State of the Art in Web Information Retrieval*

Sérgio Nunes
Doctoral Program in Informatics Engineering
Faculty of Engineering, University of Porto
Rua Dr. Roberto Frias, s/n 4200-465 Porto
PORTUGAL

June 2006†

Abstract

The main Information Retrieval (IR) concepts and theories are introduced along with the specific case of WebIR. Seminal and landmark research on WebIR is discussed along with more recent work on this subject. Several of the most active research centres, companies and researchers are presented and their work is discussed. Related research areas are also briefly described. Through all the report a special emphasis is placed on the concept of temporal analysis of the web for information retrieval.

1 Introduction

This is a state of the art report on the broad subject of Information Retrieval on the Web (WebIR). Information Retrieval deals with the representation, storage, organization and retrieval of information existing in unstructured supports. Web Information Retrieval is the application of IR to the World Wide Web (Web).

This paper starts with a brief overview of the Information Retrieval (IR) field (Section 2), including references to the main models, the common methodologies and to the evaluation strategies. In Section 3 the World Wide Web (Web) is presented considering two perspectives - Structure and Dynamics. How IR is applied to the web is the focus of Section 4, including references to typical WebIR tasks, such as web crawling, content analysis, link analysis and temporal link analysis. In Section 5 several topics related to WebIR are presented, including topic detection and tracking, information storage, bibliometrics and digital preservation.

The main research groups, researchers and projects are presented in Sections 6, 9 and 7. Commercial companies related to IR and WebIR are presented in Section 8. The main conferences and journals where research related to these topics appear are listed in Sections 10 and 11.

*Internal Technical Report submitted to fulfill the requirements of the Doctoral Program in Informatics Engineering of the Faculty of Engineering, University of Porto, Portugal.

†First published in June, 2006. Minor updates in March, 2007.

2 Information Retrieval

Information Retrieval (IR) is concerned with the creation, storage, organization, and retrieval of information. The focus of IR is on unstructured information, like document collections or the web. Structured information retrieval, usually named Data Retrieval, consists of retrieving all items that satisfy a clearly defined expression (e.g. SQL). The goal of IR is to provide users with those documents that will satisfy their information need. The information need is typically expressed in natural language, not always well structured and often semantically ambiguous.

2.1 Models

The primary goal of an IR system is to retrieve all the documents that are relevant to a user query while minimizing the number of non-relevant documents retrieved. Different approaches to this problem yield distinct information retrieval models. These models are usually classified according to their mathematical basis in three different categories [24]. Other classifications are possible, for example according to the logical view of the documents or the properties of the model.

Set-theoretic Models Documents are represented by sets that contain terms. Similarities are derived using set-theoretic operations. Implementations of these models include the Standard Boolean Model, the Extended Boolean Model and the Fuzzy Model.

Algebraic Models Documents are represented as vectors, matrices or tuples. These are transformed using algebraic operations to a one-dimensional similarity measure. Implementations include the Vector Space Model and the Generalized Vector Space Model.

Probabilistic Models Document's relevance is interpreted as a probability. Documents and queries similarities are computed as probabilities for a given query.

These models are used for content analysis in classical IR, specific models for web document modeling are discussed in Section 4. Due to its simplicity and efficient computation, the Vector Model [127] is the most widely used model in IR.

2.2 Research Methodology

This section presents a brief overview of the typical research methodology adopted in IR. Below is a description of each of these steps along with a summary.

Task Definition Tasks are defined according to research needs. See Section 4.2 for examples of tasks.

Build Collections For each task, collections of documents are built along with typical queries. For each query, domain experts select the relevant documents from the collection.

	Retrieved	Not Retrieved
Relevant	A	B
Not Relevant	C	D

Table 1: Evaluation in Information Retrieval

Run Systems Systems or algorithms are tested using the predefined queries on each collections. Results of each run are recorded.

Evaluate Results Obtained results are compared with expected results (see Section 2.3).

Research in IR is very experimental and, more important, is very broad and heterogeneous. Different IR tasks typically require very different approaches. For example, there are very few similarities between an ad-hoc retrieval task and a filtering task.

