EndorSeer: An Add-on for Browsing Digital Libraries with "Endorsed" Citations

Mandar R. Mutalikdesai Srinath Srinivasa Viswanath Gangavaram

Open Systems Lab
International Institute of Information Technology
26/C, Electronics City
Bangalore 560100, India
mandar@iiitb.ac.in, sri@iiitb.ac.in, viswanath.gangavaram@iiitb.net

Abstract

Co-citations have long been used as a measure of topical relatedness between documents. However, we have observed some characteristic patterns in the way co-citations are formed in different data corpora. Based on these observations, we propose various interpretations of what a co-citation means. For the purposes of this demo, we concern ourselves only with one of our interpretations: *co-citations as citation endorsements*. This interpretation can be used by focused surfers to browse citations that are relevant to their topic of interest in digital libraries. In this demo, we present *EndorSeer*, a Firefox add-on for CiteSeer, which emphasizes the topical relevance of outgoing citations of a given document by ranking them based on their endorsements by third-party co-citations.

1 Introduction

In the Library and Information Sciences, citation analysis is used to study semantics implicit within clusters of documents, authors and journals. One such semantic is that a citation from a document A to another document B can be seen as an implicit recommendation of B by A [2, 3, 11, 12, 19].[1]

Two or more documents are said to be *co-cited* if there is at least one other document which cites – or recommends – all of them simultaneously. Co-citation has long been used as a measure of topical relatedness among articles and authors in scientific literature [15, 16, 18, 19, 20]. While this concept of topical relatedness can be used to explain co-citations in various datasets, there are some characteristic patterns of co-citations in webpages, scientific literature and collaborative hypertext systems like Wikipedia, which may help us understand the nature of co-citations better. We have, therefore, proposed different interpretations of what a co-citation may mean [10]. In this demo, we concern ourselves only with the interpretation of a co-citation as a *citation endorsement*. We discuss this interpretation in detail below.

In scientific literature (in the field of Particle Physics) as well as on the Web, with high probability, two documents that are highly co-cited are also known to have a direct citation between them [14, 15]. However, on Wikipedia, two highly co-cited pages are known to be connected to each other via an intermediary page [14]. Therefore, we propose the following interpretation of a co-citation as being more pertinent to scientific literature and to the Web compared to Wikipedia.

Let us assume that, initially, there existed a citation from a document A to another document B, which was topically very relevant. A large number of users traversed this citation, and ended up creating their own documents on a similar topic, citing both A and B. This is akin to the "copying" model for the Web graph [7]. We speculate that this model applies to scientific literature as well. In this sense, the co-citations can be seen as an endorsement of the citation from A to B [10]. Suppose A contained outgoing citations to a number of documents, but among those, if B alone has been highly co-cited with A, we can conclude that the citation from A to B is more "important" than the rest of the citations of A. We refer to citations such as the one from A to B as *endorsed citations*.

In this demo, we propose to showcase an online tool called *EndorSeer* (available as a Mozilla Firefox[2] add-on), for browsing digital library corpora such as CiteSeer[3] using endorsed citations. The add-on can be downloaded for non-commercial usage from `http://tinyurl.com/endorseer`.

15th International Conference on Management of Data
COMAD 2009, Mysore, India, December 9–12, 2009
©Computer Society of India, 2009

[1] Assuming that nepotistic citations are filtered out.

[2] `http://www.mozilla.com/en-US/firefox/`

[3] CiteSeer is a digital library of Computer and Information Sciences literature (`http://citeseer.ist.psu.edu/`).

2 Related Work

Co-citations have long been used in scientific literature to discover clusters of related articles and authors. Small [15] found that a high degree of co-citation is a better indicator of topical relatedness than bibliographic coupling. Small [16] also studied changes in the structure of co-citation graphs of scientific literature to draw interpretations about the growth of a topic of study. White and Griffith [18] studied author co-citation graphs to analyze clusters of authors with similar interests. Zhao [20] analyzed author co-citations by considering the first five authors of a cited paper, in contrast to some of the traditional methods, where only the first author of a paper is considered in the co-citation graph.

