On Combining Text-based and Link-based Similarity Measures for Scientific Papers

Masoud Reyhani Hamedani
Computer Software Department,
Hanyang University, Seoul, Korea
masoud@agape.hanyang.ac.kr

Sang-Chul Lee
Computer Software Department,
Hanyang University, Seoul, Korea
korly@agape.hanyang.ac.kr

Sang-Wook Kim
Computer Software Department,
Hanyang University, Seoul, Korea
wook@agape.hanyang.ac.kr

ABSTRACT
In computing the similarity of scientific papers, text-based and link-based similarity measures look at only a single side of the content or citations. In this paper, we propose a new approach to compute the similarity of scientific papers accurately by combining the text-based and link-based similarity measures. Our proposed method considers the content and citations of the scientific papers simultaneously and combines the similarity scores based on the content and citations by using SVM^{rank}. The effectiveness of our proposed method is demonstrated via extensive experiments on a real-world dataset of scientific papers. The results show that more than 20% improvement in accuracy is obtained with our approach compared with previous methods.

Categories and Subject Descriptors
H.2.8 [**DATABASE MANAGEMEN**]: Database application–*Data mining*

General Terms
Algorithms, Experimentation, Measurement

Keywords
Citation, Content, Scientific Papers, Similarity

1. INTRODUCTION
Scientific papers are primary sources to share information and knowledge among scholars. Scientific literature search engines such as CiteSeerX and Google Scholar aid researchers to find papers in their areas of interest and to make sure, whether their research problems is novel by providing search or recommendation services. Similarity measure is one of the challenging issues in these systems to find relevant papers according to the user requirement.

To compute the similarity of scientific papers, text-based and link-based similarity measures can be applied. The text-based similarity measures such as Cosine [11], Dice Coefficient [11], Kullback-Leibler Distance (KLD) [5, 1, 16], and BM25 [12, 17, 3] focus on the content of the papers but neglect the citation relationship between them. The link-based similarity measures such as SimRank [8], P-Rank, and rvs-SimRank [20] consider the citation relationship between scientific papers but ignore the content. Citations are selected manually by the authors according to the content, so the content and citations are interrelated in the scientific papers. However, the text-based and link-based similarity measures consider only on a single aspect of scientific papers to compute the similarity.

In this paper, we propose a new approach to compute the similarity of scientific papers accurately by combining the text-based and link-based similarity measures. The intuition behind our method is that, to improve the accuracy of similarity measures for scientific papers both the content and citations should be taken into account *simultaneously* because they are interrelated. In order to effectively combine the text-based and link-based similarity measures, we apply a weighted linear combination by using SVM^{rank} [9] which is based on the support vector machine. To precisely evaluate the effectiveness of our proposed method, we performed extensive experiments with a real-world dataset. Our experimental results show that the accuracy of the text-based and link-based similarity measures improves dramatically by using our combination method.

The rest of the paper is organized as follows. Section 2 discusses the text-based and link-based similarity measures and their working mechanism. In Section 3, we explain our proposed method. In Section 4, we present our dataset and analyze our experimental results. Section 5 concludes the paper.

RACS'13, October 1–4, 2013, Montreal, QC, Canada.

2. RELATED WORKS

2.1 Text-based Similarity Measures
Text-based similarity measures look at a paper as a bunch of terms and two papers are more similar if they have more common terms. In the literature, there are various types of the text-based similarity measures. Cosine [11] and Dice Coefficient [11] are based on the vector space model [13]. In the vector space model, every paper is represented as a vector of index terms and each dimension belongs to a term of the paper that denotes its weight [18]. For every term, the weight calculated as TF/IDF, the product of term frequency (TF) and inverse document frequency (IDF). TF is the number of times that a term appears in a paper. IDF is a measure that indicates how much a special term is common among all the papers in the dataset. IDF is calculated by taking logarithm of dividing the total number of papers by the number of those papers that contain the term. In other words, the high TF/IDF value for term t in paper p indicates that, term t is not common among so many papers in a dataset but has high redundancy in paper p, so term t is a good characteristic of paper p.

