A Link-based Similarity Measure for Scientific Literature

Seok-Ho Yoon
Dept. of Electronics and Computer Engineering
Hanyang University
Seoul, 133-791, Korea
bogely@zion.hanyang.ac.kr

Sang-Wook Kim
Dept. of Electronics and Computer Engineering
Hanyang University
Seoul, 133-791, Korea
wook@hanyang.ac.kr

Sunju Park
School of Business
Yonsei University
Seoul, 120-749, Korea
boxenju@yonsei.ac.kr

ABSTRACT

In this paper, we propose a new approach to measure similarities among academic papers based on their references. Our similarity measure uses both in-link and out-link by transforming in-link and out-link into undirected links.

Categories and Subject Descriptors: I.5.3 [Clustering] Similarity measures

General Terms: Measurement, Reliability

Keywords: Scientific Literature, Link-based Similarity Measure

1. INTRODUCTION

As the volume of scientific literature grows fast, the demand for scientific literature retrieval service has steadily increased. One of the most popular retrieval services is to find a set of papers similar to the paper under consideration, which requires a measure that computes similarities among papers. In this paper, we point out the problems with the existing similarity measures and propose a new method for computing similarities.

2. RELATED WORK

Most prior research on similarity measures transforms the references in a paper into directed links and computes a similarity score between papers using link-based similarity measure. Typical link-based similarity measures include Bibliographic Coupling (Coupling) [1], Co-citation [2], Amsler [3], SimRank [4], rvs-SimRank, and P-Rank [5].

In Coupling, the similarity between two papers is computed based on the number of papers which are referenced by both of them (i.e, out-link) [1]. In Co-citation, the similarity between two papers is based on the number of papers that reference both papers (i.e., in-link) [2]. Amsler measures the similarity between two papers as a weighted sum of the similarity scores by Coupling and by Co-citation [3]. SimRank improves the accuracy of Coupling by computing the similarity score iteratively [4]. Rvs-SimRank and P-Rank improves Co-citation and Amsler, respectively [5].

Equation (1) describes the similarity measures mentioned above. I(a) (O(a)) denotes the set of in-link (out-link) neighbors of paper a. An individual in-link (out-link) neighbor is denoted as I_i(a) (O_i(a)). R_k(a,b) denotes the similarity score between paper a and paper b at iteration k. The relative weight of in-link and out-link is balanced by parameter $\lambda \in [0,1]$. C is a damping factor for in-link and out-link, where C $\in [0,1]$. Table 1 summarizes the existing similarity measures [5].

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

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

Table 1: Relationship among similarity measures [5]

Use of links	In-link	Out-link	Both
k=1	Cocitation C=1, λ=1	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

Scientific literature databases exhibit two interesting characteristics: (1) a paper can reference only the papers published before it (and can never reference the papers published after it) and (2) scientific literature databases often do not old papers. These two facts cause all existing similarity measures to fail in at least one of the following cases:

(P1) measuring the similarity between old, but similar papers

(P2) measuring the similarity between recent, but similar papers

(P3) measuring the similarity between two similar papers: one old, the other recent

Coupling computes the similarity score between two old but similar papers as near 0 in (P1), because there exist few papers that are referenced by both of them in the database. Similarly, co-citation computes the score as near 0 in (P2), because there exist few papers which reference both papers in the database. In (P3), because the old paper tends to have few out-links and the recent one tends to have few in-links, both Coupling and Co-citation would compute the scores as near 0 in (P3). Although the score by Amsler is not 0 in (P1) or (P2), it is an incorrect one. If the relative weights for Coupling and Co-citation are 0.5, respectively, for example, the maximum score by Amsler would be at most 0.5 in (P1) or (P2). Furthermore, Amsler computes the score as near 0 in (P3). SimRank, rvs-SimRank, and P-Rank are plagued with the same problems, since they are the iterative extensions of Coupling, Co-citation, and Amsler, respectively.

3. OUR APPROACH

Two papers A and B should be determined similar in the following three cases. First, A and B are similar if the number of papers referenced by both A and B (out-link) is high.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

Second, A and B are similar if the number of papers which reference both A and B (in-link) is high. These ideas are captured in Coupling and Co-citation, respectively. Note that, however, Coupling and Co-citation fail to capture similarity correctly in (P1) and (P2), respectively. Third, A and B are similar if many of the papers that are referenced by A reference B. The similarity score can be computed correctly in (P3), if one computes the score by counting the number of papers referenced by A that reference B (which we call 'passers'). Using a passer-based measure in (P1) or (P2) where few passers exist between two papers under consideration, however, would result in incorrect scores.

To compute the score correctly regardless of the published date of papers, therefore, one should employ all three measures–Coupling, Co-citation, and a passer-based measure. We achieve this by transforming both out-links and in-links into undirected links (i.e., disregarding the direction of references among papers) and then computing the similarity score between two papers based on the number of papers connected by both of them (which we call 'connectors'). This newly-proposed similarity measure combines all three measures properly.

