

PaperBot 

PaperBot 

Distributed Information Retrieval System

P Aditya Rao adityarao@jhu.edu

&

Naveen Natarajan nnatara2@jhu.edu

Table of Contents

Page No

Abstract	2
Introduction	2
Work done	3
Results	4
Setup Instructions	10
References	11

Abstract



Distributed information retrieval comes into play when a user wants to get information from different sources in parallel. One of the challenges of this topic is the Collection Fusion problem: The distinct result lists of the underlying information retrieval systems (IR) have to be fused to give a global relevance-ranked result list according to the user's information need.

PaperBot is a metasearch engine that blends the top web search results from Google, Microsoft and CiteSeer. It utilizes collection fusion algorithms to compile results from many of the Web's major search properties, delivering more relevant and comprehensive results every time a search is done. The search results include the algorithmic results from the most popular search engines on the Web. By accessing multiple search engines for each query, PaperBot provide a richer and more relevant spectrum of results than you would get from using any single search engine.

Introduction

Users want to retrieve documents from multiple libraries ranked by relevance concerning the users' information needs. The general ranking problem of information retrieval is enforced by the distribution of queries to different document collections. In this so-called Collection Fusion problem the local result lists retrieved from different IR systems should be merged to a global one upholding an optimal ranking concerning relevance to the user's information need.

The main problems with Collection Fusion in Distributed Information Retrieval are:

- The sources use different ranking algorithms.
- The ranking algorithms used by the sources are unknown.
- The parameters used with these algorithms (e.g., inverse document frequencies or term frequencies) cannot be obtained from the sources.

There are several strategies to perform the collection fusion. Wu and Crestani (2003) use the reference count (RC) method for automatic ranking. In this method, for a given query, they first consider the list of documents returned by a system, take each document of the list, and find references. More specifically, they count the occurrences of that document in the lists provided by other systems. They take each list one by one, find the summation of these reference counts, and rank documents using the total reference count sum for each document.

The architecture of PaperBot is shown in Fig. 1.

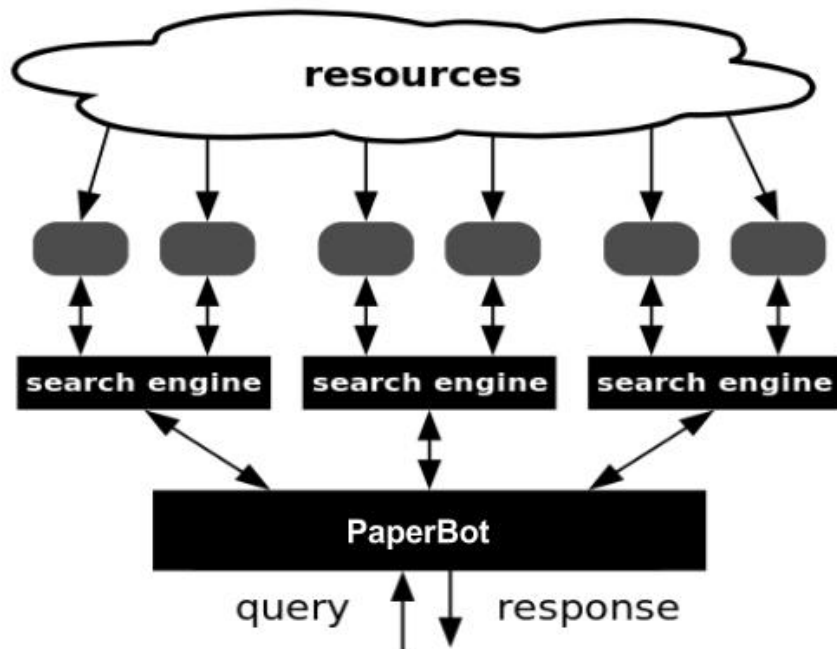


Fig 1. Architecture of PaperBot

Work done

Our implementation of meta-search (data fusion) involves three components:

1. **Query generator and dispatcher:** The request to the search engines generated based on the type of query made by the user. The generated queries are submitted to the underlying search engines- Microsoft Academic Research, CiteSeerx and Google Scholar. The queries can be of three types:
 - a. Only Keyword/Title search
 - b. Only Author search
 - c. Both Keyword/Title and Author search

We also handle the case where the queries are made with no parameters by redirecting them back to the main page.

