy. A paper can be both an expert recommendation and a classic recommendation if it is the highest rated paper at the lowest and upper levels.

The mechanics of ALEF are relatively simple and proceed in five steps. First, the teleportation weight, w_i for each node i is calculated by summing the in-and out citations. Second, the random walker teleports to a random node based on the teleportation weights.

Third, the random walker takes a step along the citation graph. Fifth, we compute the asymptotic rate at which each node is visited under this process. And fifth, these rates—which provide our rankings—are normalized so that the average score for all papers equals 1. Specifically, the article citation network can be represented as an $n \times n$ adjacency matrix, Z , where the Z_{ij} entry equals 1 if article i cites article j and 0 otherwise. The matrix is highly sparse since an individual article cites a tiny portion of all articles in the corpus.

Hierarchically Cluster Nodes

The third step is to cluster the network, using the citation graph from step 1 and node rankings from step 2. Based on these inputs, the hierarchical MapEquation uncovers the boundaries between domains, fields, subfields, etc. For simplicity, we begin with a description of the nonhierarchical map equation, which uncovers basic modular structure within networks, returning a hard partition in which each node is assigned to a single module.

The map equation exploits the duality between compressing data and revealing patterns within the data. The core idea is to compress a description of a random walk on the network. If the network has localized regions such that a random walker has a long persistence time in a small group of nodes, a random walk can be concisely encoded when this structure is exploited in the coding scheme.

Recommendation Selection

The last step involves selecting nodes for three types of recommendations: (1) Expert, (2) Classic and (3) Serendipity. Each of these types use no specific information about the user⁷. However, the different recommendation types offer results that will be most useful to different kinds of users.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

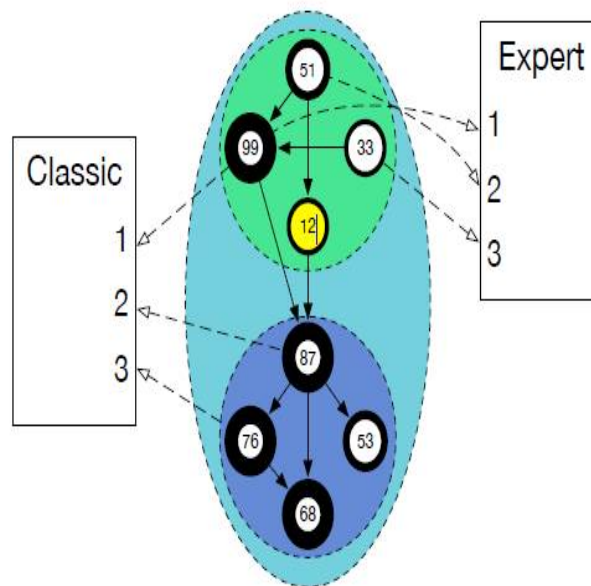


Fig. 2. Recommendation Closeup. A closeup of Fig. 14 shows how Classic and Expert recommendations are generated for a given seed paper (highlighted, score of 12). Expert recommendations are drawn from the same leaf node (sea foam green) as the seed paper, while Classic recommendations include sibling nodes as well (aqua). Papers are ordered by their Eigenfactor score, descending. The Classic recommendations provide a more diverse set of papers, while Expert recommendations are more specific.

The Expert recommendations are aimed at researchers familiar with a community of papers and authors. The algorithm aims to select papers that are highly specific to a particular sub-discipline. The Classic recommendations, on the other hand, are geared towards a graduate student or someone new to a specific field of science. The Classic papers are foundational articles in a field that every first year graduate student in the field should read. Expert and classic recommendation examples for the famous Kleinberg paper on "Authoritative sources in a hyperlinked environment" can be found in Table 2 and Table 3. The Serendipity recommendations are papers that are randomly chosen for every new user session. However, they are not randomly chosen from any part of the corpus. Rather, they are chosen from within the relevant community of papers as defined by the hierarchical network structure. The size distribution of communities for the different recommendations can be found in Figure 3.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

