

Properties of Academic Paper References

Sunghun Kim, E. James Whitehead, Jr.

Dept. of Computer Science
Baskin Engineering
University of California, Santa Cruz
Santa Cruz, CA 95060 USA
{hunkim, ejw}@cs.ucsc.edu

ABSTRACT

We propose a new method to find related papers using an input paper and its hyperlinked citation relationships rather than keywords. Such related papers are especially useful as background reading for researchers new to a research field. In this paper we introduce the background reading paper extractor (BPE), and show various properties of academic paper references.

Categories and Subject Descriptors

H.3.7 [Information Storage and Retrieval]: Digital Libraries - collection; H.5.4 [Information Interfaces and Presentation]: Hypertext/Hypermedia - navigation.

General Terms

Algorithms, Measurement, Experimentation

Keywords

Hyperlink, Paper References

1. INTRODUCTION

Suppose you are a first-year graduate student in computer science, and one day your academic advisor gives you a hypertext paper to read. You have never read any papers in this area, and as a result you don't quite understand the paper. You decide to do some background reading first. But which papers should you read and how can you find them?

We may use general or academic-focused search engines to find papers. However, a newcomer may not know the best keyword to use and may have difficulty with commonly occurring keywords. One way to find background papers is to start from a paper in this research area, and follow the references. After reading the paper, its references are followed, and so on. After some number of iterations, we will be familiar with several papers or authors that show up many times. Those are likely to be important and good papers for background reading. But reading papers and following references requires significant effort. If there were a system that takes one paper as a seed and returns several useful background-reading papers, it would be useful. We developed a background reading paper extractor (BPE) by mining hyperlink citation relationships among papers.

The properties of paper hyperlink relationships are different from that of other hyperlink systems such as web, since the reasons for

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

HT'04, August 9-13, 2004, Santa Cruz, California, USA.
Copyright 2000 ACM 1-58113-000-0/00/0000...\$5.00.

citing papers are different from other hyperlink systems [1]. Understanding the properties of academic paper references, such as the number of unique references in the paper and the number of references to converge results, is the key factor to build reference-based data mining systems such as BPE. In this paper we describe BPE and the results of running BPE on papers on the hypertext literature.

2. PAPER EXTRACTOR

BPE uses a simple algorithm to get background-reading papers, inspired by the way we find papers by following references. BPE requires a seed paper and follows the references of the paper. BPE iterates the process until it gets enough papers. The ranks of papers are calculated based on the frequency of their appearance in other papers' references.

One of the problems of this algorithm is that it always gets older papers than the input paper, since it is following references. The backward reference data might improve the results, but following the reference is good enough to find background-reading papers.

BPE queries the ACM digital library [2] to get papers and their linked citation references. Unfortunately the ACM digital library does not have complete reference data. Also, papers earlier than 1981 are not indexed in the digital library. However, we believe the available papers and reference data are enough to provide useful background-reading papers. Table 1 shows several extracted papers using an input paper [3] from the hypertext literature. The result includes several classic hypertext papers, and it shows a reasonable result even though we used a simple ranking algorithm and incomplete reference data.

Table 1. Results from a hypertext paper input

Rank	Cited	Author, Title
1	59	Jeff Conklin, Hypertext: An Introduction and Survey
2	48	Frank G. Halasz et al., Notecards in a Nutshell
3	45	Frank G. Halasz, Reflections on NoteCards: Seven Issues for the Next Generation of Hypermedia Systems
4	32	Robert M. Akscyn et al., KMS: A Distributed Hypermedia System for Managing Knowledge in Organizations
5	30	Norman Delisle et al., Neptune: A Hypertext System for CAD Applications

3. PROPERTIES OF REFERENCES

We have several questions about properties of paper references. Does the algorithm ever find the complete graph of references? How many papers should we visit to make the ranking stabilize? How many unique references does each paper have? Do different input papers produce different results?

First we observed whether iterating through references ever terminates. (i.e., do we ever find a complete graph of references?) Figure 1 shows the number of unique references as BPE visits papers. We observe that visiting 2,000 papers is not sufficient to

terminate iteration, but after visiting approximately 400 papers, the number of unique references added by each new paper is constant, average of 3.9 references.

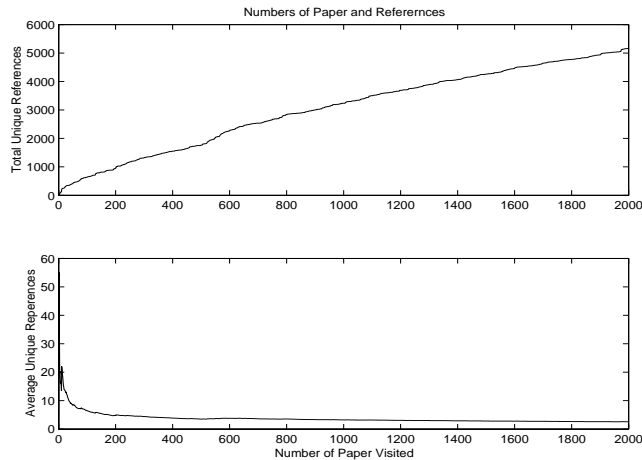


Figure 1. Unique reference numbers

But do we need a complete graph of references to get reasonable results? First, we may not need all 2,000 papers. Only the top ranked couple hundreds papers are enough. The next question is if the results converge. We fed an input paper as a seed, got the top 100 papers per iteration, and calculated change in rank (the rank delta) between the current and previous iterations. The rank delta for a single paper is the absolute value of the change in rank between current and previous iteration through BPE. We sum all of the rank differences for the top 100 most cited papers in a given BPE iteration, giving the Top 100 Rank Delta. Figure 2 shows the rank delta of results per iteration. It tells us that after having around 1,200 references, the results converge.

