IEICE TRANS. INF. & SYST., VOL.E95–D, NO.10 OCTOBER 2012

LETTER

TL-Rank: A Blend of Text and Link Information for Measuring Similarity in Scientific Literature Databases

Seok-Ho YOON[†], ***Nonmember***, **Ji-Su KIM**[†], ***Member***, **Sang-Wook KIM**[†a)], ***and*** **Choonhwa LEE**[†], ***Nonmembers***

SUMMARY This paper presents a novel similarity measure that computes similarity scores among scientific research papers. The text of a given paper in online scientific literature is often found to be incomplete in terms of its potential to be compared with others, which likely leads to inaccurate results. Our solution to this problem makes use of both text and link information of a paper in question for similarity scores in that the comparison text of the paper is strengthened by adding that of papers related to it. More accurate similarity scores can be computed by reinforcing the input with the citations of the paper as well as the citations included within the paper. The efficacy of the proposed measure is validated through our extensive performance evaluation study which demonstrates a substantial gain.

key words: *similarity score, text-based measure, link-based measure, keyword set expansion*

1. Introduction

Recently, significant research efforts have been put into the analysis of data from online literature for use in search, retrieval, and recommendation. One of the most fundamental issues in the area is how to compute similarity among papers, which is used as a key building block to implement advanced features such as clustering, recommendation, and ranking [1]. A range of current similarity measures can be categorized into two classes: text-based and link-based measures. Text-based similarity measures are basically gauged based on the number of terms in common between two papers, while link-based ones look at how many common citations are shared by them.

A paper is typically composed of three parts: title, abstract, and body. A similarity score between a pair of papers could vary when computed against different parts. At first, we investigate which part best reflects the actual similarity between two papers. We also examine what weights should be assigned to each one, if more than one part are used in the computation.

Popular literature retrieval services, including CiteSeer, Google Scholar, and MS Libra, provide a paper's title, abstract, and body in the form of text by crawling research papers from the Internet. However, they usually do not provide the full text of the paper body due to concern over copyright infringement. Also, it is not uncommon for the abstract to be missed because of crawling and parsing difficulties. This lack of information from the papers causes the low accuracy of similarity scores of text-based measures. To deal with this problem, we propose a novel similarity measure, named TL-Rank, that takes into account inter-paper citation information as well as text information.

The rest of the paper is organized as follows. In Sect. 2, we review prominent approaches to the similarity computation problem. Section 3 presents our text-based similarity study which advocates a weighted combination of multiple parts of a paper. Two versions of our TL-Rank scheme are proposed in Sect. 4, and then evaluated in a comparative performance study. Section 5 summarizes and concludes the paper.

2. Related Work

The basic idea of text-based similarity measures is that a similarity score between two documents can be estimated by comparing words occurring in the documents; in general, the more common words the two documents share, the more similar they are. There exists a range of text-based similarity measures: a Boolean model [8], probabilistic model [3]–[5], and vector model [6], [7]. Our work presented in this paper is based on the vector model. The word set from a document is described by a vector where each dimension represents the frequency of a corresponding word in the set. In general, the vector model employs TF-IDF which assigns a larger weight to less frequent words and a smaller weight to more frequent words for more accurate similarity computation [2]. Similarity between two documents A and B is represented by a similarity score between their corresponding vectors computed by Eq. (1) below.

$$S(A,B) = \frac{A \cdot B}{\| A \| \| B \|} = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (1)$$

Link-based similarity measures exploit citation relationships among documents for similarity computation. The rationale behind link-based measures is that the more citations in common, the more similar two documents should be. Some representative link-based algorithms include Bibliographic coupling [9], Co-citation [10], Amsler [11], rvs-SimRank [12], SimRank [13], and P-Rank [12]. Bibliographic coupling computes a similarity score of two documents A and B based on the number of papers commonly cited by A and B. In contrast, Co-citation looks at the number of documents that cite both of A and B. Amsler uses a weighted sum of Bibliographic coupling's and Co-citation's

Manuscript received May 28, 2012.