2. **Document selector:** Each retrieved document is assigned an initial relevance score based on its source and their ranking of the documents. Of all the retrieved documents, the ones which are to be used from each search engine are determined based on the weights assigned to them. The documents are further processed based on the regional weighting of the query parameters. The three regions under consideration are the Title, Author, and Abstract of the document. The impact of regional weightings are decided on the type of query being made.
3. **Result merger:** A new datastore is created where all the results of search engines are merged based on the following merging technique. The rank of each document is

calculated based on the relevance score from each of the sources. We consider the rank positions of documents by assigning higher weights to documents that appear in upper rank positions. The relevance judgement is impacted by the source provider as well. The datastore is a set of document titles as key and relevance scores as values. All the duplicates from the results are removed since it is a set. We sum the total relevance score for the repetitive documents from all the sources. The top results are retrieved based on this total relevance score.

Implementation Details:

PaperBot is written in Python. It can extract the publication title, list of authors, publication url, published year, number of citations and the abstract. We have used several modules to get the information and parse it. The http request are made to the search engines based on the query generated by the query generator. This is done by the using the urllib and urllib2 package in python. The parsing was done using the functions of BeautifulSoup module and simple regular expression matching. We have two web-pages:

- paper.html is an interface where the user can type the query that he wants. The submit button on this page invokes our python script - paperbot.py.
- The top ten retrieved results are displayed on a webpage generated by paperbot.py.

Results

We conducted the following tests to verify the results generated by PaperBot.

Test # 1 : Search using Keyword/Title only

More importance is given to the search term appearing in title and abstract of a retrieved document when the query consists of Title/Keyword only.

Example: When searched for the term 'borealis' using Title/Keyword search, the results from PaperBot were:

Results for query :borealis

The design of the borealis stream processing engine

Daniel J. Abadi, Yanif Ahmad, Magdalena Balazinska, Mitch Cherniack, Jeong-hyon Hwang, Wolfgang Lindner, Anurag S. Maskey, Er Rasin, Esther Ryzkina, Nesime Tatbul, Ying Xing, Stan Zdonik - 2005
Borealis is a second-generation distributed stream processing engine that is being developed ...
Citations: 132

The Planet Orbiting r Coronae Borealis

Robert Noyes Adam, Adam R. Contos, Sylvain G. Korzennik, Peter Nisenson, Timothy M. Brown, Scott D. Horner, High Altitude Observatory
Continuing precise radial velocity observations of ae Coronae Borealis have allowed ...

Cancer borealis

Qiang Fu, Lamont S. Tang, Eve Marder, Lingjun Li
doi:10.1111/j.1471-4159.2007.04482.x Mass spectrometric characterization and physiological actions of VPNDWAHFRGSWamide, a novel B type allatostatin in the crab, ...

Fluorine in R Coronae Borealis Stars

Gajendra Pandey, David L. Lambert, N. Kameswara Rao - 711
Neutral fluorine (FI) lines are identified in the optical spectra of several RCoronae Borealis ...



E-Mail :
nnataraj2@jhu.edu
adityarao@jhu.edu

The results for Google Scholar with the same query:

[HTML] [Double white dwarfs as progenitors of R Coronae Borealis stars and Type I supernovae](#)
RF Webbink - The Astrophysical Journal, 1984 - adsabs.harvard.edu

... Printed in USA DOUBLE WHITE DWARFS AS PROGENITORS OF R CORONAE BOREALIS STARS AND TYPE I SUPERNOVAE RF WEBBINK Department of Astronomy, University of Illinois Received 1983 June 13; accepted 1983 July 27 ABSTRACT Close double white ...
Cited by 1028 Related articles All 5 versions Cite

[PDF] [The design of the borealis stream processing engine](#)

DJ Abadi, Y Ahmad, M Balazinska, U Cetintemel... - 2005 - cs.harvard.edu

Abstract **Borealis** is a second-generation distributed stream processing engine that is being developed at Brandeis University, Brown University, and MIT. **Borealis** inherits core stream processing functionality from Aurora [14] and distribution functionality from Medusa [51]. ...
Cited by 821 Related articles All 60 versions Cite More ▾

[Isolation and characterization of nutrients and value-added products from snow crab \(*Chionoecetes opilio*\) and shrimp \(*Pandalus borealis*\) processing discards](#)

F Shahidi, J Synowiecki - Journal of Agricultural and Food ..., 1991 - ACS Publications