Cosine computes the similarity between two papers as the cosine of the angle between their vectors as follows:

$$\mathrm{Cos}(p_1, p_2) = \frac{\sum_{t \in (p_1 \cap p_2)} (w_{t,p_1} . w_{t,p_2})}{\sqrt{\sum_{t \in p_1} (w_{t,p_1})^2 \sum_{t \in p_2} (w_{t,p_2})^2}}, \quad (1)$$

$w_{t,p}$ is the weight of term *t* in paper *p*.

Dice Coefficient computes the similarity between two papers by considering their vectors as follows:

$$Dice\ (p_1, p_2) = \frac{2 \times \sum_{t \in (p_1 \cap p_2)} (w_{t,p_1} . w_{t,p_2})}{\sum_{t \in p_1} (w_{t,p_1})^2 + \sum_{t \in p_2} (w_{t,p_2})^2}, \quad (2)$$

Kullback-Leibler Distance (KLD) [5, 1, 16] and BM25 [12, 17, 3] are based on the probabilistic models [6]. In probabilistic models, each term appears in a paper according to a probability value. The basic assumption is that terms are distributed differently within relevant and irrelevant papers. KLD looks at the paper as a probability distribution over all the terms in the dataset and the distance between two papers is computed according to these probabilities.

$$KLD(Q, R) = \sum_{t \in V} (P(t|Q) - P(t|R)) \, log \frac{P(t|Q)}{P(t|R)}, \quad (3)$$

V is our term set. *Q* is the query paper. If term *t* does not appear in paper *R*, we should smooth paper *R* by assigning a non-zero value as the probability value of term *t* because without smoothing the value of $P(t, R)$ is zero and therefore value of $KLD(Q, R)$ would be infinity.

$$P(t|R) = \begin{cases} \beta * P(t|R) & if\ t\ occures\ in\ R \\ \varepsilon & otherwise \end{cases}, \quad (4)$$

ε is the smoothing value and could be set to a value less than the minimum existing value of $P(t|R)$. β is a normalization coefficient that has different values according to the size of the paper.

$$P(t|R) = \frac{tf(t, R)}{\sum_{t \in P} tf(t, R)}, \quad (5)$$

$$\beta = 1 - \sum_{t \in V, t \notin R} \varepsilon \ , \quad (6)$$

$tf(t, R)$ is the frequency of term *t* in paper *R*. Notice that instead of similarity, the $KLD(Q, R)$ calculates the *distance* between two papers.

BM25 puts more emphasis on the TF value and the paper length to compute the similarity between papers.

$$BM25(p, q) = \sum_{t \in q} idf_t \, . \alpha_{t,p} \, . \beta_{t,q}, \quad (7)$$

$$\alpha_{t,p} = \frac{(K_1 + 1) tf_{t,p}}{K_1 \left((1 - b) + b . \frac{|p|}{avg(L)} \right) + tf_{t,p}}, \quad (8)$$

idf_t is the inverse document frequency of term *t*, $|p|$ is the size of paper *p*, $avg(L)$ is the average size of the dataset, and $tf_{t,p}$ is the term frequency of term *t* in paper *p*. $K_1 \geq 0$ and $0 \leq b \leq 1$ are calibration values. We set $K_1 = 1.2$ and $b = 0.75$ according to [3].

$$\beta_{t,q} = \frac{(K_3 + 1) tf_{t,q}}{K_3 + tf_{t,q}}, \quad (9)$$

β normalizes the $tf_{t,q}$ of the query paper. We set $K_3 = 2$ same as [3] because in our study, the queries are almost large text (combination of title and abstract).

The text-based similarity measures use only the content of the papers to compute the similarity. Of course, content is important information that can be used to find similar papers. However, it cannot represent the authority of the paper. According to the content, two papers can be similar, even if one of them is much more authoritative than other one. Hence, they are not appropriate to compute the similarity of scientific papers because they ignore the citation relationship between the papers.