[†]The authors are with the Department of Electronics and Computer Engineering, Hanyang University, 222 Wangsimni-ro, Seongdong-gu, Seoul, 133–791, Korea.

a) E-mail: wook@hanyang.ac.kr

DOI: 10.1587/transinf.E95.D.2556

Table 1 Relationship among link-based similarity measures. (adopted from [12])

Links used k	In-link	Out-link	Both
k = 1	Co-citation C = 1, λ = 1	Bibliographic Coupling C = 1, λ = 0	Amsler C = 1, λ = 1/2
k = ∞	SimRank C = varies, λ = 1	rvs-SimRank C = varies, λ = 0	P-Rank C, λ = varies

similarity scores. Expanding Bibliographic coupling, rvs-SimRank recursively follows citation relationships while computing similarity scores. In a similar way, SimRank is an expansion of Co-citation, while P-Rank can be viewed as a recursive version of Amsler.

These link-based schemes can be expressed and compared with each other by Eq. (2) and Table 1 [12]. In the equation, $R_k(A, B)$ denotes the similarity score between documents A and B at the k'th iteration, and C ($\in [0, 1]$) is a decay factor for attenuating the similarity score over the iterations. $I_i(A)$ denotes a set of the papers connected to A through the i-th in-link, while $O_i(A)$ represents a set of papers pointed to by the i-th out-link. Also, λ determines the relative influence of the in-linked papers against the out-linked ones. With $k = 1$, $C = 1$, and $\lambda = 1$, Eq. (2) represents Co-citation. With $\lambda = 0$, it becomes Bibliographic coupling. When $k = 1$, $C = 1$, and $\lambda = 0.5$, the equation represents Amsler. With k equal to ∞, it becomes SimRank, rvs-SimRank, or P-Rank, depending on the value of λ. However, it is known that the similarity scores by SimRank, rvs-SimRank, and P-Rank tend to converge at $k = 4$ or 5 [12], [13].

$$R_0(A, B) = \begin{cases} 0 & \text{if } A \neq B \\ 1 & \text{if } A = B \end{cases},$$

$$R_{k+1}(A, B) = \lambda \times \frac{C}{|I(A)||I(B)|} \sum_{i=1}^{|I(A)|} \sum_{j=1}^{|I(B)|} R_k(I_i(A), I_j(B))$$

$$+ (1 - \lambda) \times \frac{C}{|O(A)||O(B)|} \sum_{i=1}^{|O(A)|} \sum_{j=1}^{|O(B)|} R_k(O_i(A), O_j(B)) \quad (2)$$

3. Weighted Combination of Text-Based Similarity Scores

Before introducing our new similarity measure of TL-Rank in the following section, we examine the accuracy of text-based similarities when computed against different parts of a paper: title, abstract, and body. Findings from this investigation will be used as the basis of similarity computation in the next section. The vector model along with TF-IDF [14] has been employed for the similarity measure, and our dataset crawled from MS Libra comprises 1,071,973 papers and 2,473,636 citations within them.

We have selected 20 sections in a well-known textbook of data mining [1] and extracted 124 reference papers out of the sections. For each reference paper given as a query paper, we rank top m closest ones from the entire set of reference papers according to the similarity scores obtained with the title, abstract, and body, respectively. Then, the chosen m papers are checked to see how many of them belong to the same section as the query paper, which is the accuracy indicator we use to assess the effectiveness of similarity measures [15]. This is similar, in spirit, to the recall widely used in information retrieval research [14].

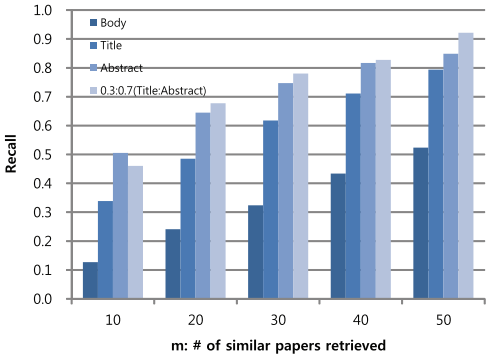

Fig. 1 Accuracy of similarities against different parts.

