Monte Carlo Methods for Web Search

Balázs Rácz

Summary of PhD thesis

Advisor:
Dr. András A. Benczúr

Data Mining and Web Search Group
Computer and Automation Research Institute of the
Hungarian Academy of Sciences

and

Department of Algebra
Budapest University of Technology and Economics

Budapest
2009

1 Introduction

One of the fastest growing sector of the software industry is that of the Internet companies, lead by the major search engines: Google, Yahoo and MSN. The importance of this field is even more emphasized by the plans of almost unprecedented magnitude that the European Union is pursuing to ease their dependence on these US-based technological firms.

The scientific and technological difficulties of this field are dominated by the mere scale: the web is estimated to contain tens to hundreds of billions of pages, with an exponential increase for over a decade and without showing any signs of that growth slowing down. At this scale, even the simplest mathematical constructs, such as a set of linear equations or a matrix inversion are turning out to be infeasible or practically unsolvable.

This thesis and the underlying publications provide solutions to certain of these scalability problems stemming from core web search engine research. The actual problems and their abstract solutions are not ours; they were described in earlier works of seminal authors of the field, generating considerable interest. Nevertheless, it was our work showing the first methods which could really scale to the size of the web without serious limitations.

A particularly important aspect of our solutions is that they are not only theoretically applicable to the web, but also very practical: they follow fairly closely and naturally fit into the architecture of a web search engine; the algorithms are parallelizable or distributed; the computational model we assumed is the one that is present in all current major data centers; and the query serving parts show characteristics very important for industrial applications, such as fault tolerance.

An important price we pay for these benefits is that out methods give approximate solutions to the abstract formulation. However, on one hand we have strict bounds on the approximation quality, on the other hand we formally prove that this is the only way to go: we give lower bounds on the resource usage of any exact method, prohibiting their application on datasets on the Web scale.

2 Overview

The thesis presents results in three groups of claims.

In the first set of claims we consider the problem of *personalized web search*, also called as personalized ranking. General web search has a static, global ranking function that the engine uses to sort the results according to some notion of relevance that depends on the query but not the user. However, relevance can easily differ from user to user, e.g. a computer geek and a history teacher may find different sites authoritative and interesting for the same query. Personalized web search allows users to specify their preference, and this preference parametrizes the ranking function. As PageRank is the most successful static ranking function, the personalized version, Personalized PageRank [23] is of particular interest. All earlier methods for computing personalized PageRank [10, 16, 17] had severe restrictions on what personalization they allowed [13]. In our work we provided the first Personalized PageRank algorithm allowing arbitrary personalization and still scaling to the full Web.

In the second set of claims we consider the problem of *similarity search in massive graphs* such as the web. Similarity search is not only motivated by advanced data mining algorithms requiring easily computable similarity functions such as clustering algorithms, but also by the 'Related pages' functionality of web search engines, where the user can query by example: supplying the URL of a web page of interest, the search engine replies by good quality pages on a similar topic. Traditional similarity functions stemming in social network analysis such as co-citation express the similarity of two nodes in a graph by using only the neighbors of the nodes in question. However, considering the size and depth (e.g. average diameter) of the web graph, this is just as inadequate as using degree as a ranking function. We consider the similarity function proposed by Jeh and Widom, SimRank [15], which is a recursive definition similar to that of PageRank. Our methods provided the first algorithm that scaled beyond graphs of a few hundred thousand nodes.

In the above areas we follow the same outline: We first give approximation algorithms for the problem, analyzing the approximation quality and convergence speed. Then we claim impossibility results about non-approximation approaches, proving prohibitive space complexity. Finally we validate the methods using experiments on real Web datasets.

In the final claims we pursue further impossibility results on similarity functions of massive graphs. We consider the decision problem: is there a pair of vertexes in a graph that share a common neighborhood of a particular size? (This is equivalent to the existence of the complete bipartite graph $K_{2,c}$ as a subgraph.) We are particularly interested in the space complexity of the problem in the data stream model: an algorithm $\mathcal{A}$ is allowed to read the set of edges of the graph sequentially, and after having one or constant many passes, it has to output the answer to the decision problem. We lower bound the temporary storage use of any such algorithm in the randomized computation model. The relevance of this problem to web search is that an algorithm $\mathcal{A}$ for the decision problem can be emulated by a search engine. During the preprocessing phase the search engine indexer can read the input a few times, producing an index database. Then the search engine query processor can answer queries only the index database, and a proper sequence of queries gives us the answer to the decision problem. Therefore any lower bound we prove on the decision problem applies either to the temporary storage requirements of the indexer, the query engine, or the index database size. A prohibitive (say, quadratic in the input size) lower bound makes it impossible to build a query engine that can feasibly serve similarity queries up to the required precision.

3 Research objectives

During our work we seek to answer the following questions with regards to the SimRank similarity function and the Personalized PageRank ranking function:

- Find a *scalable* approximation algorithm that computes these scores at query serving time.
- Prove that there exists no scalable algorithm that would compute the exact scores.
- How good is the approximation returned by our algorithm?
- What are the resource requirements of our algorithm? Show that our solution adheres the scalability requirements by conduction experimental runs on sufficiently large inputs.
- Are the mathematical definitions usable in practice? What is the quality of result lists delivered by these algorithms? Present an experimental quality evaluation on real Web datasets. Present experimental evidence that these functions are better suited to the Web than the classic solutions.
- Parameter tuning: How shall we set the parameters in our solution to gain sufficient quality results with acceptable resource consumption?
- Define new functions and analyze them according to the above criterion.

It is easy to see that the central concept to our research goals is that of *scalable algorithms*. Due to the sheer size of the Web as a dataset many specialized systems and architectures were created to deal with this challenge [4, 2, 8]. We consider an algorithm scalable for web search engines if it fulfills the following requirements [24, 19]:

- **Precomputation:** The method consists of two parts: an off-line precomputation phase, which is allowed to run for about a day to precompute an index database, and an on-line query serving part, which can access only the index database, and needs to answer a query within a few hundred milliseconds.

- **Time:** The index database is precomputed within the time of a sorting operation, up to a constant factor. To serve a query the index database can only be accessed a constant number of times.
- **Memory:** The algorithms run in *external memory*: the available main memory is constant, so it can be arbitrarily smaller than the size of the Web graph. In some cases we will consider semi-external-memory algorithms [21] with linear memory requirement in the number of vertexes in the web graph, with a small constant factor.
- **Parallelization:** Both precomputation and query part can be implemented to utilize the computing power and storage capacity of thousands of servers interconnected with a fast local network.

4 Research methods

The algorithms we developed can be classified as fingerprint-based data mining algorithms, the textbook example of which was established by Broder [6]. These methods operate by expressing the result as an expectation of a random variable, and then by taking N independent sample (fingerprint) we estimate the result via Monte Carlo method.

To create probabilistic reformulations of PageRank and similar problems we heavily rely on the random walk-based expression of PageRank: on one hand the stationary distribution of the uniform random walk (Markov-chain) on the graph [23], on the other hand the ending point of the random walk with uniform starting point and geometrically distributed length [9].