Co-citations have also been analyzed in the context of the Web to discover pages with related content. Dean and Henzinger [3] proposed (and Davison [2] endorsed) that two pages are related if they are highly co-cited. Hou and Zhang [5] also used co-citations to find semantically relevant pages. Reddy and Kitsuregawa [13] used co-citations to discover Web communities. Hyperlink-Induced Topic Search (HITS) [6] utilized the bipartite structures at the core of Web communities to determine good hub pages and authority pages pertinent to a given query. These bipartite cores correspond to co-citations of authorities by hubs. Efron [4] used co-citations to determine the political orientation of webpages. Thelwall and Wilkinson [17] used co-citations along with bibliographic couplings and direct citations to find similar websites within the UK academic Web. Larson [8] used co-citations as a measure of relatedness, and visualized clusters of pages on various topics using multi-dimensional scaling. Moise, et al. [9] proposed the idea of "focused co-citation." They argued that due to the presence of several webpages with no particular topical focus, just counting the number of co-citations between pairs of pages is not a good enough measure of relatedness. They proposed that given a page A, any other page B that is co-cited with A should contribute to the topical focus of A proportionally to the joint probability of co-citation of A and B.

In comparison to existing literature on co-citation analysis, we look towards using co-citations as a distinguishing feature for outgoing citations in a given document. The "distinguished" citations can then be used for focused resource discovery.

3 Citation Endorsement

Given a document C, let C^O be the set of all documents cited by C such that $C \notin C^O$. Let $C^I = \{D | C \in D^O\}$ be the set of all documents that cite C.

Given a pair of documents $\{A, B\}$ such that $B \in A^O$, the *endorsement probability* of the citation (A, B), denoted by $\rho(A, B)$, is given by

$$\rho(A,B) = \frac{|A^I \cap B^I|}{\sum\limits_{\forall X \in A^O} |A^I \cap X^I|} \qquad (1)$$

The idea behind calculating endorsement probabilities for the outgoing citations of a document is to quantify how "distinguishable" each citation is from the others. Consider the following perspective for citation endorsements. Assume a category of users who browse a digital library like CiteSeer (or the Web) for researching a particular topic, and end up creating their own document on that topic. The significance that such users accord to other documents depends upon the relevance of those documents to their topic of interest. Given that such a user has cited a document A, the endorsement probabilities of the outgoing citations from A can be seen as a relative measure of the user's tendency to also cite any of the out-neighbors of A.

With respect to the copying model [7], the endorsement probability of a citation can be seen as the probability that a topically focused user will "copy" that citation. Similarly, the citation-endorsement probabilities can be seen as the propensity that a topically focused crawler would index an outgoing citation relative to the other outgoing citations from its current document.

In this demo, we use the endorsement probabilities of the outgoing citations from a given document to highlight the relevance of those citations to the topic of the given document. We have developed a Firefox add-on client for CiteSeer, called EndorSeer, which lists the outgoing citations of a given paper in the descending order of their endorsement probabilities, thereby helping the user maintain topical focus while browsing CiteSeer.

4 An Overview of EndorSeer

The overall functioning of EndorSeer is illustrated in figure 1. We extract the citation relationships between documents from the data dump provided by CiteSeer under the Open Archives Initiative [1]. We then compute the co-citations for every pair of documents incident on a citation, and store this information in our database. We also compute and store the endorsement probabilities for each of these citations as described in section 3.

Using EndorSeer as an add-on to a Web browser, when a user requests the CiteSeer page pertaining to a paper k, the Web browser contacts the CiteSeer server and fetches information about paper k as usual. Then, the add-on contacts our server and fetches the endorsed citations emanating from paper k. The user gets to seamlessly view the endorsed citations of paper k in addition to the information about paper k given by CiteSeer as usual, in the same browser win-