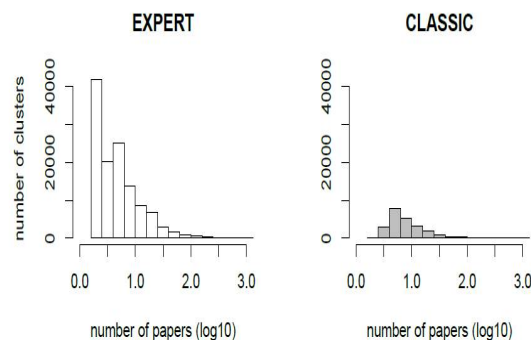


Fig. 3. Cluster size distribution. The bar plots show the number of clusters with a given number of papers. For the expert recommendation, clusters are extracted from the lowest part of the tree; therefore, there are more clusters with smaller number of papers. For the classic recommendation, the algorithm zooms out one level of the tree. This produces fewer clusters with more papers in each cluster. We did not include singletons – clusters with only one paper – in this figure. There were approximately 80k singletons in the AMiner tree.

Existing system:

There are collaborative filtering approaches, which rely on finding similar users and using their ratings to provide recommendations. They require no knowledge about the items being recommended, and with sufficient user ratings can infer properties about the items in question. Content based methods are ideal for datasets where there are few users, however there is substantial difficulty in automated feature extraction. Term frequency-inverse document frequency (TF-IDF) is a commonly used technique

Disadvantages

- ✚ Limited good rating coverage ie the number of ratings should dominate the number of items being rated to perform well
- ✚ The problem with using a PageRank approach on article-level citation graphs is that the “random walker” on the graph will move inexorably backwards in time, and as a result will over-weight older papers
- ✚ In scientific literature, synonymy is an issue. Like usage data, full text is difficult to obtain from publisher

Proposed Methodology

Eigen factor recommendation algorithm finds relevant papers, given a seed paper³. The data required to generate these recommendations is a citation graph. The output consists of a list of paper IDs and a set of recommended papers associated with each paper ID. The number of recommendations to be obtained can be set by the user, and range from one to hundreds of recommendations.

Advantages

- ✚ Higher recommendation coverage
- ✚ Produced comparable recommendations as measured by click-through rate

III. LITERATURE SURVEY

We present various inequalities for Euler's beta function of n variables. One of our theorems states that the inequalities $\frac{1}{n} \sum_{i=1}^n x_i B(x_1, \dots, x_n/b_n)$ hold for all $x_i \in (0, 1)$ ($i = 1, \dots, n$; $n \geq 3$) with the best possible constants $\frac{1}{n}$ and $b_n \geq 1 - \frac{1}{n}$. This extends a recently published result of Dragomir et al., who investigated (1) for the special case $n = 2$.

Understanding how topics in scientific literature evolve is an interesting and important problem. Previous work simply models each paper as a bag of words and also considers the impact of authors. However, the impact of one document on another as captured by citations, one important inherent element in scientific literature, has not been considered. In this paper, we address the problem of understanding topic evolution by leveraging citations, and develop citation-aware approaches.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

We propose an iterative topic evolution learning framework by adapting the Latent Dirichlet Allocation model to the citation network and develop a novel inheritance topic model. We evaluate the effectiveness and efficiency of our approaches and compare with the state of the art approaches on a large collection of more than 650,000 research papers in the last 16 years and the citation network enabled by CiteSeerX. The results clearly show that citations can help to understand topic evolution better.

Knowledge discovery from scientific articles has received increasing attentions recently since huge repositories are made available by the development of the Internet and digital databases. In a corpus of scientific articles such as a digital library, documents are connected by citations and one document plays two different roles in the corpus: document itself and a citation of other documents. In the existing topic models, little effort is made to differentiate these two roles.

We believe that the topic distributions of these two roles are different and related in a certain way. In this paper we propose a Bernoulli Process Topic (BPT) model which models the corpus at two levels: document level and citation level. In the BPT model, each document has two different representations in the latent topic space associated with its roles. Moreover, the multilevel hierarchical structure of the citation network is captured by a generative process involving a Bernoulli process.