2.3 Evaluation

Evaluation of results plays a major role in every research activity. In the field of IR, tests are typically performed using three common elements: a dataset, a set of queries and a set of answers. The answers are the relevant documents as pre-determined by domain experts. The system or algorithm to be tested is run using these elements.

Before the 1990s, this type of IR evaluation was carried out by individuals and small groups [144]. Due to the high cost of building and maintaining evaluation sets, these were relatively small in scope. In 1991, with the creation of the TREC conference series (see Section 10), this problem was addressed and large test collections were created and made available for researchers worldwide. Voorhees [140] presents a review of this paradigm in the context of the evaluation conferences.

In Table 1 four typical IR values are defined. A, for example, represents the number of documents that were both retrieved and relevant to the query. $A+C$ is the total number of retrieved documents. $A + B + C + D$ is, obviously, the number of documents in the collection. Below is a list of classic IR measures for evaluating system's effectiveness.

Recall $\frac{A}{A+B}$ Proportion of relevant documents that are retrieved.

Precision $\frac{A}{A+C}$ Proportion of retrieved documents that are relevant.

Miss $\frac{B}{A+B}$ Proportion of relevant documents that not are retrieved.

False Alarm or Fallout $\frac{C}{C+D}$ Proportion of non-relevant material that is retrieved.

Richness or Generality $\frac{A+B}{A+B+C+D}$ Proportion of the collection that is relevant.

Due to the dynamic nature of the web, to the fact that only part of the documents are indexed and to the broad range of users accessing search engines,

these measures are less fitted for WebIR evaluation. Also, in such a highly interconnected and vast collection, relevance is a much more subjective measure. For example, non-relevant documents connected to relevant documents are partially relevant. Since the web is mostly used by non-experts in information searching, criterias related to presentation and user effort assume greater importance. For example, good *response times* are reportedly [102] one of the most important features to users.

Due to the lack of widely accepted benchmarks and methods for search engine evaluation, coverage (or index size) is frequently used as a measure to compare services [26]. However, coverage is mostly related to the performance of the crawler rather than the performance of the retrieval system in itself [78]. When this measure is used, page duplication and spam need to be taken into account. For example, a more aggressive spam filtering strategy might result in a smaller (higher quality) index.

Gordon et al. [70] distinguish two types of search engine evaluation: *testimonials* and *shootouts*. The former is based on direct experience and the informal evaluation of features lists, thus highly subjective. The latter is more closely related to traditional IR methodologies. Shootouts use labor intensive methods based on sets of queries or on fixed TREC style datasets. Hawking et al. [78] present a detailed survey on these methods and results from several studies. Nevertheless, the results of these evaluations do not have a statistical support [26].

On a different level, and when details are available, search engine algorithms are compared using different factors, including the computational resources required, the similarity among rankings within experimental setups and the subjective judgment about highly ranked pages [55].

Automated means for large-scale evaluation of search results are needed and remains an open problem [126].

3 World Wide Web

The World Wide Web was invented in 1990 by Tim Berners-Lee et al. [33]. It is now the biggest service on the Internet, either in the number of users and in the amount of data managed. In ten years it has grown from a small service used by researchers and academics to a worldwide system used by millions of people.

The remainder of this chapter is organized in two sections, focusing on the structure of the web and the dynamics of the web.

3.1 Web Struture

This section presents several studies related to the structure of the web. The focus of research has been on measuring the size of the web and its link structure. The web can be characterized from multiple perspectives using numerous metrics. Due to the large dimension and permanent evolution of the web, this is a difficult task.

In 1997, Bharat et al. [34] used five search engines to estimate the relative size and overlap of public search engines. After generating a lexicon of about 400,000 terms, queries were performed in each of the selected search engines. Then, sampled URLs were selected to check for containment in all the search

Work	Period	Estimated Size
Bharat [34]	1997-11	200,000,000
Lawrence [108]	1999-02	800,000,000
Gulli [73]	2005-01	11,500,000,000

Table 2: Studies on the Web's Size

engines. Using statistical analysis they estimate the size of the static public web to be at least 200 million documents. In 2005, Gulli et al. [73] adopted the same methodology and state that the indexable web includes more than 11.5 billion pages.