To show the infeasibility of exact computation of the measures in question we prove lower bounds on the space complexity of the problems. We use the methodology developed for analyzing graph algorithms in the data stream model [14], where we reduce the problems at hand to communication complexity games [20], mostly the bit-vector probing problem.

In the experimental evaluation of Personalized PageRank we compare the closeness of approximation to the algorithm by Jeh and Widom [16]. To compare the resulting ranking orders we use the methodology applied in PageRank research [18, 11, 26].

In the experimental evaluation of similarity search functions we use the methodology developed by Haveliwala [12], where we utilize a high-quality Internet Directory, the Open Directory Project (DMOZ) [22]. Taking the category classification of the directory as a base truth, we quantify the quality of similarity search functions by how close it can reproduce the category classification.

5 New Results

Claim 1: Monte Carlo algorithm for computing Personalized PageRank

The main problem of the currently prevalent model of Web Search is that the user has to express her information need as a keyword query. This is a very difficult task, especially for the average user. If the query is too specific, contains too many words, there is a good chance that the page the user is looking for does not match it, because it happens to phrase the information with different wording – this is the problem of *recall*. On the other hand, if the query is too generic, contains too few words, then millions of other pages will match it, and from this long result list it is quite impossible to select the page that the user will be interested in – this is the problem of *precision*.

Due to the recall problem the users' behavior has shifted to phrasing simple, very short search queries, accepting the large multitude of results. Therefore the algorithms behind the search engine will have the main objective to tackle the precision problem by presenting the result list in an order to the user where the most relevant pages are in the top few results.

The ranking problem has been studied extensively, and the solutions can be classified according to several aspects. A *local* ranking algorithm considers a single page at a time, whereas a *global* ranking algorithm runs on the entire dataset. A *static* ranking computes a fixed ranking from the dataset and applies this ranking for every query, whereas *dynamic* ranking algorithms are query-dependent. In practice we typically use a mixture of algorithms, for example a local static algorithm for identifying and filtering malicious web pages (e.g. malware), a local dynamic algorithm to score the keyword matches in the page (e.g. weight matches in the title or in large font higher), and a global static algorithm to represent the popularity of the result page on the entire Web (to capture the quality of the page).

In this last category of ranking algorithms the most widely researched method is *PageRank* [23], since many believe it to be the driving factor behind the quality and popularity of the leading search engine, Google. PageRank is based on the following assumption:

> A hyperlink $u \rightarrow v$ is the vote of page u for the quality of content of page v.

This intuition is applied recursively in the definition of PageRank in that the PageRank value of a web page v can be computed from the PageRank values of the pages linking to v.

A major drawback of the PageRank algorithm is that it is static, it computes the relevance of a web page as one single number, and applies the same decision to all queries, no matter if it is an American computer scientist or a Mongolian history teacher asking. This drawback is fixed by personalization, where we can compute the relevance values based on the judgment of a subset of a Web, and aim to have the ability to set this subset individually for each user.

The main difficulty of Personalized PageRank [5, 23] that the starting point, the personalization is only available at query time. This makes the usual PageRank calculation methods infeasible, since they typically require several hours of computation, and even the most patient users cannot be expected to wait that long in hope for the benefits of personalization. Several groups have been seeking scalable methods for personalization [10, 16, 17, 13], but all of these prior work have had significant restrictions on how the personalization can be expressed. The main result of this claim is an algorithm that allows unrestricted personalization:

> **Claim 1.1 [J4, C9].** A scalable randomized algorithm for computing Personalized PageRank scores that returns an unbiased estimation for any personalization starting point with constant many database accesses from an index database with a size linear in the number of web pages. Improvement of the approximation quality by utilizing the database records for the neighbors of the starting page.

Since the Personalized PageRank values are linear in the weighted starting distribution vector [10], we can reach arbitrary personalization based on this result.

Of course for the feasibility of the above method we need to be able to compute the index database using a scalable method. I have given two solutions to this problem, of which on can select based on the available resources.

> **Claim 1.2 [J4, C9].** External memory indexing method computing the index database of Claim 1.1 on a graph with V nodes and average degree d, using M internal memory with $\Theta(V(N \log_M NV + Ld))$ I/O operations, where N is a constant controlling the approximation quality, and L is a constant appropriate for the mixing speed of the graph.

Substituting the values of the constant resulting from the experimental evaluation in the thesis ($L = 20$, $d = 10$, $N = 100$, $V = 10^{10}$, $M = 1\text{GB}$) we get a total I/O requirement of 256 TB, which can be performed using 60 disk in a day. The actual space used is 8 TB, and since the algorithm only uses external memory sort and merge to run, the disk access can be performed in blocks of up to several hundred megabytes, thereby reaching the peak data transfer speed of modern disks.

Claim 1.3 [J4, C9]. Indexing method for computing the database of Claim 1.1 using K computers interconnected with a fast local-area network, where the total memory of the computers is sufficient to store the entire Web graph, with the expected total communication of $\Theta(NV)$.

In the recent years very sophisticated methods were developed for storing the Web graph in main memory [1, 3], which require only a few bits per link. However, using a much simpler approach allowing faster processing we can still perform the computation using 100 typical workstation-sized machines. Substituting the above mentioned constants and using everyday network technologies the indexing can be completed with 100 machines in about an hour.

Claim 2: Analyzing and improving Monte Carlo methods for computing the SimRank similarity function

As we mentioned in the introduction of Claim 1, one of the main problems of Web search is that of the difficulty of formulating keyword queries (from the perspective of the user), and the difficulty of understanding the keyword queries (from the perspective of search engines). A possible solution to this problem is to ask for more data from the user when she specifies a search query. Of course we don't want to complicate the search workflow and disrupt its fluency by clarification questions or complicated UI, therefore it is especially useful if the search query contains some implicit extra information.

One possibility of such implicit extra information is *search by example*. In this mode of operation the user specifies an existing web page as a query instead of some keywords, and expects a response of a list of web pages in the same topic. This functionality has been available since the beginning on the search result pages of search engines under the link "Similar Pages". Despite this being probably the most often displayed link today (since it appears many times on all search result pages) it receives relatively little traffic, most probably because the current algorithms return results of varying quality.

It is reasonable to assume that advanced link-mining algorithms will revolutionize search by example just as PageRank has revolutionized the ranking problem. This is why the primary focus of our research has been the SimRank similarity function [15], which defines the similarity of two web pages (or nodes in an arbitrary graph) with a recursive definition similar to PageRank.

The major difficulty with the SimRank similarity function is that while one can use the naive power-iteration method to compute PageRank, this is absolutely infeasible for SimRank, since the resource requirements would be quadratic in the number of web pages. Previous results using aggressive heuristics were only able to apply SimRank on graphs with about 200,000 nodes.

The first SimRank algorithm that is truly scalable to the size of the Web (as defined in our research objectives) was developed by my co-author Dániel Fogaras [J5, C10]. This is a randomized approximation algorithm that computes fingerprints for each node in the graph, and then gives unbiased estimation on the SimRank value using these fingerprints. Using Monte Carlo method, with N fingerprints we can get sufficient precision:

Claim 2.1 [J5, C10]. Analysis of the convergence speed of the fingerprint-based SimRank approximation method, and proof that for any fixed absolute error the error probability converges to zero exponentially in the number N of fingerprints taken, uniformly over the all nodes and all graphs. Proof that for the top query problem ignoring a fixed absolute error the expected recall converges to 1 exponentially and uniformly over all nodes and all graphs.