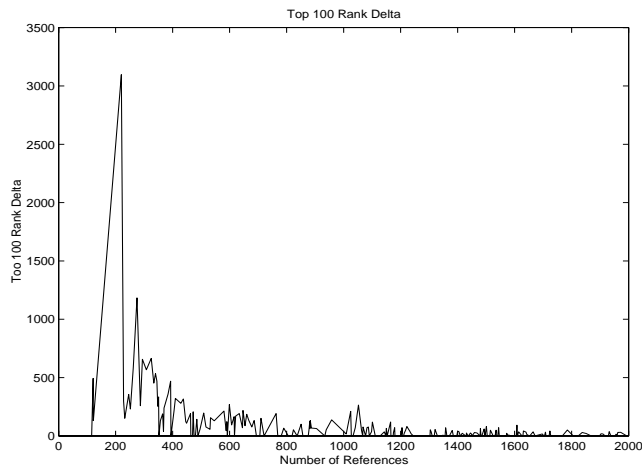


Figure 2. Rank Delta per iteration

Finally we studied if a seed paper affects the results. This question is important to see if a seed paper can decide the domain of output papers. We picked four papers from the Hypertext Conference 2003 in four different sections; Emergent web patterns (P1) [4], Hypermedia semantics (P2) [5], Link aggregation (P3) [6], and Hypermedia creation (P4) [7]. We picked one paper (P5) [8] from the Software Configuration Management domain. We iterated until we had 1,000 references for each paper and compared the Top 100 Rank Delta to see how

different the results are. The rank delta comparison is shown in Table 2. Even though the papers are from the same hypertext domain, the extracted papers from different sections are slightly different. The rank delta with P5 shows that papers from two different domains produce very different results. Overall we see that the extracted papers depend on the seed paper.

Table 2. Top 100 Rank Delta comparison with different inputs

	P1	P2	P3	P4	P5
P1	0	3743	2956	3767	4707
P2		0	3145	2600	4537
P3			0	3945	4427
P4				0	4663
P5					0

4. RELATED WORK

CiteSeer provides various statistics based on citations such as most cited documents, and most cited authors [9]. It also provides Computer Science Dictionary that is automatically created based on citation numbers. BPE provides papers lists based on your most interesting paper as an input. It is useful to get some related background papers without knowing exact keywords or categories.

5. CONCLUSIONS

We showed an algorithm to extract background-reading papers from a given paper. The result in Table 1 is satisfactory in spite of limitations of reference data and the algorithm. The paper reference properties show various results including that looking only 1,200 references is good enough to get converged results. Also we showed that the seed paper decides the domain of output papers.

To produce more accurate results, we need to find more sophisticated algorithms such as PageRank [10] for academic papers. Using backward reference data, citations can be used to get related papers as well.

6. REFERENCES

- [1] D. O. Case and G. M. Higgins, "How Can We Investigate Citation Behavior? A Study of Reasons for Citing Literature in Communication," *J. American Society for Information Science*, 51(7), pp. 635-645, 2000.
- [2] ACM, "ACM Digital Library," <http://portal.acm.org/dl.cfm>, 2004.
- [3] J. Whitehead, "As We Do Write: Hyper-terms for Hypertext," *ACM SIGWEB Newsletter*, vol. 9, no. 2-3, pp. 8 - 18, 2000.
- [4] A. A. Macedo, K. N. Truong, J. A. Camacho-Guerrero, and M. d. G. Pimentel, "Automatically Sharing Web Experiences through a Hyperdocument Recommender System," *Proc. Hypertext 2003*, pp. 48-56, Nottingham, UK, 2003.
- [5] C. C. Marshall and F. M. Shipman, "Which Semantic Web?," *Proc. Hypertext 2003*, pp. 57-66, Nottingham, UK, 2003.
- [6] N. Eiron and K. S. McCurley, "Untangling Compound Documents on the Web," *Proc. Hypertext 2003*, pp. 85-94, Nottingham, UK, 2003.
- [7] T. Miles-Board, Deveril, J. Lansdale, L. Carr, and W. Hall, "Decentering the Dancing Text: From Dance Intertext to Hypertext," *Proc. Hypertext 2003*, pp. 108-119, Nottingham, UK, 2003.
- [8] A. Sarma, Z. Noroozi, and A. v. d. Hoek, "Palantir: Raising Awareness Among Configuration Management Workspaces," *Proc. ICSE 2003*, pp. 444-454, Portland, Oregon, 2003.
- [9] "Computer and Information Science Papers CiteSeer Publications ResearchIndex," <http://citeseer.ist.psu.edu/cis>, 2004.
- [10] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," Stanford Digital Library Technologies Project 1998.