Byproducts from shrimp and different parts of crab contained, on a dry basis, from 17.0 to 32.2% of chitin and from 3.4 to 14.7 mg/100 g of carotenoid pigments, mostly astaxanthin and its esters. Chitin was produced at a yield of about 86%, and carotenoids were ...
Cited by 230 Related articles All 3 versions Cite

[CITATION] The use of *Clintonia borealis* and other indicators to gauge impacts of white-tailed deer on plant communities in northern Wisconsin, USA

CP Balgooyen, DM Waller - Natural Areas Journal, 1995

Cited by 103 Related articles Cite More ▾

[Synopsis of biological data on the pink shrimp, *Pandalus borealis* Kroyer, 1838](#)




SE Shumway, HC Perkins, DF Schick, AP Stickney - 1985 - aquaticcommons.org

This synopsis of the literature was designed to summarize the biological and biochemical studies involving *Pandalus borealis* as well as to provide a summary of the literature.

The results for *Microsoft Academia* with the same query

Academic > Results for "borealis" in All Fields of Study Subscribe

Were you looking for these authors:

 Borealis AB
  Aurora Borealis
  A. B. Borealis

Publications (4286) any time

[The Design of the Borealis Stream Processing Engine](#) (Citations: 403) View...
 Daniel J. Abadi, Yanif Ahmad, Magdalena Balazinska, Ugur Cetintemel, Mitch Cherniack, Jeong-hyon Hwang, Wolfgang Lindner, Anurag Maskey, Alex Rasin, Esther Rykina, Nesime Tatbul, Ying Xing
borealis is a second-generation distributed...versity, brown university, and mit. **borealis** inherits core st... functionality...
 Conference: Conference on Innovative Data Systems Research - CIDR, pp. 277-289, 2005

[Fault-tolerance in the Borealis distributed stream processing system](#) (Citations: 90)
 Magdalena Balazinska, Hari Balakrishnan, Samuel Madden, Michael Stonebraker
 We present a replication-based approach to fault-tolerant distributed stream processing in the face of node failures, network failures, and network partitions. Our approach aims to reduce the degree of inconsistency in the system while guaranteeing that available inputs capable of being processed are processed within a specified time threshold. This threshold allows a user to trade availability for ...
 Conference: International Conference on Management of Data - SIGMOD, pp. 13-24, 2005

[Population differentiation in randomly amplified polymorphic DNA of red-cockaded woodpeckers *Picoides borealis*](#) (Citations: 52)
 S. M. HAIG, J. M. RHYMER, D. G. HECKEL
 Journal: Molecular Ecology - MOL ECOL, vol. 3, no. 6, pp. 581-595, 1994

The results for *CiteSeerX* with the same query

Results 1 - 10 of 685

Next

[Cancer borealis](#)
 by Qiang Fu, Lamont S. Tang, Eve Marder, Lingjun Li
 "... in the crab, Cancer **borealis** Qiang Fu,* Lamont S. Tang, Eve Marder and Lingjun Li* *School of Pharmacy ..."
[Abstract](#) - [Add to MetaCart](#)

[The design of the borealis stream processing engine](#)
 by Daniel J. Abadi, Yanif Ahmad, Magdalena Balazinska, Mitch Cherniack, Jeong-hyon Hwang, Wolfgang Lindner, Anurag S. Maskey, Er Rasin, Esther Ryzkina, Nesime Tatbul, Ying Xing, Stan Zdonik - *In CIDR , 2005*
 "... The Design of the **Borealis** Stream Processing Engine Daniel J. Abadi 1 , Yanif Ahmad 2 , Magdalena ..."
[Abstract](#) - [Cited by 132 \(8 self\)](#) - [Add to MetaCart](#)

[Borealis Developer's Guide Borealis Team](#)
 by unknown authors
 "... **Borealis** Developer's Guide **Borealis** Team May 31, 2006Contents 1 Introduction 4 1.1 System ..."
[Abstract](#) - [Add to MetaCart](#)

Test # 2 : Search using Author only

More importance is given to the search term appearing in author field of a retrieved document when the query consists of Author only.