2.2 Link-based Similarity Measures

The link-based similarity measures compute the similarity of scientific papers by considering the citation graph where nodes represent the papers and edges denote the citation relationship between the papers. In the literature, there are various types of link-based similarity measures such as Co-citation [14], Coupling [10], Amsler [2], SimRank [8], P-Rank and rvs-SimRank [20]. Coupling takes into account only out-links, so the similarity between two papers is based on the number of papers that are *directly* cited by both of them. Unlike Coupling, Co-citation considers only in-links, and similarity between two papers depends on the number of papers that *directly* cite both. To compute the similarity, Amsler combines the similarity scores obtained by Co-citation and Coupling. In other words, Amsler considers the number of papers that are directly cited by both papers and the number of papers that directly cite both of them.

SimRank considers the in-links *recursively* and neglects the out-links, and similarity between two papers is based only on the number of papers that cite both of them. SimRank improves the accuracy of Co-citation. P-Rank considers both the in-links and out-links *recursively* to compute similarity, so the similarity between two papers is based on the number of papers that cite both of them and are cited by both. P-Rank improves the accuracy of Amsler. In addition, P-Rank provides a unified formulation that other link-based similarity measures such as Co-citation, Coupling, Amsler, and SimRank are its special cases. rvs-SimRank is another special case of P-Rank that computes the similarity between papers only based on the out-links recursively, so the similarity between two papers is based only on the number of papers that are cited by both of them. However, SimRank, rvs-SimRank, and P-Rank are the most famous link-based similarity measures in the literature.

The P-Rank unified formulation is as follows:

$$R_0(p, q) = \begin{cases} 0 & if\ (p \neq q) \\ 1 & if (p = q) \end{cases} \quad (10)$$

$$R_{k+1}(p, q) = \lambda \times \frac{C}{|I(p)||I(q)|} \sum_{i=1}^{|I(p)|} \sum_{j=1}^{|I(q)|} R_k \left(I_i(p), I_j(q) \right) + (1 - \lambda) \times \frac{C}{|O(p)||O(q)|} \sum_{i=1}^{|O(p)|} \sum_{j=1}^{|O(q)|} R_k \left(O_i(p), O_j(q) \right),$$

$R_k(p, q)$ denotes the P-Rank score between papers *p* and *q* on the iteration *k*. *I(p)* is the set of the papers that cited paper *p*. *O(p)* is the set of the papers are cited by *p*, $0 \leq \lambda \leq 1$ is the weighting parameter, and $0 \leq C \leq 1$ is a damping factor. If λ=1, Equation (10) equals to SimRank, and if λ=0, it equals to rvs-SimRank.

All of the link-based similarity measures analyze the citation graph to compute the similarity between scientific papers by ignoring the content of the papers. However, citations cannot clearly represent the content of a paper. Like text-based similarity

measures, we argue that link-based similarity measures are not appropriate to compute the similarity of scientific papers.

3. OUR PROPOSED METHOD

Scientific papers contain two interrelated information: content and citations. Therefore, to improve the accuracy of similarity measures both of them should be taken into account. The content of a paper represents its context that can be used to find similar papers. All the text-based similarity measures use the content to compute the similarity of the scientific papers. On the other hand, the citations are selected manually by the authors as a set of the related and authoritative scientific works. Also, the number of citations to a paper indicates its authority [15]. Hence, citation relationship can also be used to compute the similarity of the scientific papers. All the link-based similarity measures focus on the citation relationship between the papers to compute the similarity. However, both the text-based and link-based similarity measures consider only one aspect of the scientific papers.

We propose a new method to *effectively* compute the similarity of the scientific papers by *combining* the text-based and link-based similarity measures. We *simultaneously* consider both the content and citations to compute the similarity of scientific papers. First, we compute the similarity between two papers by applying the text-based and link-based similarity measures *separately*. Then, to effectively compute the similarity between them, we combine the similarity scores obtained by the text-based and link-based similarity measures according to a weighted linear combination as follows:

$$Sim(p,q) = w_1 Text\_Sim(p,q) + w_2 Link\_Sim(p,q), \qquad (11)$$

where $Text\_Sim(p,q)$ denotes the similarity score of papers p and q that computed by a text-based similarity measure; and $Link\_Sim(p,q)$ denotes the similarity score of papers p and q which is calculated by a link-based similarity measure. w_i is a weighting value. Instead of considering each similarity score as an *equally significant* factor to compute the similarity between p and q, we consider a *weight* as the *degree of importance* for each text-based and link-based similarity score to combine.