The important consequence of this claim is that with a fixed error the number N of fingerprints can be considered constant, independently of the query or even the growth of the graph (i.e., even asymptotically).

Despite having fairly strong theorems about the convergence speed a natural question arises whether there is an algorithm performing exact computation or we have to do with approximate solutions? My lower bound theorems answer this question:

> **Claim 2.2 [J5, C10].** Lower bound on the index database size, in that any SimRank algorithm supplying exact results on arbitrary graphs will require index database of $\Omega(V^2)$ on some graphs with V nodes, whereas any approximation algorithm will require $\Omega(V)$ sized index database.

The direct corollary of this is that we can't hope for a generic solution for graphs sized as the Web, since the required index database exceeds the total storage capacity ever manufactured. Our approximation similarity search method is on the other hand space-optimal up to a logarithmic factor using the following representation:

> **Claim 2.3 [J5, C10].** Compact representation for the fingerprint paths generated by the [P]SimRank algorithm of [C10] that encodes the coupled fingerprint paths in two cells per node.

This compact representation requires asymptotically $O(V \log V)$ storage, which means that substituting the usual constants ($V = 10^{10}$, $N = 100$) the similarity database for the entire Web consumes 8 TB of space.

Our algorithms show very important properties from the industrial perspective:

> **Claim 2.4 [J5].** Preparation of our algorithms for industrial [2] application: parallelization, fault tolerance, load balancing and dynamic adaptation to workload. Incremental indexing methods for updating the index. Experimental proof that the total serving capacity of a cluster is linear in the number of computing nodes in the cluster.

Claim 3: On the common neighborhood problem

In this claim we consider an abstract problem, which can be considered a complexity theory interpretation of the graph-based similarity search problem. Buchsbaum, Giancarlo and Westbrook considered in [7] the following decision problem in the data stream model: Given a directed graph and a constant c, decide whether the graph has a $\overrightarrow{K_{2,c}}$ as a directed subgraph, i.e., is there a pair of nodes with at least c common neighbors?

The data stream model presents the input graph on a one-way read-only input tape to the algorithms. Two interesting cases are usually considered: in the single-pass model the input tape can be advanced only in one direction, i.e. the input can be read through only once. This model is especially suited for application where a large quantity of continuously streaming data has to be processed, since these streams are typically not possible to be stored and processed offline due to the mere data volume. The general case allows a "rewind" operation on the input take, which the algorithm can trigger $O(1)$ times, i.e., the input can be read through constant many times. This is a good model for data residing of secondary storage, where the cost of random access is infeasible. This is true for the current hard disk technologies.

The interesting question in the data stream model is always the temporary storage requirement, to give lower bounds on the internal storage requirement of any algorithm.

Unfortunately one of the basic lemmas in the the above quoted paper [7] has an incorrect proof that cannot be fixed easily.

> **Claim 3.1 [J2].** Correct proof for the single-pass data stream model results of [7].

Using the new proof methodology we can give stronger bounds in both a single and the $O(1)$-pass model. The new bounds are tight up to a logarithmic factor, i.e. we also give algorithms that solve the common neighborhood problem with a logarithmic factor more storage.

Claim 3.2 [J2]. Lower bound on the common neighborhood problem in the single-pass data stream model, in that the temporary storage requirement for graphs with n vertexes and neighborhood threshold c is $\Omega(\sqrt{c}n^{3/2})$. Algorithm for solving the common neighborhood problem with $O(\sqrt{c}n^{3/2}\log n)$ space.

Claim 3.3 [J2]. Lower bound on the common neighborhood problem in the $O(1)$-pass data stream model, in that the temporary storage requirement for graphs with n vertexes and neighborhood threshold c is $\Omega(\sqrt{c}n^{3/2})$. Algorithm for solving the common neighborhood problem with $O(\sqrt{c}n^{3/2}\log n)$ space.

Application of results

The results in Claim 2 were implemented by Dániel Fogaras as part of the research grant "Analog", and the result scan be tried on a brawl of the .hu domain from 2004 on the website `www.hasonlo.hu`. The following table shows and example query.

	Similarity query result for query `www.bkv.hu` using the PSimRank similarity function	Description
1	`www.bkv.hu/`	public transport company of Budapest
2	`www.malev.hu/`	Hungarian Airlines
3	`www.elvira.hu/`	online timetable for the Hungarian Railways
4	`www.mahart.hu/`	Hungarian Ship Lines
5	`www.turizmusonline.hu/adatbazis/kutatas_fejlesztes.php`	Tourism Office
6	`www.turizmusonline.hu/heti_turizmus/bemutatkozo.php`	Tourism Office
7	`www.volan.hu/`	Hungarian Coach Lines
8	`www.idojaras.hu/`	weather
9	`www.met.hu/`	weather
10	`www.worldtimeserver.com/`	

Acknowledgment

Above all I would like to thank my co-author and at the time fellow PhD student Dániel Fogaras for introducing me to the field of link-based search algorithms and the several years of fruitful collaboration that followed.

I would like to say special thanks to the leading research scientists of the Data Mining and Web Search Group: András Benczúr, who was my advisor, and András Lukács. They have supported my career during my time at the Hungarian Academy of Sciences, and supplied me with a constant stream of very diverse tasks ranging from CS research through software engineering of large-scale infrastructure systems to R&D grants work and public procurement, which have developed a versatile set of skills that turned out to be very helpful in my further career. I am thankful to my colleagues and co-authors in the research group, Károly Csalogány and Tamás Sarlós for implementing some of our algorithms and doing experimental evaluation, and for all the feedback they gave during the research work that kept us on the right course. The quality of our manuscripts and papers were singificantly improved by comments from the aforementioned people, as well as Lajos Rónyai, Dániel Marx, Glen Jeh, Andrew Twigg, Adam L. Buchsbaum and Raffaele Giancarlo. I would like to especially thank to Ferenc Bodon, with whom I worked together for a long time on a field of research outside this thesis.

I am especially grateful to Professors Lajos Rónyai, András Recski and Bálint Tóth, who, as the heads of the three departments related to my field (Algebra, Computer Science and Probability, respectively) have supported my development with advice, challenges and opportunities.

Publications

Publication score sccording to the system by the Habilitation and Doctoral Committee: 29.66 points.
Number of publications (total): 16
Number of reviewed publications: 14
Number of known cictations (total): 124
Number of known independent citations: 109

Journal papers

[J1] András Benczúr, István Bíró, Károly Csalogány, Balázs Rácz, Tamás Sarlós, and Máté Uher. PageRank és azon túl: Hiperhivatkozások szerepe a keresésben. *Magyar Tudomány*, 2006(11):1325, 2006. L 2 / 6 = 0,33 pont.

[J2] A. L. Buchsbaum, R. Giancarlo, and B. Rácz. New results for finding common neighborhoods in massive graphs in the data stream model. *Theoretical Computer Science*, 407(1-3):302–309, 2008. LR 6 / 3 = 2 pont.