Example: When searched for the author name 'David Yarowsky' using Author Name search, the results for PaperBot were :

Results for author :david yarowsky

[Unsupervised Word Sense Disambiguation Rivaling Supervised Methods](#)
 David Yarowsky - 1995
 This paper presents an unsupervised learning algorithm for sense disambiguation that, when trained on unannotated English text, rivals the performance of supervised techniques that require time-consuming hand annotations. The algorithm is based on two powerful constraints---that words tend to have one sense per discourse and one sense per collocation---exploited in an iterative bootstrapping procedure. Tested accuracy exceeds ...
 Citations: 743

[Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora](#)
 David Yarowsky - 1992
 This paper describes a program that disambiguates English word senses in unrestricted text using statistical models of the major Roget's Thesaurus categories. Roget's categories serve as approximations of conceptual classes. The categories listed for a word in Roget's index tend to correspond to sense distinctions; thus selecting the most likely category provides a useful level of sense ...
 Citations: 439

[Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French](#)
 David Yarowsky - 1994
 This paper presents a statistical decision procedure for lexical ambiguity resolution. The algorithm exploits both local syntactic patterns and more distant collocational evidence, generating an efficient, effective, and highly perspicuous recipe for resolving a given ambiguity. By identifying and utilizing only the single best disambiguating evidence in a target context, the algorithm avoids the problematic complex modeling of statistical dependencies. ...


 E-Mail :
 nnatara2@jhu.edu
 adityarao@jhu.edu


The results for Google Scholar with the same author name

[Unsupervised word sense disambiguation rivaling supervised methods](#)
D Yarowsky - Proceedings of the 33rd annual meeting on ..., 1995 - dl.acm.org
Abstract This paper presents an unsupervised learning algorithm for sense disambiguation that, when trained on unannotated English text, rivals the performance of supervised techniques that require time-consuming hand annotations. The algorithm is based on two ...
Cited by 1554 Related articles All 72 versions Cite

[Word-sense disambiguation using statistical models of Roget's categories trained on large](#)
D Yarowsky - Proceedings of the 14th conference on Computational ..., 1992 - dl.acm.org
Abstract This paper describes a program that disambiguates English word senses in unrestricted text using statistical models of the major Roget's Thesaurus categories. Roget's categories serve as approximations of conceptual classes. The categories listed for a ...
Cited by 845 Related articles All 38 versions Cite More ▼

[A method for disambiguating word senses in a large corpus](#)
WA Gale, KW Church, D Yarowsky - Computers and the Humanities, 1992 - Springer
Abstract Word sense disambiguation has been recognized as a major problem in natural language processing research for over forty years. Both quantitative and qualitative methods have been tried, but much of this work has been stymied by difficulties in acquiring ...
Cited by 608 Related articles All 20 versions Cite

The results for Microsoft Academia with the same author name

[Unsupervised Word Sense Disambiguation Rivaling Supervised Methods](#) (Citations: 743)  View...
David Yarowsky
This paper presents an unsupervised learning algorithm for sense disambiguation that, when trained on unannotated English text, rivals the performance of supervised techniques that require time-consuming hand annotations. The algorithm is based on two powerful constraints---that words tend to have one sense per discourse and one sense per collocation---exploited in an iterative bootstrapping procedure. Tested accuracy exceeds ...
Conference: Meeting of the Association for Computational Linguistics - ACL, pp. 189-196, 1995

[Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora](#)
(Citations: 439)
David Yarowsky
This paper describes a program that disambiguates English word senses in unrestricted text using statistical models of the major Roget's Thesaurus categories. Roget's categories serve as approximations of conceptual classes. The categories listed for a word in Roget's index tend to correspond to sense distinctions; thus selecting the most likely category provides a useful level of sense disambiguation ...
Conference: International Conference on Computational Linguistics - COLING, pp. 454-460, 1992

[A method for disambiguating word senses in a large corpus](#) (Citations: 310)
William A. Gale, Kenneth W. Church, **David Yarowsky**
Journal: Computers and The Humanities - COMPUT HUM, 1993

[Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French](#)

The results for CiteSeerX with the same author name

Results 1 - 10 of 61 Next

[Unsupervised word sense disambiguation rivaling supervised methods](#)
 by David Yarowsky - *IN PROCEEDINGS OF THE 33RD ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, 1995
 "... UNSUPERVISED WORD SENSE DISAMBIGUATION RIVALING SUPERVISED METHODS David Yarowsky Department ..."
 Abstract - Cited by 383 (4 self) - Add to MetaCart

[Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora](#)
 by David Yarowsky, 1992
 "... David Yarowsky AT&T Bell Laboratories 600 Mountain Avenue Murray Hill N J, 07974 yarowskyresearch ..."
 Abstract - Cited by 265 (10 self) - Add to MetaCart

[One sense per discourse](#)
 by William A. Gale, Kenneth W. Church, David Yarowsky - *In DARPA Speech and Natural Language Workshop*, 1992
 "... One Sense Per Discourse William A. Gale Kenneth W. Church David Yarowsky AT&T Bell Laboratories ..."
 Abstract - Cited by 172 (5 self) - Add to MetaCart

[Decision Lists For Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French](#)

Test # 3 : Search using Keyword/Title and Author

More importance is given to the keyword/title appearing in title and author name appearing in author region of a retrieved document when the query consists of Title/Keyword and Author both.