Figure 1 compares the recall of the results obtained using the different parts of the papers. The abstract comparison results in the highest accuracy, when a single part is used. It is surprising that the body comparison performs worse than any others. Even though the body includes richer information than the other two, it may also have a lot of general terms not directly related to the main issues dealt with by the paper. Therefore, the body is inappropriate to be used for the purpose of similarity computation. On the other hand, while the title contains key terms expressing the paper's essential content, still other important terms tend to be missed due to its conciseness. Our study reveals that, in most cases, the abstract provides just sufficient terms that cover the essence of the paper, leading to a better result than others.

We further observe that the accuracy of similarity scores can be enhanced by using multiple parts used in the computation. In addition to the single part cases, Fig. 1 also shows the case of a combination of title and abstract. A weight ratio of 0.3 : 0.7 for the title and abstract combination achieves the best result, outperforming the abstract-only case by around 5%. However, adding the body part worsens the accuracy, because it likely contains a large number of general terms, not directly related to the main issues of the paper. It is noteworthy that both title and abstract with weight ratio 0.3 : 0.7 have been used for all the similarity computation results presented from now on.

4. TL-Rank: A Similarity Term Expansion Scheme

4.1 Term Set Expansion

As shown in the previous section, the abstract provides the most important terms in computing text-based similarity scores among research papers. However, unfortunately, they are not always available due to crawling and parsing difficulties. It is not rare to see some crawled papers whose

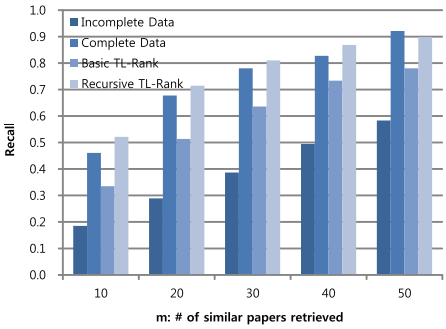

Fig. 2 Similarity accuracy vs. information completeness.

abstracts are unavailable to a certain extent or even completely. One of our analysis studies shows that crawled abstracts have 52% of the words on average, compared to their original. It is obvious that this loss of information should adversely affect the accuracy of similarity scores, which is confirmed by our results in Fig. 2. "Complete" in the graph represents the case where the missing terms of the abstracts have manually been filled in, while "Incomplete" is the as-is case. The graph demonstrates that the similarity accuracy may be improved by up to 50%, if the loss can somehow be made up. It is noted that this experiment was conducted in the same way as in the previous section.

When a paper is cited by another paper, we may expect the same or synonymous terms would appear in their titles and abstracts. Based on this expectation, we propose a new measure, called TL-Rank, that expands the term set of paper by adding the terms from the titles and abstracts of all the papers in citation relationship with that paper. For example, suppose that paper B cites paper C and is being cited by paper A. As the term set for paper B in similarity computation, we use not only the terms from B but also the terms out of A and C. Therefore, TL-Rank can be viewed as a similarity measure that makes use of both text and link information. The similarity computation is performed using the term set augmented by incorporating terms from related papers. This expansion results in a dramatic improvement over "Incomplete" case, as shown by TL-Rank in Fig. 2. It is noted that, for this experiment, we set the ratio of weights assigned to the terms from the original paper, cited papers, and citing papers to $1:1:1$. Also, the weight ratio for titles and abstracts is set to $0.3:0.7$ as in Fig. 1.

In addition to the quantitative results above, Table 2 assesses the three cases from a qualitative comparison perspective. Given one of the most prominent papers [16] in the area of data clustering as a query paper, we let each scheme select 10 papers that they believe are the most similar to the query paper. Boldface words in the table indicate keywords in the area, while the underlined are considered keywords for other areas. We can see that all the selections by "Complete" and TR-Rank cases include the word "clustering" (which is a keyword in the area of data clustering research.) In contrast, the result of "Incomplete" case is contaminated by keywords for other areas such as sequential pattern mining. Notice that there are cross-over research efforts by which some clustering algorithms use sequential pattern mining, and vice versa. This can cause confusion of similarity measures, when title and abstract texts are not sufficient enough for the computation.

Table 2 Qualitative evaluation of similarity schemes.