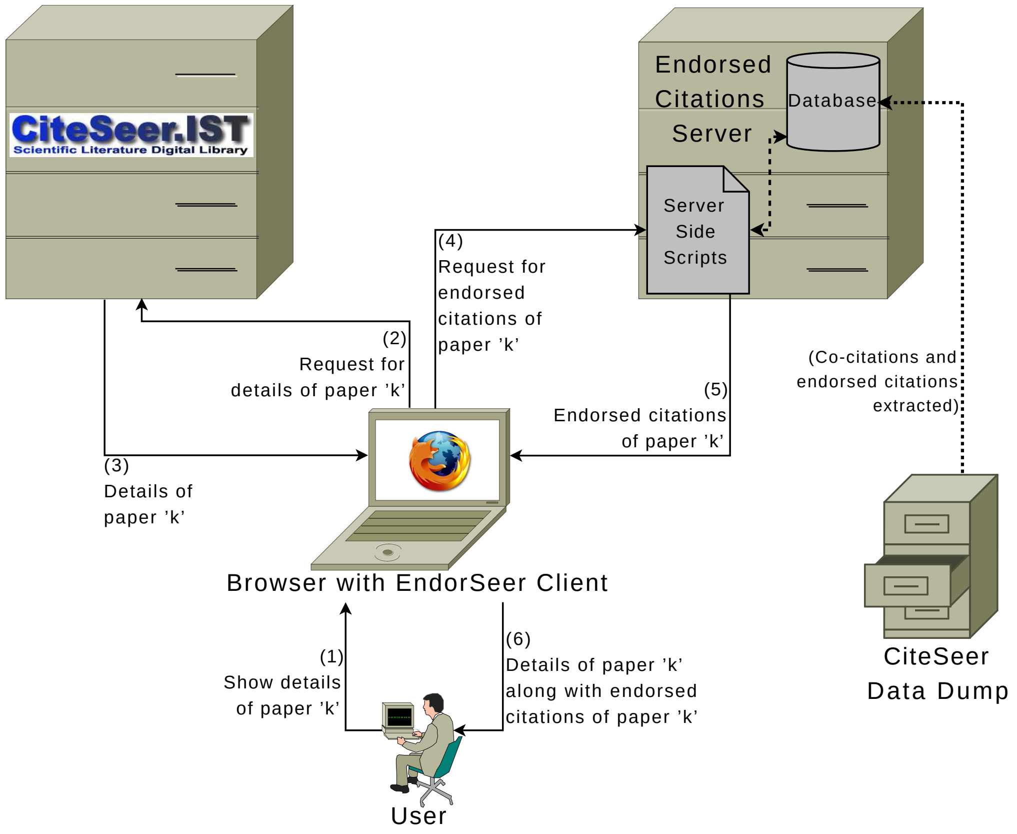

CiteSeer.IST
Scientific Literature Digital Library
Endorsed Citations Server
Database
Server Side Scripts
(4) Request for endorsed citations of paper 'k'
(2) Request for details of paper 'k'
(5) Endorsed citations of paper 'k'
(Co-citations and endorsed citations extracted)
(3) Details of paper 'k'
Browser with EndorSeer Client
CiteSeer Data Dump
(1) Show details of paper 'k'
(6) Details of paper 'k' along with endorsed citations of paper 'k'
User

Figure 1: A step-by-step illustration of the functioning of EndorSeer

dow or tab. The endorsed citations are listed in the descending order of their endorsement probabilities, ρ.

In figure 2, we show a screenshot of the functioning of the EndorSeer add-on. The user has accessed the CiteSeer page for the paper titled *The Case For Reliable Concurrent Multicasting Using Shared Ack Trees.* In addition to the details of the paper fetched from the CiteSeer server such as title, authors, abstract and citations, the endorsed citations of this paper fetched from our server by the EndorSeer add-on are also displayed in the browser window.

According to the add-on, the citation to the paper titled *A Reliable Multicast Framework for Light-weight Sessions and Application Level Framing* is the most topically relevant to the given paper among all its citations, with an endorsement probability of 0.203822, and has hence been highlighted accordingly. Similarly, eleven other citations have been endorsed with varying probabilities. If the user now clicks on any one of these citations, the details of the corresponding paper along with its endorsed citations are in turn displayed. Thus, the user can browse CiteSeer with topical focus using the endorsed citations.

A key requirement of EndorSeer is the availability of an offline data dump of the digital library for which it is intended, as also a refreshing strategy for that data dump. For the future, we intend to work on efficient mechanisms to crawl the CiteSeer website from time to time to refresh our offline dataset, so that papers that are freshly added to CiteSeer, along with the resultant citations and co-citations, can also be considered while computing endorsed citations on our server.