Example: When searched using both Title/Keyword as 'overlay networks' and the author name 'Yair Amir' using Author Name search, the results we got for PaperBot are



Results for query :overlay networks and author :yair amir

[Reliable Communication in Overlay Networks](#)
 Yair Amir, Claudiu Danilov - 2003
 Reliable point-to-point communication is usually achieved in overlay networks by applying TCP/IP on the end nodes of a connection. This paper presents an hopby-hop reliability approach that considerably reduces the latency and jitter of reliable connections. Our approach is feasible and beneficial in overlay networks that do not have the scalability and interoperability requirements of the global Internet....
 Citations: 24

[A Low Latency, Loss Tolerant Architecture and Protocol for Wide Area Group Communication](#)
 Yair Amir Claudiu Danilov Jonathan Robert Stanton - 2000
 ...
 Citations: 99

[A Cost-Benefit Flow Control for Reliable Multicast and Unicast in Overlay Networks](#)
 Yair Amir, Baruch Awerbuch, Claudiu Danilov, Jonathan Stanton - 2005
 Abstract — When many parties share network resources on an overlay network, mechanisms must exist to allocate the resources and protect the network from overload. Compared to large physical networks



E-Mail :
nnatara2@jhu.edu 
adityarao@jhu.edu 

The results for Google Scholar with the same author name and same query

[A low latency, loss tolerant architecture and protocol for wide area group communication](#)

Y Amir, C Danilov, J Stanton - Dependable Systems and ..., 2000 - [ieeexplore.ieee.org](#)

... **Yair Amir**, Claudiu Danilov, Jonathan Stanton Department of Computer Science The Johns Hopkins University Baltimore, Maryland 21218 USA {yairamir, claudiu, jonathan}@cs.jhu.edu Abstract Group communication systems are proven tools upon which to build **fault-tolerant** ...

Cited by 177 Related articles All 19 versions Cite More▼

[Scaling Byzantine Fault-Tolerant Replication to Wide Area Networks](#)

Y Amir, C Danilov, D Dolev, J Kirsch... - ... and Networks, 2006. ..., 2006 - [ieeexplore.ieee.org](#)

Abstract This paper presents the first hierarchical Byzantine **fault-tolerant** replication architecture suitable to systems that span multiple wide area sites. The architecture confines the effects of any malicious replica to its local site, reduces message complexity of wide ...

Cited by 45 Related articles All 30 versions Cite More▼

[The Totem single-ring ordering and membership protocol](#)

Y Amir, LE Moser, PM Melliar-Smith... - ACM Transactions on ..., 1995 - [dl.acm.org](#)

... messages broadcast and/or computations required. Recent protocols for **fault-tolerant** distributed systems [Amir et al. 1992b; Birman and van Renesse 1994; Kaashoek and Tanenbaum 1991; Melliar-Smith et al. 1990; Peterson et al. ...

Cited by 337 Related articles All 26 versions Cite

[Fast message ordering and membership using a logical token-passing ring](#)

Y Amir, LE Moser, PM Melliar-Smith... - ... the 13th International ..., 1993 - [ieeexplore.ieee.org](#)

... **Y. Amir**, LE Moser, PM Melliar-Smith, DA Agarwal, P. Ciarfella Department of Electrical and Computer Engineering University of California, Santa Barbara, CA 93106 Abstract Many protocols exist to support the maintenance of consistency of data in **fault-tolerant** distributed sys ...

Cited by 169 Related articles All 10 versions Cite More▼

[\[PDF\] The spread wide area group communication system](#)

Y Amir, J Stanton - 1998 - [ics.uci.edu](#)

Page 1. 1 The Spread Wide Area Group Communication System **Yair Amir** and Jonathan Stanton Department of Computer Science The Johns Hopkins University {yairamir, jonathan}@cs.jhu.edu Abstract Building a wide area group communication system is a challenge. ...