[J3] Z. Dezső, E. Almaas, A. Lukács, B. Rácz, I. Szakadát, and A.-L. Barabási. The dynamics of information access on the web. *Physical Review E*, 73(6), 2006. LR 6 / 6 = 1 pont.

[J4] D. Fogaras, B. Rácz, K. Csalogány, and T. Sarlós. Towards fully personalizing PageRank: Algorithms, lower bounds and experiments. *Internet Mathematics*, 2(3), 2005. LR 6 * 33% = 2 pont.

[J5] Dániel Fogaras and Balázs Rácz. Practical algorithms and lower bounds for similarity search in massive graphs. *IEEE Transactions on Knowledge and Data Engineering*, 19(5):585–598, 2007. L 6 / 2 = 3 pont.

[J6] Alexei Vazquez, Balázs Rácz, András Lukács, and Albert-László Barabási. Impact of non-poissonian activity patterns on spreading processes. *Physical Review Letters*, 98(15):158702, 2007. LR 6 / 4 = 1,5 pont.

Papers in conference proceedings

[C7] A. A. Benczúr, K. Csalogány, K. Hum, A. Lukács, B. Rácz, Cs. I. Sidló, and M. Uher. Architecture for mining massive web logs with experiments. In *Proceedings of the HUBUSKA Open Workshop on Generic Issues of Knowledge Technologies*, 2005. 2 / 6 = 0,33 pont.

[C8] D. Fogaras and B. Rácz. A scalable randomized method to compute link-based similarity rank on the web graph. In *Proceedings of the Clustering Information over the Web workshop, Conference on Extending Database Technology*, 2004. L 4 / 2 = 2 pont.

[C9] D. Fogaras and B. Rácz. Towards fully personalizing PageRank. In *Proceedings of the 3^{rd} Workshop on Algorithms and Models for the Web-Graph (WAW2004), in conjunction with FOCS 2004.*, 2004. L 4 / 2 = 2 pont.

[C10] D. Fogaras and B. Rácz. Scaling link-based similarity search. In *Proceedings of the 14^{th} Int'l World Wide Web Conference*, 2005. L 6 / 2 = 3 pont.

[C11] Balázs Rácz. nonordfp: An FP-Growth variation without rebuilding the FP-tree. In *FIMI '04, Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations*, 2004. L 4 pont.

[C12] B. Rácz, F. Bodon, and L. Schmidt-Thieme. On benchmarking frequent itemset mining algorithms. In *Proceedings of the 1st International Workshop on Open Source Data Mining, in conjunction with ACM SIGKDD*, 2005. L 4 * 50% = 2 pont.

[C13] B. Rácz, A. Lukács, and Cs. I. Sidló. Two-phase data warehouse optimized for data mining. In *Proceedings of the First International Workshop on Business Intelligence for the Real-Time Enterprise, in conjunction with VLDB 2006*, 2006. L 4 * 50% = 2 pont.

[C14] T. Sarlós, A. A. Benczúr, K. Csalogány, D. Fogaras, and B. Rácz. To randomize or not to randomize: Space optimal summaries for hyperlink analysis. In *Proceedings of the 15th Int'l World Wide Web Conference*, 2006. L 6 / 4 = 1,5 pont.

Technical report

[T15] B. Rácz. Adatbányászati és többváltozós statisztikai modellek elektronikus újságok látogatottsági adatainak elemzésére., 2002. TDK dolgozat, 1 pont.

[T16] B. Rácz. Tömörítés és hosszútávú tárolás. Technical Report 4, Adatrosta könyvtár, 2003. 2 pont.

Miscellaneous

[M17] B. Rácz and A. Lukács. High density compression of log files. In *Proceedings of the Data Compression Conference*, page 557, 2004. (poster), L 0 pont.

Citations

Scaling Link-Based Similarity Search [C10]

31 independent, 33 total citations

[H1] Mohammed Al-Badawi, Dr. Siobhán North, and Dr. Barry Eaglestone. Indexing xml databases: Classifications, problems identification and a new approach. Technical report, The University of Sheffield, 2007.

[H2] Ilaria Bartolini and Paolo Ciaccia. Towards an effective semi-automatic technique for image annotation. In Michelangelo Ceci, Donato Malerba, and Letizia Tanca, editors, *SEBD*, pages 258–265, 2007.

[H3] Ilaria Bartolini and Paolo Ciaccia. Imagination: Exploiting link analysis for accurate image annotation. In *Adaptive Multimedial Retrieval: Retrieval, User, and Semantics: 5th International Workshop, AMR 2007, Paris, France, July 5-6, 2007 Revised Selected Papers*, pages 32–44, Berlin, Heidelberg, 2008. Springer-Verlag.

[H4] Luca Becchetti, Paolo Boldi, Carlos Castillo, and Aristides Gionis. Efficient semi-streaming algorithms for local triangle counting in massive graphs. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 16–24, New York, NY, USA, 2008. ACM.

[H5] András A. Benczúr, Károly Csalogány, and Tamás Sarlós. Link-based similarity search to fight web spam. In *AIRWeb 2006, Proceedings of the Second International Workshop on Adversarial Information Retrieval on the Web*, pages 9–16, 2006.

[H6] A. Broder. CS598E course material on Princeton University, 2005. `http://www.cs.princeton.edu/courses/archive/spr05/cos598E/bib/Princeton.pdf`.

[H7] Aurel Cami and Narsingh Deo. Techniques for analyzing dynamic random graph models of web-like networks: An overview. *Networks*, 51(4):211–255, 2008.

[H8] Paul Alexandru Chirita, Claudiu S. Firan, and Wolfgang Nejdl. Personalized query expansion for the web. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 7–14, New York, NY, USA, 2007. ACM.

[H9] K. Csalogány, A. Benczúr, D. Siklósi, and L. Lukács. Semi-supervised learning: A comparative study for web spam and telephone user churn. In *Proc. of Graph Labelling Workshop and Web Spam Challenge 2007 in conjunction with ECML/PKDD 2007*, 2007.

[H10] D. Fogaras, B. Rácz, K. Csalogány, and T. Sarlós. Towards fully personalizing PageRank: Algorithms, lower bounds and experiments. *Internet Mathematics*, 2(3), 2005.

[H11] Monika Henzinger. Hyperlink analysis on the world wide web. In *HYPERTEXT '05: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, pages 1–3, New York, NY, USA, 2005. ACM.

[H12] Srivatsa Iyengar. Entity reconciliation in spin. M.Tech project report, Indian Institute of Technology Bombay, 2005.

[H13] Quanzhi Li and Yi-fang Brook Wu. People search: Searching people sharing similar interests from the web. *J. Am. Soc. Inf. Sci. Technol.*, 59(1):111–125, 2008.

[H14] Yu-Ru Lin, Hari Sundaram, Yun Chi, Junichi Tatemura, and Belle L. Tseng. Detecting splogs via temporal dynamics using self-similarity analysis. *ACM Trans. Web*, 2(1):1–35, 2008.

[H15] Dmitry Lizorkin, Pavel Velikhov, Maxim Grinev, and Denis Turdakov. Accuracy estimate and optimization techniques for simrank computation. *Proc. VLDB Endow.*, 1(1):422–433, 2008.