The distribution parameters of the BPT model are estimated by a variation approximation approach. In addition to conducting the experimental evaluations on the document modeling task, we also apply the BPT model to a well known scientific corpus to discover the latent topics. The comparisons against state-of-the-art methods demonstrate a very promising performance.

IV. SYSTEM ARCHITECTURE DESIGN

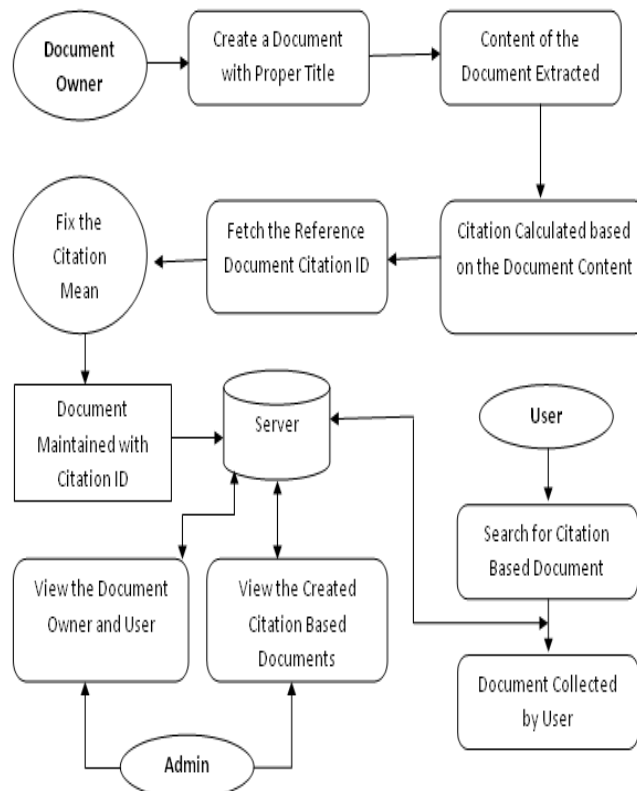


Fig.4. System Architecture Design

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

V. EXPERIMENTAL RESULTS

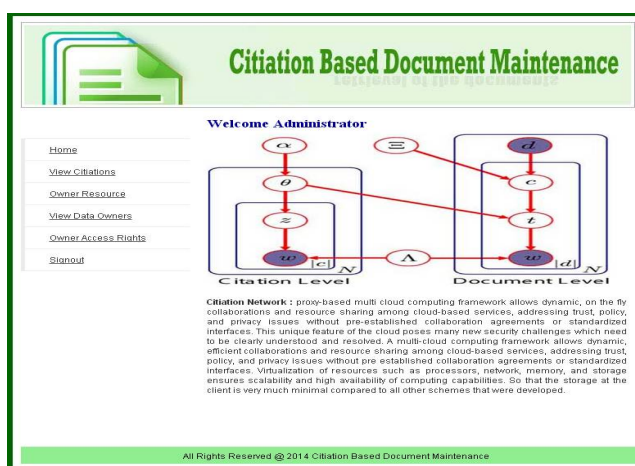


Fig.5. Home Page

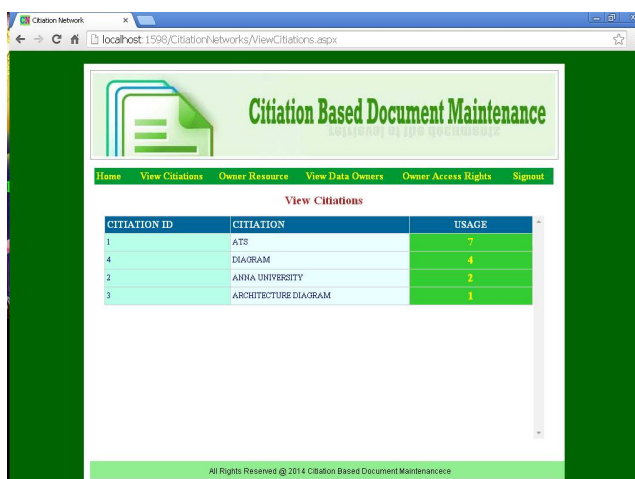


Fig.6. View Citations

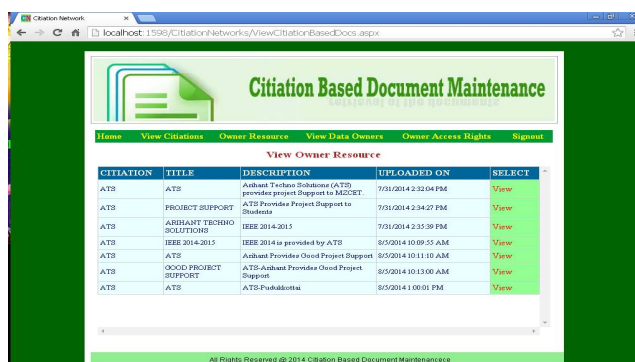



Fig.7. View Owner Resource

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)


Website: www.ijirset.com

Vol. 6, Issue 7, July 2017



ID	NAME	E-MAIL-ID	MOBILE	CITY	STATE	COUNTRY
1	S VENKATESAN	veskai08031987@gmail.com	7296283782	PUDUKKOTTAI	TAMILNADU	INDIA
2	SINDHU	sindhu08031987@gmail.com	9327847387	KARAIKUDI	TAMILNADU	INDIA
3	OKKULT	gola@gmail.com	9327847387	PUDUKKOTTAI	TAMILNADU	INDIA
4	A.J.ARUN	ajaru@gmail.com	9327847387	PUDUKKOTTAI	TAMILNADU	INDIA
5	P.SUDHANA	psuga@gmail.com	9456456453	PUDUKKOTTAI	TAMILNADU	INDIA

Fig.8. View Data Owners



ID	NAME	E-MAIL-ID	MOBILE	CITY	Activate	Deactivate
1	S VENKATESAN	veskai08031987@gmail.com	7296283782	PUDUKKOTTAI	Activate	Deactivate
2	SINDHU	sindhu08031987@gmail.com	9327847387	KARAIKUDI	Activate	Deactivate
3	OKKULT	gola@gmail.com	9327847387	PUDUKKOTTAI	Activate	Deactivate
4	A.J.ARUN	ajaru@gmail.com	9327847387	PUDUKKOTTAI	Activate	Deactivate
5	P.SUDHANA	psuga@gmail.com	9456456453	PUDUKKOTTAI	Activate	Deactivate

Fig.9. Activate/Deactivate Data Owners

VI. CONCLUSION AND FUTURE SCOPE