The results for Microsoft Academia with the same author name and same query

[Transis: A Communication Subsystem for High Availability](#) (Citations: 298)

Yair Amir, Danny Dolev, Shlomo Kramer, Dalia Malki

This paper describes Transis, a communication subsystem for high availability. Transis is a transport layer that supports reliable multicast services. The main novelty is in the efficient implementation using broadcast. The basis of Transis is automatic maintenance of dynamic membership. The membership algorithm is symmetrical, operates within the regular flow of messages, and overcomes partitions and re-merging. The higher layer provides various multicast services for sets of ...

Conference: Symposium on **Fault-Tolerant** Computing - FTCS, pp. 76-84, 1992

[The Totem single-ring ordering and membership protocol](#) (Citations: 225) [View...](#)

Yair Amir, Louise E. Moser, P. M. Melliar-Smith, Deborah A. Agarwal, P. Ciarfella

... **fault-tolerant** distributed systems are becoming more...

Journal: ACM Transactions on Computer Systems - TOCS, vol. 13, no. 4, pp. 311-342, 1995

[Fast Message Ordering and Membership Using a Logical Token-Passing Ring](#) (Citations: 101)

Yair Amir, Louise E. Moser, P. M. Melliar-Smith, Deborah A. Agarwal, P. Ciarfella

... maintenance of consistency of data in **fault-tolerant** distributed systems; these protocols are...

Conference: International Conference on Distributed Computing Systems - ICDCS, pp. 551-560, 1993

[Scaling Byzantine Fault-Tolerant Replication to Wide Area Networks](#) (Citations: 28)

Yair Amir, Claudiu Danilov, Danny Dolev, Jonathan Kirsch, Josh Olsen, David Zage

... the first hierarchical byzantine **fault-tolerant** replication architecture suitable to systems... compared with a at byzantine **fault-tolerant** approach...

Conference: Dependable Systems and Networks - DSN, 2005

[Scaling Byzantine Fault-Tolerant Replication to Wide Area Networks](#) (Citations: 16)

Yair Amir, Claudiu Danilov, Jonathan Kirsch, Danny Dolev, Cristina Nita-rotau, Josh Olsen, David John Zage

The results for Citeseer with the same author name and same query

Results 1 - 10 of 37
Transis: A Communication Sub-System for High Availability by Yair Amir, Danny Dolev, Shlomo Kramer, Dalia Malki , 1992 "... Transis: A Communication Sub-System for High Availability Yair Amir, Danny Dolev, Shlomo Kramer ..." Abstract - Cited by 337 (46 self) - Add to MetaCart
Replication Using Group Communication Over a Partitioned Network by Yair Amir , 1995 "... "Doctor of Philosophy" Yair Amir Submitted to the Senate of the Hebrew University of Jerusalem (1995). ii ..." Abstract - Cited by 81 (19 self) - Add to MetaCart
A Low Latency, Loss Tolerant Architecture and Protocol for Wide Area Group Communication by Yair Amir, Claudiu Danilov, Jonathan Stanton - In Proceedings of the International Conference on Dependable Sys Networks , 2000 "... A Low Latency, Loss Tolerant Architecture and Protocol for Wide Area Group Communication Yair ..." Abstract - Cited by 80 (14 self) - Add to MetaCart
Membership Algorithms for Multicast Communication Groups by Yair Amir, Danny Dolev, Shlomo Kramer, Dalia Malki - In 6th Intl. Workshop on Distributed Algorithms proceedings (LCNS , 1992

Setup Instructions

There is already a web interface running on the CS department server. The link is <http://cs.jhu.edu/~nnataraj/paper.html>. You can use this link or set it up on your own.

In order to set it up on your own follow the instructions given below:

- 1) Unzip the paperbot.zip folder.
- 2) Go inside the public_html folder and copy all the contents from it to the public_html folder on your server.
- 3) Set the permissions for all files. There will be a folder named cgi-bin, make sure to change the permissions for that folder also.
- 4) Now go to any web browser and go to <http://yourserver.com/paper.html> and you can run it.

I would strongly suggest you to use the link on the CS server

<http://cs.jhu.edu/~nnataraj/paper.htm>

References

1. Wu, S., & Crestani, F. (2003). *Methods for ranking information retrieval systems without relevance judgments*. In *Proceedings of the ACM symposium on applied computing conference* (pp. 811–816).
2. *Automatic ranking of information retrieval systems using data fusion* Rabia Nuray , Fazli Can 2 June 2005
3. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008