Similar to that the accuracy of Co-citation is improved by SimRank through iteration, the proposed measure (which we call 'Inter-Connection') can be improved through iteration. Equation (2) represents Inter-Connection. Compared to Equation (1), both in-links and out-links are transformed into undirected links in Equation (2). L(a) denotes the set of undirected link neighbors of paper a.

$$R_{k+1}(a,b) = \frac{C}{|L(a)||L(b)|} \sum_{i=1}^{|L(a)|} \sum_{j=1}^{|L(b)|} R_k(L_i(a), L_j(b)) \quad (2)$$

Co-citation (SimRank) and Coupling (rvs-SimRank) compute the score between papers using either in-link or out-link but not both. Some may insist that Amsler (P-Rank) use both in-link and out-link [5], but since the score using in-link and the other using out-link are computed separately (and using a single type of links may result in incorrect scores as shown above), the weighted sum of two scores may be lower than what it should have been. By comparison, Inter-Connection measures the similarity between papers using in-links, out-links, and passers altogether at the same time.

One could have computed three scores using Coupling, Co-citation, and a passer-based measure and generated a weighted sum of them, but this would suffer the same problem faced by Amsler. That is, one of the scores may be near 0, which results in the score that is much lower than the correct value. Inter-Connection, however, has the effect of increasing the relative weight of Co-citation in the case of (P1), increasing the weight of Coupling in (P2), and increasing the weight of the passer-based measure in (P3). Inter-Connection therefore is a proper way to measure the similarity between papers, regardless of the difference in their publish dates.

4. EXPERIMENTS

Our experiments ran on about 1 million papers from DBLP[1] and reference information crawled from Libra[2]. In order to evaluate the accuracy of Inter-Connection, we did the following. First, we selected five well-known fields in data mining (sequential pattern mining, link mining, spatial database, web mining, and multi-relational data mining) and selected the reference papers at the end of each chapter for each field from a textbook [6]. The references included both old and recent papers. Second, we used one of the references to be a query paper and found the k highest scoring papers (where k can be 10, 20, 30, 40, 50) by each similarity measure. Third, we computed the precision of each similarity measure by comparing the k highest scoring papers to those in the reference list of the field of the query page. Forth, we repeated the second and third steps until all references were used as a query page.

Figure 1 shows the precision of each similarity measure. Due to space limitation, only the results of Coupling, Co-citation, Amsler, and Inter-Connection are shown. The precision of Inter-Connection is the highest, because compared to prior measures, Inter-Connection can compute the similarity scores properly in all (P1), (P2), and (P3) cases.

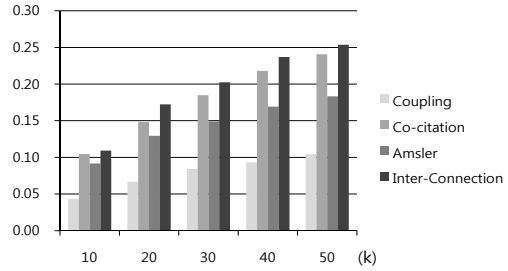

Figure 1: Comparison among similarity measures.

5. CONCLUSION

We propose Inter-Connection, a new link-based similarity measure for computing the similarity score between papers in a scientific literature database. Inter-Connection disregards the direction of references among papers by transforming in-link and out-link into undirected links.

Experimental results show the accuracy of Inter-Connection is higher than those of existing similarity measures.

6. ACKNOWLEDGMENTS

This work was supported by NHN Corp and partially by NRF (Grant No. 2008-0061006). Any opinions, findings, and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

7. REFERENCES

[1] M. Kessler, "Bibliographic Coupling Between Scientific Papers," *Journal of the American Documentation*, Vol. 14, No. 1, pp. 10-25, 1963.

[2] H. Small, "Co-citation in the 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.

[3] R. Amsler, *"Application of citation-based automatic classification. Technical report,"* The University of Texas at Austin Linguistics Research Center, 1972.

[4] G. Jeh and J. Widom, "SimRank: A Measure of Structural-Context Similarity," In *Proc. Int'l. Conf. on Special Interest Group on Knowledge Discovery and Data*, pp. 538-543, 2002.

[5] P. Zhao, J. Han, and Y. Sun, "P-Rank: a Comprehensive Structural Similarity Measure over Information Networks," In *Proc. Int'l. Conf. on Information and Knowledge Management*, pp. 553-562, 2009.

[6] J. Han and M. Kamber, *Data Mining: Concepts and Techniques(2nd Edition)*, Morgan Kaufmann, 2006.

[1] http://www.informatic.uni-trier.de/ley/db/setminus
[2] http://academic.research.microsoft.com/setminus