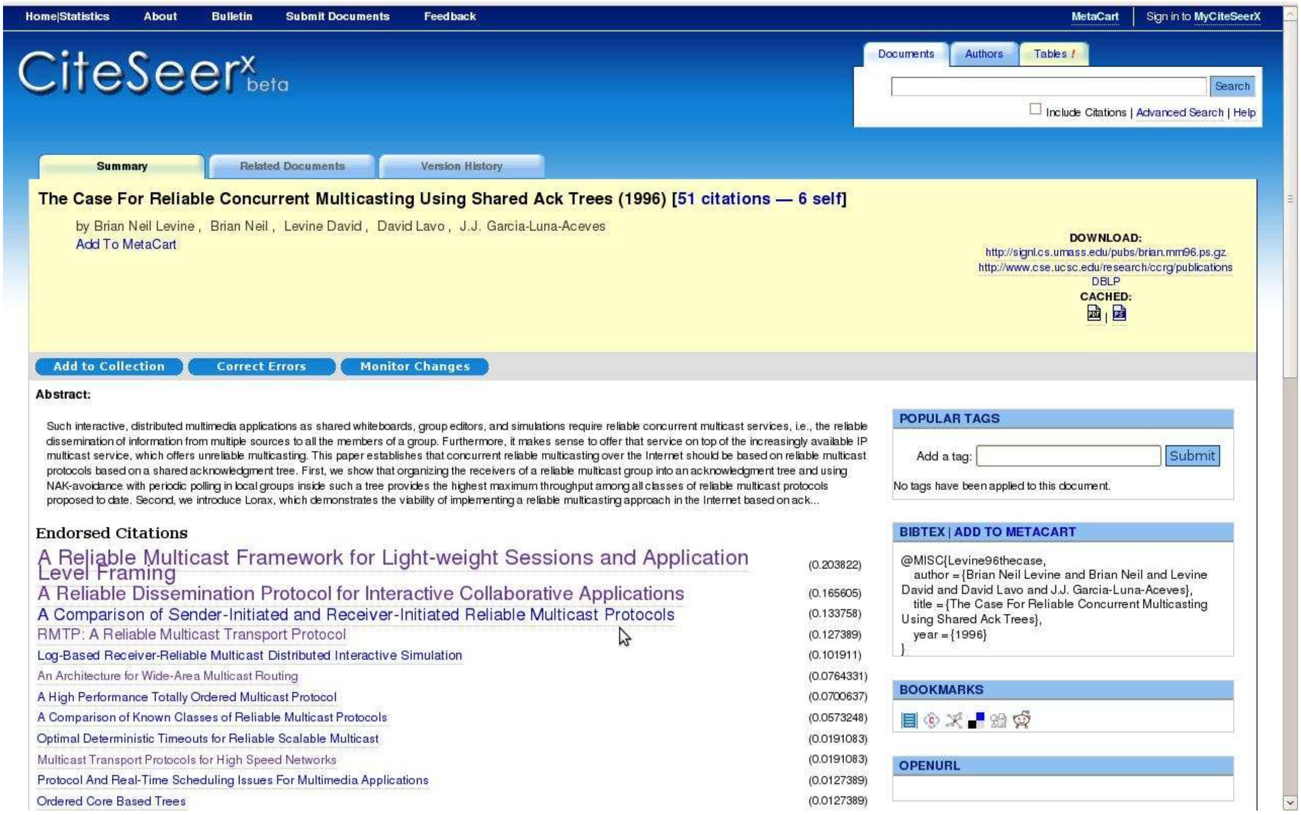
References

[1] *CiteSeer.IST scientific literature digital library (open archives initiative).* The dublin core standard with additional metadata fields, including citation relationships (References and IsReferencedBy), author affiliations, and author addresses. Available at `http://cs1.ist.psu.edu/public/oai/oai_citeseer.tar.gz` (accessed 6 February 2009).

[2] B. Davison. Topical locality in the Web. SIGIR 2000.

[3] J. Dean, M. R. Henzinger. Finding related pages in the World Wide Web. Computer Networks, 31(11-16), 1999.

[4] M. Efron. Using co-citation information to estimate political orientation in Web documents. KAIS, 9(4), 2006.