[H16] Siddhartha Reddy, K. Srinath, Srinivasa Mandar, and R. Mutalikdesai. Measures of ignorance on the web. In *Proceedings of the International Conference on Management of Data COMAD 2006*, pages 140–149, 2006.

[H17] Elisa Rondini. Semi-automatic techniques for the semantic annotation of multimedia databases. Master's thesis, University of Bologna, 2005. (in Italian).

[H18] Takehiko Sakamoto and Keishi Tajima. Improved methods of structure calculation using link-based similarity functions. In *Proc. of the 18th IEICE Data Engineering Workshop*, 2007. (in Japanese).

[H19] T. Sarlós, A. A. Benczúr, K. Csalogány, D. Fogaras, and B. Rácz. To randomize or not to randomize: Space optimal summaries for hyperlink analysis. In *Proceedings of the 15th Int'l World Wide Web Conference*, 2006.

[H20] Allan M. Schiffman. Hierarchy in web page similarity link analysis. Technical Report 06-02, Carnegie Mellon University and CommerceNet Labs, May 2006.

[H21] Song, Ma, Lian-Li, and Zhang Zhijun. The study on the comprehensive computation of the documents similarity. *Computer Engineering and Applications*, 42(30), 2006. (in Chinese).

[H22] Jessica Staddon. Sponsored ad-based similarity: an approach to mining collective advertiser intelligence. In *ADKDD '08: Proceedings of the 2nd International Workshop on Data Mining and Audience Intelligence for Advertising*, pages 50–56, New York, NY, USA, 2008. ACM.

[H23] K Venkatraman. Pagerank by distributed computing: An empirical analysis. Technical report, Quotient Inc., 2008.

[H24] Benjamin N. Waber, John J. Magee, and Margrit Betke. Web mediators for accessible browsing. In Constantine Stephanidis and Michael Pieper, editors, *Universal Access in Ambient Intelligence Environments*, volume 4397 of *Lecture Notes in Computer Science*, pages 447–466. Springer, 2006.

[H25] Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. Block-based similarity search on the web using manifold-ranking. In *Web Information Systems - WISE 2006*, volume 4255/2006 of *Lecture Notes in Computer Science (LNCS)*, pages 60–71. Springer-Verlag, 2006.

[H26] Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. Towards a unified approach to document similarity search using manifold-ranking of blocks. *Inf. Process. Manage.*, 44(3):1032–1048, 2008.

[H27] Haixuan Yang. *Machine learning models on random graphs*. PhD thesis, The Chinese University of Hong Kong (People's Republic of China), 2007. Adviser: King, Irwin and Lyu, Michael R.

[H28] Haixuan Yang, Irwin King, and Michael R. Lyu. Predictive random graph ranking on the web. In *In Proceedings of the IEEE World Congress on Computational Intelligence (WCCI*, pages 3491–3498, 2006.

[H29] Yunming Ye, Yan Li, Xiaofei Xu, Joshua Huang, and Xiaojun Chen. MFCRank: A web ranking algorithm based on correlation of multiple features. In *Computational Linguistics and Intelligent Text Processing*, volume 3878/2006 of *Lecture Notes in Computer Science (LNCS)*, pages 378–388. Springer-Verlag, 2006.

[H30] X Yin. *Scalable Mining and Link Analysis Across Multiple Database Relations.* PhD thesis, University of Illinois at Urbana-Champaign, 2007.

[H31] Xiaoxin Yin and Jiawei Han. Exploring the power of heuristics and links in multi-relational data mining. In *LNAI 4994, Proceedings of the 17th Int'l Symposioum in Foundations of Intelligent Systems*, pages 17–27, 2008.

[H32] Xiaoxin Yin, Jiawei Han, and Philip S. Yu. Linkclus: efficient clustering via heterogeneous semantic links. In *VLDB '06: Proceedings of the 32nd international conference on Very large data bases*, pages 427–438. VLDB Endowment, 2006.

[H33] Yangbo Zhu. Distributed pagerank computation in search engine confederation. Master's thesis, Carnegie Mellon University, 2006.

Towards scaling fully personalized pagerank: Algorithms, lower bounds, and experiments. [J4]

13 independent citations

[H34] Marin Bertier, Davide Frey, Rachid Guerraoui, Anne-Marie Kermarrec, and Vincent Leroy. Personalized web search by gossiping with unknown social acquaintances. Technical Report 6878, Institut National de Recherche en Informatique et en Automatique, 2009.

[H35] Marin Bertier, Rachid Guerraoui, Anne-Marie Kermarrec, and Vincent Leroy. Toward personalized query expansion. In *Social Network Systems 2009*, 2009. to appear.

[H36] Paolo Boldi, Francesco Bonchi, Carlos Castillo, Debora Donato, Aristides Gionis, and Sebastiano Vigna. The query-flow graph: model and applications. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 609–618, New York, NY, USA, 2008. ACM.

[H37] Paolo Boldi, Roberto Posenato, Massimo Santini, and Sebastiano Vigna. Traps and pitfalls of topic-biased pagerank. In *Algorithms and Models for the Web-Graph: Fourth International Workshop, WAW 2006, Banff, Canada, November 30 - December 1, 2006. Revised Papers*, volume 4936 of *Lecture Notes in Computer Science*, pages 107–116, Berlin, Heidelberg, 2008. Springer.

[H38] Aurel Cami and Narsingh Deo. Techniques for analyzing dynamic random graph models of web-like networks: An overview. *Networks*, 51(4):211–255, 2008.

[H39] Soumen Chakrabarti. Dynamic personalized pagerank in entity-relation graphs. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 571–580, New York, NY, USA, 2007. ACM.

[H40] Prasad Chebolu and Páll Melsted. Pagerank and the random surfer model. In *SODA '08: Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1010–1018, Philadelphia, PA, USA, 2008. Society for Industrial and Applied Mathematics.

[H41] William W. Cohen. Graph walks and graphical models, 2007. Machine Learning Department, Carnegie Mellon University.

[H42] Einat Minkov. *Adaptive Graph Walk Based Similarity Measures in Entity-Relation Graphs.* PhD thesis, Carnegie Mellon University, 2008.

[H43] Purnamrita Sarkar, Andrew W. Moore, and Amit Prakash. Fast incremental proximity search in large graphs. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 896–903, New York, NY, USA, 2008. ACM.

[H44] Aixin Sun, Maggy Anastasia Suryanto, and Ying Liu. Blog classification using tags: An empirical study. In *10th International Conference on Asian Digital Libraries, ICADL 2007, Hanoi, Vietnam, December 10-13, 2007, Proceedings*, volume 4822 of *Lecture Notes in Computer Science*, pages 307–316. Springer, 2007.

[H45] Jonathan Traupman. Resisting sybils in peer-to-peer markets. In *Trust Management*, volume 238 of *IFIP International Federation for Information Processing*, pages 269–284. Springer, 2007.

[H46] Jonathan David Traupman. *Robust reputations for peer-to-peer markets.* PhD thesis, University of California at Berkeley, Berkeley, CA, USA, 2007. Adviser J.D. Tygar.

Towards scaling fully personalized pagerank. [C9]

17 independent citations

[H47] Sinan Al-Saffar and Gregory Heileman. Experimental bounds on the usefulness of personalized and topic-sensitive pagerank. In *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 671–675, Washington, DC, USA, 2007. IEEE Computer Society.