In order to indicate the weights, we use SVM^{rank} [9] which is based on the support vector machine (SVM). By using SVM^{rank}, we can automatically calculate the *optimal* values of w_1 and w_2 in Equation (11). SVM^{rank} is a free software and publicly available for noncommercial use[1], has high accuracy, and performs the training phase very quickly [4]. SVM^{rank} is a SVM based algorithm for predicting multivariate or structured outputs. According to [4], SVM^{rank} solves a maximum-margin optimization problem by finding a hyperplane. In our case, the hyperplane is a vector of weights that is represented as $W = \langle w_1, w_2 \rangle$ and provides an ideal separation between relevant and irrelevant papers in training set by finding the optimal values of w_1 and w_2. The training set is used for SVM^{rank} training process and contains training instances. Each training instance represents a paper in regarding to a query as follows:

$$\langle r, qid, text\_sim(p,q), link\_sim(p,q) \rangle, \qquad (12)$$

where r is either "0" or "1" that indicates whether paper p is relevant or irrelevant to query q, qid is the query number, $text\_sim(p,q)$ is the similarity score of p and q which is computed by a text-based similarity measure , $link\_sim(p,q)$ is the similarity score of p and q which is computed by a link-based similarity measure, and $0 \le text\_sim(p,q), link\_sim(p,q) \le 1$. SVM^{rank} indicates the hyperplane according to pairwise preference constraints [9]. The preference constraint is included for all pairs of training instances in the training set that have different value of r and same value of qid.

4. EXPERIMENTS

4.1 Dataset

We crawled information of 1,071,973 papers from DBLP and obtained their citation information from MS Academic Search. In order to make a precise ground truth, we used a famous data mining textbook [7]. We considered all the papers in the bibliographic section of every chapter. Finally, we selected 11 research topics from different chapters whose related papers exist in our dataset: data processing, mining frequent patterns and association rules, classification, clustering, mining data streams, link mining, graph mining, data cubes, spatial database, OLAP and data warehouse, and web mining. Therefore, our ground truth contains 11 sets and for each paper in our dataset, we only have title, abstract, and citation information. We do not have access to the body of the papers because of the copyright issue. However, according to [19], among title, abstract, and body, the combination of title and abstract is more appropriate to compute the similarity of scientific papers.

4.2 Results and Analyses

To evaluate the effectiveness of our proposed method and analyze the experimental results, we used MAP, precision at 10 top results (P@10), and recall at 10 top results (R@10) as our evaluation measures. We tried to find a link-based similarity measure that has the best accuracy with our real-world dataset. P-Rank, SimRank, and rvs-SimRank compute the similarity of scientific papers by traversing the citation graph recursively. Therefore, for each similarity measure we tried to indicate the best iteration that makes more accurate results in terms of MAP, precision and recall.

Figure 1 shows the accuracy of SimRank in the different iterations. SimRank had the best accuracy in the iteration 1 (SimRank@1). The difference between accuracy of SimRank in the iteration 3 to 5 was not so tangible; precision and recall had exactly the same values. However, in terms of MAP, the accuracy in the iteration 4 is worse than iterations 3 and 5.

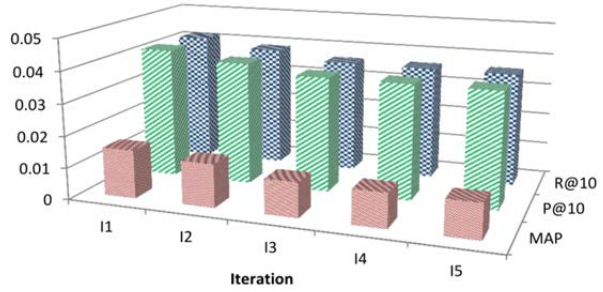

Figure 1: Accuracy of SimRank in different iterations.