The Case For Reliable Concurrent Multicasting Using Shared Ack Trees (1996) [51 citations — 6 self]

by Brian Neil Levine , Brian Neil , Levine David , David Lavo , J.J. Garcia-Luna-Aceves

DOWNLOAD:
http://signl.cs.umass.edu/pubs/brian.mm96.ps.gz
http://www.cse.ucsc.edu/research/ccrg/publications
DBLP
CACHED:

Abstract:

Such interactive, distributed multimedia applications as shared whiteboards, group editors, and simulations require reliable concurrent multicast services, i.e., the reliable dissemination of information from multiple sources to all the members of a group. Furthermore, it makes sense to offer that service on top of the increasingly available IP multicast service, which offers unreliable multicasting. This paper establishes that concurrent reliable multicasting over the Internet should be based on reliable multicast protocols based on a shared acknowledgment tree. First, we show that organizing the receivers of a reliable multicast group into an acknowledgment tree and using NAK-avoidance with periodic polling in local groups inside such a tree provides the highest maximum throughput among all classes of reliable multicast protocols proposed to date. Second, we introduce Lorax, which demonstrates the viability of implementing a reliable multicasting approach in the Internet based on ack...

Endorsed Citations

A Reliable Multicast Framework for Light-weight Sessions and Application Level Framing	(0.203822)
A Reliable Dissemination Protocol for Interactive Collaborative Applications	(0.165605)
A Comparison of Sender-Initiated and Receiver-Initiated Reliable Multicast Protocols	(0.133758)
RMTP: A Reliable Multicast Transport Protocol	(0.127389)
Log-Based Receiver-Reliable Multicast Distributed Interactive Simulation	(0.101911)
An Architecture for Wide-Area Multicast Routing	(0.0764331)
A High Performance Totally Ordered Multicast Protocol	(0.0700637)
A Comparison of Known Classes of Reliable Multicast Protocols	(0.0573248)
Optimal Deterministic Timeouts for Reliable Scalable Multicast	(0.0191083)
Multicast Transport Protocols for High Speed Networks	(0.0191083)
Protocol And Real-Time Scheduling Issues For Multimedia Applications	(0.0127389)
Ordered Core Based Trees	(0.0127389)

POPULAR TAGS

Add a tag:

No tags have been applied to this document.

BIBTEX | ADD TO METACART

```
@MISC{Levine96thecase,
   author = {Brian Neil Levine and Brian Neil and Levine David and David Lavo and J.J. Garcia-Luna-Aceves},
   title = {The Case For Reliable Concurrent Multicasting Using Shared Ack Trees},
   year = {1996}
}
```

Figure 2: Screenshot of a CiteSeer page with EndorSeer showing endorsed citations for the paper. The citations shown by CiteSeer by default have not been captured in this screenshot.

[5] J. Hou, Y. Zhang. Effectively finding relevant Web pages from linkage information. TKDE, 15(4), 2003.

[6] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. JACM, 46, 1999.

[7] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, E. Upfal. The Web as a graph. PODS 2000.

[8] R. R. Larson. Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace. ASIS 1996.

[9] G. Moise, J. Sander, D. Rafiei. Focused co-citation: Improving the retrieval of related pages on the Web. WWW 2003.

[10] M. R. Mutalikdesai, S. Srinivasa. Co-citations as citation endorsements. JIS, under review, 2009.

[11] L. Page, S. Brin, R. Motwani, T. Winograd. The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford University, 1998.

[12] H. W. Park, M. Thelwall. Hyperlink analyses of the World Wide Web: A review. JCMC, 8(4), 2003.

[13] P. K. Reddy, M. Kitsuregawa. Inferring Web communities through relaxed co-citation and dense bipartite graphs. Database Engineering Workshop, 2001.

[14] S. Reddy, S. Srinivasa, M. R. Mutalikdesai. Measures of "ignorance" on the Web. COMAD 2006.

[15] H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. JASIS, 24(4), 1973.

[16] H. Small. Macro-level changes in the structure of co-citation clusters: 1983-1989. Scientometrics, 26(1), 1993.

[17] M. Thelwall, D. Wilkinson. Finding similar academic Web sites with links, bibliometric couplings and co-links. IPM, 40(3), 2004.

[18] H. D. White, B. C. Griffith. Author co-citation: A literature measure of intellectual structure. JASIS, 32(3), 1981.

[19] H. D. White, K. W. McCain. Bibliometrics. ARIST, 24, 1989.

[20] D. Zhao. Going beyond counting first authors in author co-citation analysis. ASIST 2005.