Rank	Incomplete	Complete	TL-Rank
1	Exploratory Mining and Pruning Optimizations of Constrained Association Rules	**Clustering** by pattern similarity in large data sets	Constraint-based **Clustering** in Large Databases
2	CLOSET: An Efficient Algorithm for Mining Frequent Closed Itemsets	OPTICS: Ordering Points To Identify the **Clustering** Structure	A Density-Based Algorithm for Discovering **Clusters** in Large Spatial Databases with Noise
3	Fast Algorithms for Projected **Clustering**	A Density-Based Algorithm for Discovering **Clusters** in Large Spatial Databases with Noise	Fast Algorithms for Projected **Clustering**
4	Automatic Subspace **Clustering** of High Dimensional Data for Data Mining Applications	CURE: An Efficient **Clustering** Algorithm for Large Databases	**Clustering** Through Decision Tree Construction
5	Mining Generalized Association Rules	Cure: An Efficient **Clustering** Algorithm for Large Databases	OPTICS: Ordering Points To Identify the **Clustering** Structure
6	Finding Interesting Rules from Large Sets of Discovered Association Rules	ROCK: A Robust **Clustering** Algorithm for Categorical Attributes	An Efficient Approach to **Clustering** in Large Multimedia Databases with Noise
7	What Makes Patterns Interesting in Knowledge Discovery Systems	An Efficient Approach to **Clustering** in Large Multimedia Databases with Noise	CACTUS - **Clustering** Categorical Data Using Summaries
8	A New Framework For Itemset Generation	Fast Algorithms for Projected **Clustering**	Scaling **Clustering** Algorithms to Large Databases
9	Knowledge Discovery in Large Spatial Databases: Focusing Techniques for Efficient Class Identification	Finding Generalized Projected **Clusters** In High Dimensional Spaces	CURE: An Efficient **Clustering** Algorithm for Large Databases
10	LOF: Identifying Density-Based Local Outliers	Scaling **Clustering** Algorithms to Large Databases	BIRCH: An Efficient Data **Clustering** Method for Very Large Databases

4.2 Recursive Expansion

As discussed above, TL-Rank tries to compensate for any loss in the original paper by importing terms from its citing and cited papers. However, the utility of it will be limited considerably, if the imports face the same problem of incomplete text. This problem can be dealt with effectively, if we apply the term expansion in a recursive fashion. In other words, the term set expansion is repeatedly performed over several hops in a citation relationship network. The idea is similar, in spirit, to the expansions of Bibliographic coupling, Co-citation, and Amsler to their respective recursive versions of rvs-SimRank, SimRank, and P-Rank. If k denotes the number of recursive expansion steps, then TL-Rank above corresponds to $k = 1$. Results being presented here is the case of k being set to 2. We have found that going further does not necessarily lead to a better result because of noisy information introduced by the expansion process. Different weights assigned to the term set from cited papers, a query paper, and citing papers affect resultant similarity scores. Varying the weight combination, we have found that the ratio of $2:1:1$ yields best result.

Figure 2 compares the results of "Incomplete", "Complete", and TL-Rank cases. Results from the recursive expansion scheme are referred to as Recursive TL-Rank in the graph, while the case of $k = 1$ is indicated simply as TL-Rank. The accuracy of Recursive TL-Rank increases substantially up to 3.3 times compared to that of the incomplete abstract case. Moreover, our Recursive TL-Rank shows the results comparable with the complete abstract case, demon-

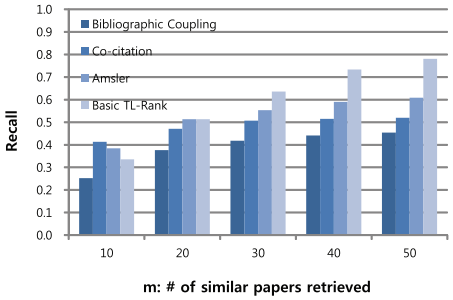

Fig. 3 Comparison with non-recursive link-based similarity measures.

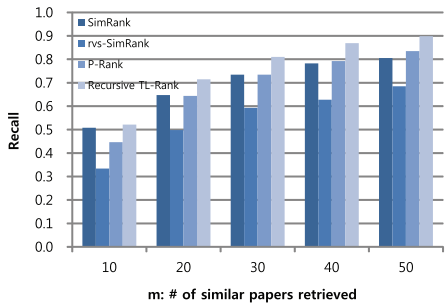

Fig. 4 Comparison with recursive link-based similarity measures.

strating that the scheme is able to adequately make up for incomplete paper information in scientific literature databases. Enrichment of similarity terms via recursive expansion leads to a higher recall for the result.