We present various inequalities for Euler's beta function of n variables. One of our theorems states that the inequalities $\frac{1}{n} \sum_{i=1}^n x_i \leq \frac{1}{n} \sum_{i=1}^n x_i^2$ hold for all $x_i \in [1, n]$ with the best possible constants D and B . This extends a recently published result of Dragomir et al., who investigated the special case $n = 2$.

Understanding how topics in scientific literature evolve is an interesting and important problem. Previous work simply models each paper as a bag of words and also considers the impact of authors. However, the impact of one document on another as captured by citations, one important inherent element in scientific literature, has not been considered.

In this system, we address the problem of understanding topic evolution by leveraging citations, and develop citation-aware approaches. We propose an iterative topic evolution learning framework by adapting the Latent Dirichlet

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Allocation model to the citation network and develop a novel inheritance topic model. We evaluate the effectiveness and efficiency of our approaches and compare with the state of the art approaches on a large collection of more than 650,000 research papers in the last 16 years and the citation network enabled by CiteSeerX. The results clearly show that citations can help to understand topic evolution better.

Knowledge discovery from scientific articles has received increasing attentions recently since huge repositories are made available by the development of the Internet and digital databases. In a corpus of scientific articles such as a digital library, documents are connected by citations and one document plays two different roles in the corpus: document itself and a citation of other documents. In the existing topic models, little effort is made to differentiate these two roles.