[H48] Sinan al Saffar and Gregory L. Heileman. Semantic impact graphs for information valuation. In *DocEng '08: Proceeding of the eighth ACM symposium on Document engineering*, pages 209–212, New York, NY, USA, 2008. ACM.

[H49] Reid Andersen, Christian Borgs, Jennifer Chayes, John Hopcraft, Vahab S. Mirrokni, and Shang hua Teng. Local computation of pagerank contributions. In *Proceedings of the Third Workshop on Algorithms and Models for the Web Graph*, pages 150–165, 2007.

[H50] Reid Andersen, Christian Borgs, Jennifer Chayes, John Hopcroft, Kamal Jain, Vahab Mirrokni, and Shanghua Teng. Robust pagerank and locally computable spam detection features. In *AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pages 69–76, New York, NY, USA, 2008. ACM.

[H51] Reid Andersen and Fan Chung. Detecting sharp drops in pagerank and a simplified local partitioning algorithm. In *Theory and Applications of Models of Computation, Proceedings of TAMC 2007*, pages 1–12, 2007.

[H52] Reid Andersen, Fan Chung, and Kevin Lang. Local graph partitioning using pagerank vectors. In *FOCS '06: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 475–486, Washington, DC, USA, 2006. IEEE Computer Society.

[H53] K. Avrachenkov, N. Litvak, D. Nemirovsky, and N. Osipova. Monte carlo methods in pagerank computation: When one iteration is sufficient. *SIAM J. Numer. Anal.*, 45(2):890–904, 2007.

[H54] András A. Benczúr, Károly Csalogány, and Tamás Sarlós. On the feasibility of low-rank approximation for personalized pagerank. In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 972–973, New York, NY, USA, 2005. ACM.

[H55] András A. Benczúr, Károly Csalogány, Tamás Sarlós, and Máté Uher. Spamrank - fully automatic link spam detection. In *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.

[H56] Maurice Coyle and Barry Smyth. Supporting intelligent web search. *ACM Trans. Interet Technol.*, 7(4):20, 2007.

[H57] David Gleich and Marzia Polito. Approximating personalized pagerank with minimal use of web graph data. *Internet Mathematics*, 3(3):257–294, 2006.

[H58] Rahul Sami and Andy Twigg. Lower bounds for distributed markov chain problems. *The Computing Research Repository*, abs/0810.5263, 2008.

[H59] Yang Sun, Huajing Li, Isaac G. Councill, Jian Huang, Wang-Chien Lee, and C. Lee Giles. Personalized ranking for digital libraries based on log analysis. In *WIDM '08: Proceeding of the 10th ACM workshop on Web information and data management*, pages 133–140, New York, NY, USA, 2008. ACM.

[H60] Hanghang Tong, Christos Faloutsos, Brian Gallagher, and Tina Eliassi-Rad. Fast best-effort pattern matching in large attributed graphs. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 737–746, New York, NY, USA, 2007. ACM.

[H61] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. Fast random walk with restart and its applications. In *ICDM '06: Proceedings of the Sixth International Conference on Data Mining*, pages 613–622, Washington, DC, USA, 2006. IEEE Computer Society.

[H62] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. Random walk with restart: fast solutions and applications. *Knowledge and Information Systems*, 14(3):327–346, 2008.

[H63] LV Yuanhua. A study of personalized information retrieval based on implicit feedback. Master's thesis, Institute of Software, Chinese Academy of Sciences, 2007. (in Chinese).

To Randomize or Not To Randomize: Space Optimal Summaries for Hyperlink Analysis [C14]

9 independent, 14 total citations

[H64] Reid Andersen, Christian Borgs, Jennifer Chayes, John Hopcraft, Vahab S. Mirrokni, and Shang hua Teng. Local computation of pagerank contributions. In *Proceedings of the Third Workshop on Algorithms and Models for the Web Graph*, pages 150–165, 2007.

[H65] Reid Andersen, Christian Borgs, Jennifer Chayes, John Hopcroft, Kamal Jain, Vahab Mirrokni, and Shanghua Teng. Robust pagerank and locally computable spam detection features. In *AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pages 69–76, New York, NY, USA, 2008. ACM.

[H66] Reid Andersen, Fan Chung, and Kevin Lang. Local partitioning for directed graphs using pagerank. In *Algorithms and Models for the Web-Graph*, volume 4863 of *Lecture Notes in Computer Science*, pages 166–178. Springer, 2007.

[H67] András A. Benczúr, Károly Csalogány, and Tamás Sarlós. Link-based similarity search to fight web spam. In *AIRWeb 2006, Proceedings of the Second International Workshop on Adversarial Information Retrieval on the Web*, pages 9–16, 2006.

[H68] Paul Alexandru Chirita, Claudiu S. Firan, and Wolfgang Nejdl. Personalized query expansion for the web. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 7–14, New York, NY, USA, 2007. ACM.

[H69] K. Csalogány, A. Benczúr, D. Siklósi, and L. Lukács. Semi-supervised learning: A comparative study for web spam and telephone user churn. In *Proc. of Graph Labelling Workshop and Web Spam Challenge 2007 in conjunction with ECML/PKDD 2007*, 2007.

[H70] Atish Das Sarma, Sreenivas Gollapudi, and Rina Panigrahy. Estimating pagerank on graph streams. In *PODS '08: Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 69–78, New York, NY, USA, 2008. ACM.

[H71] P. Ferragina and A. Gulli. A personalized search engine based on web-snippet hierarchical clustering. *Softw. Pract. Exper.*, 38(2):189–225, 2008.

[H72] Paolo Ferragina and Antonio Gulli. Snaket: A personalized search-result clustering engine. *CEPIS Upgrade: Next Generation Web Search*, 8(1):20–27, 2007.

[H73] D. Fogaras, B. Rácz, K. Csalogány, and T. Sarlós. Towards fully personalizing PageRank: Algorithms, lower bounds and experiments. *Internet Mathematics*, 2(3), 2005.

[H74] David Gleich and Marzia Polito. Approximating personalized pagerank with minimal use of web graph data. *Internet Mathematics*, 3(3):257–294, 2006.

[H75] Miklos Kurucz, Andras Benczur, Karoly Csalogany, and Laszlo Lukacs. Spectral clustering in telephone call graphs. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 82–91, New York, NY, USA, 2007. ACM.

[H76] Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *FOCS '06: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 143–152, Washington, DC, USA, 2006. IEEE Computer Society.

[H77] Rebecca S. Wills and Ilse C. F. Ipsen. Ordinal ranking for google's pagerank. *SIAM Journal on Matrix Analysis and Applications*, 30(4):1677–1696, 2009.

nonordfp: An FP-Growth variation without rebuilding the FP-tree [C11]

8 independent, 9 total citations

[H78] Ferenc Bodon. A trie-based apriori implementation for mining frequent item sequences. In *OSDM '05: Proceedings of the 1st international workshop on open source data mining*, pages 56–65, New York, NY, USA, 2005. ACM.