In order to demonstrate the effectiveness of our proposal, we compare its performance with those of representative link-based similarity measures. First, the non-recursive version of TL-Rank is compared with Bibliographic coupling, Co-citation, and Amsler. As shown in Fig. 3, it outperforms existing link-based similarity measures at best by 58% when top 50 papers are chosen.

Figure 4 depicts the similarity accuracy of our Recursive TL-Rank along with the current recursive link-based similarity measures. It is noted that, in general, recursive versions show better results than their corresponding non-recursive versions, which is attributed to the enrichment of paper similarity information by recursive expansions. As expected, the result demonstrates that our proposal is able to attain higher recall by benefiting from richer information reinforced by incorporating related papers' title and abstract rather than by using just citation relations. In summary, the novelty of our solution is that it effectively blends together text and link information for computing similarity scores of scientific research papers.

5. Conclusions

This paper first analyzes the accuracy of text-based similarity measures using a paper's different parts, such as title, abstract, and body, which reveals that a weighted combination of title and abstract yields the best result. Based on the findings, we further propose the TL-Rank scheme which merges the comparison term set of a paper in question with those of related papers being cited by or citing it. TL-Rank utilizes a blend of text and link information by considering text information from not only the target but also papers related to it. Our performance study demonstrates the efficacy of the proposal, surpassing current state-of-the-art techniques.

References

[1] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Second ed., Morgan Kaufmann, San Francisco, 2006.
[2] G. Salton, E.A. Fox, and H. Wu, "Extended Boolean information retrieval," Commun. ACM, vol.26, no.11, pp.1022–1036, Nov. 1983.
[3] N. Fuhr, "Probabilistic models in information retrieval," Comput. J., vol.35, no.3, pp.243–255, Dec. 1992.
[4] S.E. Robertson and K.S. Jones, "Relevance weighting of search terms," J. American Society for Information Sciences, vol.27, no.3, pp.129–146, May/June 1976.
[5] S.K.M. Wong and Y.Y. Yao, "On modeling information retrieval with probabilistic inference," ACM Trans. Information Systems, vol.13, no.1, pp.39–68, Jan. 1995.
[6] G. Salton and M.E. Lesk, "Computer evaluation of indexing and text processing," J. ACM, vol.15, no.1, pp.8–36, Jan. 1968.
[7] G. Salton, The SMART Retrieval System - Experiments in Automatic Document Processing, Prentice-Hall, New Jersey, 1971.
[8] G. Salton and M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.
[9] M. Kessler, "Bibliographic coupling between scientific papers," J. American Documentation, vol.14, no.1, pp.10–25, April 1963.
[10] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," J. American Society for Information Science, vol.24, no.4, pp.265–269, July/Aug. 1973.
[11] R. Amsler, "Application of citation-based automatic classification," Technical Report 72-14, The University of Texas at Austin Linguistics Research Center, 1972.
[12] P. Zhao, J. Han, and Y. Sun, "P-Rank: A comprehensive structural similarity measure over information networks," Proc. 18th ACM Int'l. Conf. on Information and Knowledge Management, pp.553–562, Hong Kong, China, Nov. 2009.
[13] G. Jeh and J. Widom, "SimRank: A measure of structural-context similarity," Proc. 8th ACM Int'l. Conf. on Knowledge Discovery and Data, pp.538–543, Alberta, Canada, July 2002.
[14] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, Addison Wesley, Boston, 1999.
[15] S. Yoon, S. Kim, and S. Park, "A link-based similarity measure for scientific literature," Proc. 19th Int'l Conf. on World Wide Web, pp.1213–1214, Raleigh, USA, April 2010.
[16] G. Karypis, R.H. Han, and V. Kumar, "CHAMELEON: A hierarchical clustering algorithm using dynamic modeling," Computer, vol.32, no.8, pp.68–75, Aug. 1999.