Figure 2 shows the accuracy of rvs-SimRank in terms of MAP, precision, and recall in the different iterations. rvs-SimRank had the best accuracy in the iteration 2 (rvs-SimRank@2). The accuracy of rvs-SimRank for iterations 3, 4, and 5 in terms of precision and recall are exactly same. However, the iteration 4 has better accuracy in terms of MAP.

[1] http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

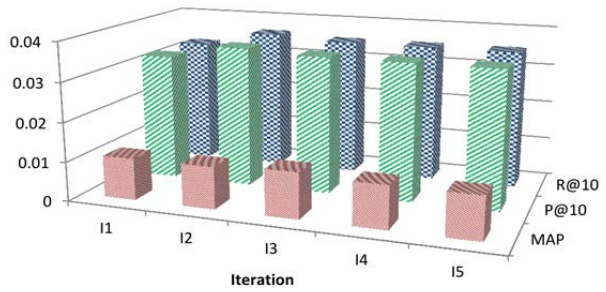

Figure 2: Accuracy of rvs-SimRank in different iterations.

Figure 3 shows the accuracy of P-Rank in terms of MAP, precision, and recall in the different iterations. P-Rank had the best accuracy in the iteration 5 (P-Rank@5). Unlike SimRank and rvs-SimRank, the accuracy of P-Rank was improved by traversing the citation graph more deeply.

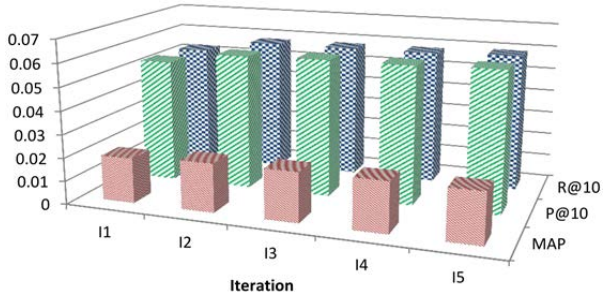

Figure 3: Accuracy of P-Rank in different iterations.

In order to select the best link-based similarity measure, we compared the accuracy of P-Rank, SimRank and rvs-SimRank in their optimal iteration. Figure 4 shows the comparison of SimRank@1, rvs-SimRank@2, and P-Rank@5 accuracy.

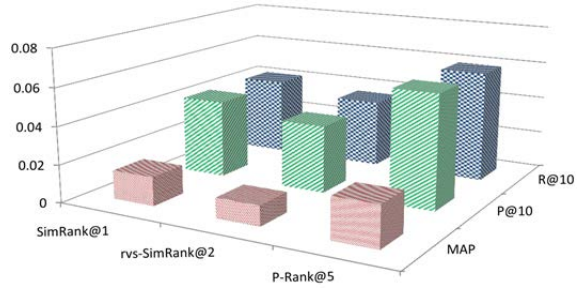

Figure 4: Accuracy of SimRank@1, rvs-SimRank@2, and P-Rank@5.

P-Rank outperformed the SimRank and rvs-SimRank in terms of MAP, precision, and recall. The reason is clear because the P-Rank considers both the citations to a paper and its references to the other papers simultaneously to compute the similarity. Therefore, as the best link-best similarity measure in our real-world dataset, we selected P-Rank in the iteration 5 to combine with the text-based similarity measures.