We believe that the topic distributions of these two roles are different and related in a certain way. In this paper we propose a Bernoulli Process Topic (BPT) model which models the corpus at two levels: document level and citation level. In the BPT model, each document has two different representations in the latent topic space associated with its roles. Moreover, the multilevel hierarchical structure of the citation network is captured by a generative process involving a Bernoulli process. The distribution parameters of the BPT model are estimated by a variational approximation approach. In addition to conducting the experimental evaluations on the document modeling task, we also apply the BPT model to a well known scientific corpus to discover the latent topics. The comparisons against state-of-the-art methods demonstrate a very promising performance.

REFERENCES

- [1] D. J. de Solla Price, "Networks of scientific papers," *Science*, vol. 169, pp. 510–515, 1965.
- [2] L. Bornmann and R. Mutz, "Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references," *Journal of the Association for Information Science and Technology*, 2015.
- [3] A. E. Jinha, "Article 50 million: an estimate of the number of scholarly articles in existence," *Learned Publishing*, vol. 23, no. 3, pp. 258–263, 2010.
- [4] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: Extraction and mining of academic social networks," in *KDD'08*, 2008, pp. 990–998.
- [5] P. B. Kantor, L. Rokach, F. Ricci, and B. Shapira, *Recommender systems handbook*. Springer, 2011.
- [6] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, pp. 42–49, 2009. [Online]. Available: [https://datajobs.com/data-science-repo/Recommender-Systems-\[Netflix\].pdf](https://datajobs.com/data-science-repo/Recommender-Systems-[Netflix].pdf)
- [7] L. L' u, M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, and T. Zhou, "Recommender Systems," p. 97, Feb. 2012. [Online]. Available: <http://arxiv.org/abs/1202.1112>
- [8] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods and metrics for cold-start recommendations," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2002, pp. 253–260.
- [9] X. Su and T. M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques," *Advances in Artificial Intelligence*, vol. 2009, no. Section 3, pp. 1–19, 2009. [Online]. Available: <http://www.hindawi.com/journals/aai/2009/421425/>
- [10] O. K' uc, " uktunc, E. Saule, K. Kaya, and U. C, ataly " urek, "Direction awareness in citation recommendation," vol. i, pp. 3–8, 2012. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary>
- [11] M. M. Kessler, "Bibliographic Coupling Between Scientific Papers," *American Documentation* (pre-1986), vol. 14, no. 1, p. 10, 1963.
- [12] H. Small, "Co-Citation in Scientific Literature: A new measure of the relationship between two documents," *Journal of the American Society for Information Science*, vol. 24, no. 4, pp. 265–269, 1973.
- [13] H. White and B. Griffith, "Author cocitation: A literature measure of intellectual structure," *Journal of the American Society for Information Science*, vol. 32, pp. 163–171, 1981.
- [14] J. M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.
- [15] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Seventh International World-Wide Web Conference (WWW 1998)*, 1998. [Online]. Available: <http://ilpubs.stanford.edu:8090/361/>
- [16] J. West, T. Bergstrom, and C. Bergstrom, "The eigenfactor metrics: A network approach to assessing scholarly journals," *College and Research Libraries*, vol. 71, no. 3, pp. 236–244, 2010.
- [17] J. D. West, M. C. Jensen, R. J. Dandrea, G. J. Gordon, and C. T. Bergstrom, "Author-level Eigenfactor metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 4, pp. 787–801, Apr. 2013. [Online]. Available: <http://doi.wiley.com/10.1002/asi.22790>
- [18] J. Bollen, M. A. Rodriguez, and H. Van de Sompel, "Journal status," *Scientometrics*, vol. 69, no. 3, pp. 669–687, 2006.
- [19] S. J. Liebowitz and J. P. Palmer, "Assessing the relative impacts of economics journals," *Journal of Economic Literature*, pp. 77–88, 1984.
- [20] G. Pinski and F. Narin, "Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics," *Information Processing & Management*, vol. 12, no. 5, pp. 297–312, 1976.