[H79] Ferenc Bodon and Lars Schmidt-thieme. The relation of closed itemset mining, complete pruning strategies and item ordering in apriori-based fim algorithms. In *In Proc. PKDD*, pages 437–444, 2005.

[H80] Li Liu, Eric Li, Yimin Zhang, and Zhizhong Tang. Optimization of frequent itemset mining on multiple-core processor. In *VLDB '07: Proceedings of the 33rd international conference on Very large data bases*, pages 1275–1285. VLDB Endowment, 2007.

[H81] Wim Pijls and Walter A. Kosters. Mining frequent itemsets: A perspective from operations research. Technical Report 24, Econometric Institute, Erasmus University Rotterdam, 2008.

[H82] Balázs Rácz, Ferenc Bodon, and Lars Schmidt-Thieme. On benchmarking frequent itemset mining algorithms: from measurement to analysis. In *OSDM '05: Proceedings of the 1st international workshop on open source data mining*, pages 36–45, New York, NY, USA, 2005. ACM.

[H83] Peter Schonhofen and Andras A. Benczur. Feature selection based on word-sentence relation. In *ICMLA '05: Proceedings of the Fourth International Conference on Machine Learning and Applications*, pages 37–42, Washington, DC, USA, 2005. IEEE Computer Society.

[H84] Takeaki Uno, Masashi Kiyomi, and Hiroki Arimura. Lcm ver.3: collaboration of array, bitmap and prefix tree for frequent itemset mining. In *OSDM '05: Proceedings of the 1st international workshop on open source data mining*, pages 77–86, New York, NY, USA, 2005. ACM.

[H85] Liqiang War and Liu Da-Xin. Study on fast algorithms for frequent itemset mining. *Journal of Harbin Engineering University*, 28(3), 2008. (in Chinese).

[H86] Zhong-ping Zhang, Li Yan, Zhi-jie Lin, and Ai-jie Wang. Frequent itemsets mining algorithm based on index arrays. *Compuer Application Research*, 26(1), 2009. (in Chinese).

Architecture for mining massive web logs with experiments [C7]

2 independent, 3 total citations

[H87] E Khorram and S M Mirzababaei. Finding an optimized discriminate function for internet application recognition. *Proceedings of the World Academy of Science, Engineering and Technology*, 4:160–163, 2005.

[H88] M Rahmati and S M Mirzababaei. Data mining on the router logs for statistical application classification. In *Proceedings of the Fourth World Enformatika Conference*, 2005.

[H89] Cs. I. Sidló and A. Lukács. Shaping sql-based frequent pattern mining algorithms. In *Knowledge Discovery in Inductive Databases*, volume 3933 of *Lecture Notes in Computer Science*, pages 188–201. Springer, 2006.

On benchmarking frequent itemset mining algorithms: from measurement to analysis [C12]

1 independent, 2 total citations

[H90] Ferenc Bodon. A trie-based apriori implementation for mining frequent item sequences. In *OSDM '05: Proceedings of the 1st international workshop on open source data mining*, pages 56–65, New York, NY, USA, 2005. ACM.

[H91] Wim Pijls and Walter A. Kosters. Mining frequent itemsets: A perspective from operations research. Technical Report 24, Econometric Institute, Erasmus University Rotterdam, 2008.

High-density compression of log files [M17]

3 independent, 4 total citations

[H92] S Grabowski and S Deorowicz. Web log compression. *Automatyka / Akademia Gorniczo-Hutnicza im. Stanislawa Staszica w Krakowie*, 11(3):417–424, 2007. (in English).

[H93] Kimmo Hatonen. *Data mining for telecommunications network log analysis.* PhD thesis, Department of Computer Science, University of Helsinki, 2009.

[H94] B. Rácz, A. Lukács, and Cs. I. Sidló. Two-phase data warehouse optimized for data mining. In *Proceedings of the First International Workshop on Business Intelligence for the Real-Time Enterprise, in conjunction with VLDB 2006*, 2006.

[H95] Przemyslaw Skibinski and Jakub Swacha. Fast and efficient log file compression. In *CEUR Workshop Proceedings of 11th East-European Conference on Advances in Databases and Information Systems (ADBIS 2007*, 2007.

Impact of Non-Poissonian Activity Patterns on Spreading Processes [J6]

6 independent, 9 total citations

[H96] Julian Candia, Marta C Gonzalez, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-Laszlo Barabasi. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015 (11pp), 2008.

[H97] A. Gautreau, A. Barrat, and M. Barthelemy. Microdynamics in stationary complex networks. *ArXiv e-prints*, November 2008.

[H98] K.-I. Goh and A.-L. Barabási. Burstiness and memory in complex systems. *EPL*, 81(4):48002, feb 2008.

[H99] A. Grabowski, N. Kruszewska, and R. A. Kosiński. Properties of on-line social systems. *European Physical Journal B*, 66:107–113, November 2008.

[H100] J. Gu, W. Li, and X. Cai. The effect of the forget-remember mechanism on spreading. *European Physical Journal B*, 62:247–255, March 2008.

[H101] Wei Hong, Xiaopu Han, Tao Zhou, and Binghong Wang. Heavy-tailed statistics in short-message communication. *Chinese Physics Letters*, 26(2):028902 (3pp), 2009.

[H102] A. Grabowskiand R. Kosinski. The SIRS model of epidemic spreading in virtual society. *Acta Physica Polonica A*, 114(3):589–596, 2008.

[H103] J. G. Oliveira and A. Vazquez. Impact of interactions on human dynamics. *Physica A Statistical Mechanics and its Applications*, 388:187–192, January 2009.

[H104] T. Zhou, H. A. T. Kiet, B. J. Kim, B.-H. Wang, and P. Holme. Role of activity in human dynamics. *EPL (Europhysics Letters)*, 82(2):28002 (5pp), 2008.

The Dynamics of Information Access on the Web [J3]

19 independent, 20 total citations

[H105] Pierpaolo Andriani and Bill McKelvey. Beyond gaussian averages: redirecting international business and management research toward extreme events and power laws. *Journal of International Business Studies*, 38(7):1212–1230, December 2007.

[H106] Julian Candia, Marta C Gonzalez, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-Laszlo Barabasi. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015 (11pp), 2008.

[H107] Raul Caruso. Information and global security, a cautionary tale. *NewsNotes of the Economists for Peace and Security*, November 2006.

[H108] C. Castellano, S. Fortunato, and V. Loreto. Statistical physics of social dynamics. *Review of Modern Physics*. (to appear).

[H109] Peter Geczy, Noriaki Izumi, Shotaro Akaho, and Koiti Hasida. Knowledge worker intranet behaviour and usability. *Int. J. Bus. Intell. Data Min.*, 2(4):447–470, 2007.

[H110] Peter Géczy, Noriaki Izumi, Shotaro Akaho, and Kôiti Hasida. Human-centric design of percipient knowledge distribution service. In *WSKS '08: Proceedings of the 1st world summit on The Knowledge Society*, pages 31–40, Berlin, Heidelberg, 2008. Springer-Verlag.

[H111] K.-I. Goh and A.-L. Barabási. Burstiness and memory in complex systems. *EPL*, 81(4):48002, feb 2008.

[H112] X.-P. Han, T. Zhou, and B.-H. Wang. Modeling human dynamics with adaptive interest. *New Journal of Physics*, 10(7):073010–+, July 2008.