Figure 5 shows the results of P-Rank, text-based similarity measures based on the vector space model, and their weighted linear combinations according to our proposed method (Cosine+P-Rank, Dice+P-Rank). P-Rank had the worse accuracy in the terms of MAP, precision, and recall in comparing with the text-based similarity measures. The reason is that, P-Rank focuses

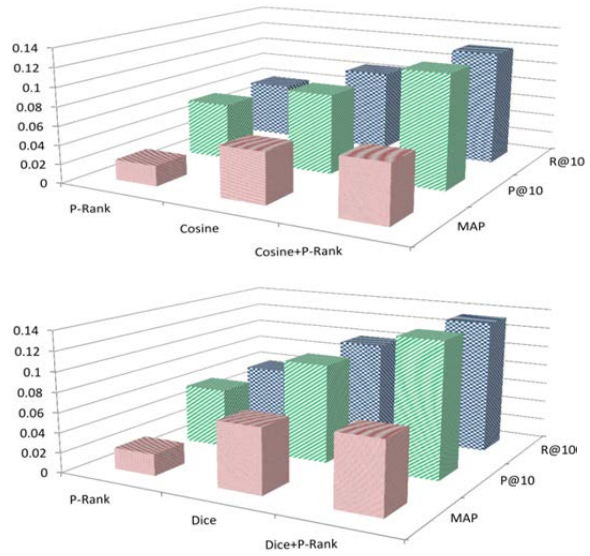

Figure 5: Accuracy of Cosine, Dice, P-Rank, and their weighted linear combination.

only on the citation relationship of the scientific papers and analyzes the citation graph to compute the similarity. The citations are selected manually by the authors as a set of the related and authoritative scientific works; however, they cannot represent the context of the paper clearly. The citations denote a collection of the papers that are related to a paper but *partially*. In other words, a paper contains citations and each citation is related only to a special part of the paper, not whole of the content. Cosine and Dice had the better accuracy in terms of MAP, precision and recall rather P-Rank. The reason is that, the text-based similarity measures consider the content of the scientific papers to compute the similarity. The content of a paper is important information and represents the context of the paper more clearly than citation information. Therefore, the text-based similarity measures are more accurate than link-based similarity measures to compute the similarity of scientific papers.

Our proposed method absolutely outperformed the Cosine, Dice, and P-Rank in terms of MAP, precision, and recall. The text-based similarity measures consider only the content of the scientific papers to compute the similarity and neglect the citation information between them. On the other hand, the link-based similarity measures consider only the citation relationship between papers and analyze the citation graph to compute the similarity of scientific papers by ignoring their content. However, scientific papers contain the content and citations which are interrelated and cannot be supposed isolated, so in order to compute the similarity accurately, both of them should be taken into account. Our proposed method considers the content and citations of scientific papers simultaneously to compute the similarity by linearly combining the results of the text-based similarity measures and P-Rank according to an *optimal* weighing scheme. Our proposed method improved the accuracy of Cosine and Dice around 31% and 24%, respectively.

Figure 6 shows the results of P-Rank, text-based similarity measures based on the probabilistic models, and their weighted linear combination according to our proposed method (BM25+P-Rank, KLD+P-Rank). Figures 5 and 6 demonstrate the same results. BM25 and KLD outperformed P-Rank and our proposed method outperformed all of them in terms of MAP, precision, and recall with the same reasons that we mentioned in analyzing Fig-

ure 5. Our proposed method improved the accuracy of BM25 and KLD around 20% and 25%, respectively.

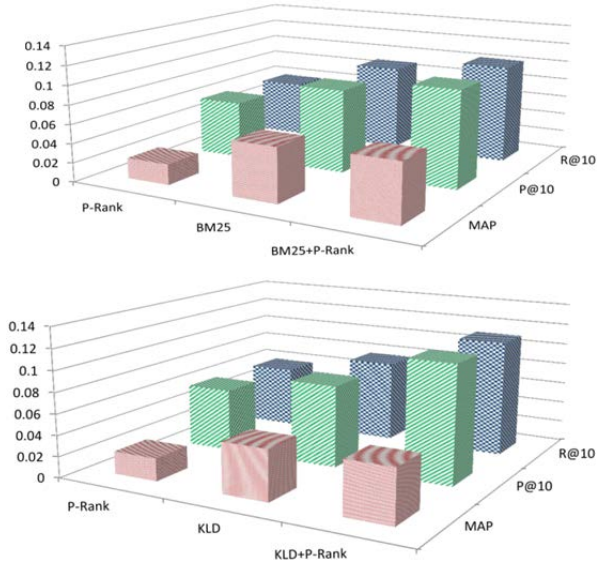

Figure 6: Accuracy of BM25, KLD, P-Rank, and their weighted linear combination.