[H113] Wei Hong, Xiaopu Han, Tao Zhou, and Binghong Wang. Heavy-tailed statistics in short-message communication. *Chinese Physics Letters*, 26(2):028902 (3pp), 2009.

[H114] Yoram M. Kalman. *Silence in Text Based Computer Mediated Communication: The Invisible Component*. PhD thesis, University of Haifa, 2007.

[H115] Yoram M. Kalman, Gilad Ravid, Daphne R. Raban, and Sheizaf Rafaeli. Pauses and response latencies: A chronemic analysis of asynchronous cmc. *Journal of Computer-Mediated Communication*, 12(1):1–23, 2009.

[H116] Andreas Kaltenbrunner. *Dynamics of message interchange between stochastic units in the contexts of human communication behaviour and spiking neurons*. PhD thesis, Universitat Pompeu Fabra, 2007.

[H117] Andreas Kaltenbrunner, Vicenc Gomez, and Vicente Lopez. Description and prediction of slashdot activity. In *LA-WEB '07: Proceedings of the 2007 Latin American Web Conference*, pages 57–66, Washington, DC, USA, 2007. IEEE Computer Society.

[H118] Andreas Kaltenbrunner, Vicenç Gómez, Ayman Moghnieh, Rodrigo Meza, Josep Blat, and Vicente López. Homogeneous temporal activity patterns in a large online communication space. In *Proceedings of the BIS 2007 Workshop on Social Aspects of the Web*, volume 245 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2007.

[H119] J. G. Oliveira and A. Vazquez. Impact of interactions on human dynamics. *Physica A Statistical Mechanics and its Applications*, 388:187–192, January 2009.

[H120] Filippo Radicchi. Human activity in the web, 2009.

[H121] D. Ralt. No netting, health and stress - studying wellness from a net perspective. *Med. Hypotheses*, 70(1):85–91, 2008.

[H122] Xiaodan Song, Yun Chi, Koji Hino, and Belle L. Tseng. Information flow modeling based on diffusion rate for prediction and ranking. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 191–200, New York, NY, USA, 2007. ACM.

[H123] B. Ulicny, K. Baclawski, and A. Magnus. New metrics for blog mining. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 6570 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, April 2007.

[H124] T. Zhou, H. A. T. Kiet, B. J. Kim, B.-H. Wang, and P. Holme. Role of activity in human dynamics. *EPL (Europhysics Letters)*, 82(2):28002 (5pp), 2008.

References

[1] Micah Adler and Michael Mitzenmacher. Towards compressing web graphs. In *Data Compression Conference*, pages 203–212, 2001.

[2] Luiz André Barroso, Jeffrey Dean, and Urs Hölzle. Web Search for a Planet: The Google Cluster Architecture. *IEEE Micro*, 23(2):22–28, 2003.

[3] Paolo Boldi and Sebastiano Vigna. The WebGraph framework I: Compression techniques. Technical Report 293-03, Universita di Milano, Dipartimento di Scienze dell'Informazione, 2003.

[4] E. Brewer. Lessons from giant-scale services.

[5] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

[6] Andrei Z. Broder. On the Resemblance and Containment of Documents. In *Proceedings of the Compression and Complexity of Sequences (SEQUENCES'97)*, pages 21–29, 1997.

[7] A. L. Buchsbaum, R. Giancarlo, and J. R. Westbrook. On finding common neighborhoods in massive graphs. *Theoretical Computer Science*, 299(1-3):707–18, 2004.

[8] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified data processing on large clusters. In *Proceedings of the 6th USENIX Symposium on Operating System Design and Implementation (OSDI)*, San Francisco, CA, USA, 2004. USENIX Association.

[9] Dániel Fogaras. Where to start browsing the web? In *Proceedings of the 3rd International Workshop on Innovative Internet Community Systems (I2CS)*, volume 2877/2003 of *Lecture Notes in Computer Science (LNCS)*, pages 65–79, Leipzig, Germany, June 2003. Springer-Verlag.

[10] Taher H. Haveliwala. Topic-sensitive PageRank. In *Proceedings of the 11th World Wide Web Conference (WWW)*, pages 517–526, 2002.

[11] Taher H. Haveliwala. Efficient encodings for document ranking vectors. In *Proceedings of the 4th International Conference on Internet Computing (IC)*, pages 3–9, Las Vegas, Nevada, USA, 2003.

[12] Taher H. Haveliwala, Aristides Gionis, Dan Klein, and Piotr Indyk. Evaluating strategies for similarity search on the web. In *Proceedings of the 11th World Wide Web Conference (WWW)*, pages 432–442, 2002.

[13] Taher H. Haveliwala, Sepandar Kamvar, and Glen Jeh. An analytical comparison of approaches to personalizing PageRank. Technical Report 2003-35, Stanford University, 2003.

[14] Monika R. Henzinger, Prabhakar Raghavan, and Sridhar Rajagopalan. Computing on data streams. In *External Memory Algorithms, DIMACS Book Series vol. 50.*, pages 107–118. American Mathematical Society, 1999.

[15] Glen Jeh and Jennifer Widom. SimRank: A measure of structural-context similarity. In *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 538–543, 2002.

[16] Glen Jeh and Jennifer Widom. Scaling personalized web search. In *Proceedings of the 12th World Wide Web Conference (WWW)*, pages 271–279. ACM Press, 2003.

[17] Sepandar Kamvar, Taher H. Haveliwala, Christopher Manning, and Gene Golub. Exploiting the block structure of the web for computing PageRank. Technical Report 2003-17, Stanford University, 2003.

[18] Sepandar D. Kamvar, Taher H. Haveliwala, Christopher D. Manning, and Gene H. Golub. Extrapolation methods for accelerating PageRank computations. In *Proceedings of the 12th World Wide Web Conference (WWW)*, pages 261–270. ACM Press, 2003.

[19] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the Web for emerging cyber-communities. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11–16):1481–1493, 1999.

[20] Eyal Kushilevitz and Noam Nisan. *Communication Complexity.* Cambridge University Press, 1997.

[21] Ulrich Meyer, Peter Sanders, and Jop Sibeyn. *Algorithms for Memory Hierarchies, Advanced Lectures.* Springer-Verlag, Berlin, 2003.

[22] Open Directory Project (ODP). `http://www.dmoz.org`.

[23] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford University, 1998.

[24] Christopher R. Palmer, Phillip B. Gibbons, and Christos Faloutsos. ANF: A fast and scalable tool for data mining in massive graphs. In *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 81–90. ACM Press, 2002.

[25] Tamás Sarlós, András A. Benczúr, Károly Csalogány, Dániel Fogaras, and Balázs Rácz. To randomize or not to randomize: Space optimal summaries for hyperlink analysis. In *Proceedings of the 15th International World Wide Web Conference (WWW)*, pages 297–306, 2006. Full version available at http://www.ilab.sztaki.hu/websearch/Publications/.

[26] Pavan Kumar C. Singitham, Mahathi S. Mahabhashyam, and Prabhakar Raghavan. Efficiency-quality tradeoffs for vector score aggregation. In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB)*, pages 624–635, 2004.