5. CONCLUSIONS

In this paper, we proposed a new method to compute the similarity of scientific papers accurately by considering the content and citations simultaneously. Text-based and link-based similarity measures consider only one aspect of the scientific papers, content or citations, respectively. Our proposed method combines the similarity scores computed by the text-based and link-based similarity measures according to a weighted linear combination. To perform the weighted linear combination, we used SVM^{rank} which is based on the support vector machine. Our extensive experimental results on a real-world dataset of scientific papers show that our proposed method improves the accuracy of similarity measures for scientific papers dramatically.

6. ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2012R1A1A2007817).

7. REFERENCES

[1] A. Aktolga, I. Ros, and Y. Assogba. Detecting Outlier Sections in US Congressional Legislation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 235-244, 2011.

[2] R. Amsler. *Application of Citation-based Automatic Classification*. Technical report, The University of Texas at Austin Linguistics Research Center, 1972.

[3] A. Baron, A. Eiselt, and P. Rosso. Monolingual Text Similarity Measures: A Comparison of Models over Wikipedia Articles Revisions, In *Proceedings of the 25th European Conference on IR Research*, pages 305-319, 2003.

[4] F. Belem, E. Martins, T. Pontes, J. Almeida, and M. Goncalves. Associative Tag Recommendation Exploiting Multiple Textual Features, In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1033-1042, 2011.

[5] B. Bigi. Using Kullback-Leibler Distance for Text Categorization. In *Proceedings of the 25th European Conference on IR Research*, pages 305-319, 2003.

[6] N. Fuhr. Probabilistic Models in Information Retrieval, *J. The Computer*, Vol.35, No.3, pages 243–255, Dec. 1992.

[7] J. Han, and M. Kamber. *Data Mining: Concepts and Techniques*, Second Edition, Morgan Kaufmann, San Francisco, 2006.

[8] J. Jeh, and J. Widom. SimRank: A Measure of Structural-Context Similarity. In *Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 538-543, 2002.

[9] T. Joachims. Optimizing Search Engines using Clickthrough Data. In *Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pages 133-142, 2002.

[10] M. Kessler. Bibliographic Coupling between Scientific Papers, *J. The American Documentation*, Vol. 14, No. 1, pages 10-25, 1963.

[11] D. Lin. An Information-Theoretic Definition of Similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296-304, 1998.

[12] Y. Lv, and C. Zhai. When Documents are very Long, BM25 Fails. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1103-1104, 2011.

[13] G. Salton, and M. E. Lesk. Computer Evaluation of Indexing and Text Processing, *J. ACM*, Vol.15, No.1, pages 8–36, Jan. 1968.

[14] H. Small. Co-citation in the Scientific Literature: A New Measure of the Relationship between Two Documents, *J. The American Society for Information Science*, Vol. 24, No. 4, pages 265-269, 1973.

[15] T. Strohman, W. Croft, and D. Jensen. Recommending Citations for Academic Papers, In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, page 705-706, 2007.

[16] B. Tan, X. Shen, and C. Zhai. Mining Long-Term Search History to Improve Search Accuracy. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 718-723, 2006.

[17] X. Wan. A Novel Document Similarity Measure Based on Earthmover's Distance, *J. Information Sciences*, Vol.177, No.18, pages 3718-3730, Sep. 2007.

[18] R.B. Yates, and B.R. Neto. *Modern Information Retrieval*, Addison Wesley, Boston, 1999.

[19] S. Yoon, S. Kim, and J. Kim. On Computing Text-based Similarity in Scientific Literature, In *the Proceedings of the 20th International Conference Companion on World Wide Web, WWW,* pages 169-170, 2011.

[20] P. Zhao, H. Han, and S. Yizhou. P-Rank: a Comprehensive Structural Similarity Measure over Information Networks. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 553-562, 2009.