One of the most complete studies on the web's structure was done by Broder et al. [39] in 1999. Two main contributions were presented in this work, the estimation that the fraction of web pages with i in-links is proportional to $\frac{1}{i^{2.1}}$ and the proposal of a coherent macrostructure for the web. In this proposed structure (called "bow-tie model"), a single large strongly connected component (called MAIN) is identified. This "'giant strongly connected component' (SCC) is at the heart of the web". Two other components are recognized - IN and OUT. IN consists of pages that can reach the SCC, but cannot be reached from it, and OUT consists of pages that are accessible from the SCC, but do not link back to it. It is worth transcribing a significant statement from this work - "the web is not the ball of highly-connected spaghetti we believed it to be; rather, the connectivity is strongly limited by a high-level global structure" [39].

In a 1999 study published by Nature, Lawrence et al. [108] studied the distribution of information on the web. The publicly indexable web was estimated to contain 800 million pages, encompassing about 15 terabytes [1] (TB) of data or about 6 TB of plain text (after removing HTML tags). It was found that 83% of sites contains commercial content, 6% contains scientific or educational content and only 1.5% contains pornographic content.

Bharat et al. [35] also studied the replication of content on the web and found that 10% of the web's content was mirrored. Previous studies [40, 129] have found a large amount of page duplication on the web.

Kumar et al. [106] explore the web graph to find emerging communities. In their work, the co-citation concept is used to find nodes that share common interest. More than 100.000 small emerging communities were enumerated. It was found that the web harbors a large number of communities, each of which manifests itself as a set of interlinked web pages.

A detailed characterization of a national community web is presented by Gomes et al. [66]. In this study it is observed that most sites are small virtual hosts under the `.pt` domain and that the number of sites under construction is very high.

On a related topic is the work published by Google [68] on the statistical analysis of HTML content.

In Table 2, a brief summary of the cited works and their estimation on the web's size is presented. Regarding web's size, notice that the deep web might be at least 100 time larger that the static web [77]. Deep Web is a term commonly used to refer to the vast repositories of content that are inaccessible to search engines, such as documents in databases.

[1] 1 terabyte = 2^{40} bytes = $1,000,000,000,000$ bytes.

3.2 Web Dynamics

In this section, studies related to the change, activity or progress of the web are presented. The evolution of the web has been studied with the main goal of modeling its behavior. Research has been focused on the persistence of both content and URL through time. Several values for URL's half-life [2] have been advanced. It is commonly accepted that the web is a highly dynamic environment.

Ntoulas et al. [122] have analyzed the evolution of both content and link structure of web pages, specially focusing on aspects of potential interest to search engine designers. 150 web sites were crawled and indexed weekly over the course of one year (starting in late 2002), generating more than 3.3 TB of data. For each web page two measures were analyzed: change frequency (when) and change degree (how much). The degree of change was measured with high detail, using the TF.IDF weighted cosine distance (see Section 4.4) to compute this change. Their main findings were: only 20% of pages lasted one year; after a year, 50% of the content on the web is new; link structure is more dynamic than page content; 25% of links change every day. For example, it was found that creation of new pages is much more frequent than updating existing pages. This was one of the first works to study the evolution of web link structure experimentally. It was observed that existing pages are being removed from the web and replaced by new ones at a very rapid time. These figures have a direct impact on search engine design, namely crawling strategies and scheduling.

Gomes et al. [67] have modeled URL and content persistence using data from Tumba's archive. Their study uses an exhaustive crawl of a regional web domain (`.pt`), while previous studies have relied on a selection of specific URLs. It was found that persistent URL tend to be static, short and linked from other sites. While persistent content tends to be small, not dynamically generated and have a Last-Modified header defined. Interestingly, it was found that lasting contents tend to be referenced by different URL during their lifetime (domain or content management system changes are among the main reasons).

Several studies [111, 133] have estimated the half-life of URL to be between four and five years. In [111] it was found that more than half the pages being tracked in the .COM domain disappeared in 24 months. As time passes, link rot is expected to increase [55]. A recent work from Gomes et al. [67] states that URL and content persistence is much lower that previous estimates. According to this work, URL half-life is 2 month, while web site half-life is 556 days.

Recent studies [105] have analyzed the Blogspace and note that blogs exhibit a striking temporal characteristic. In this work it is found that 2001 marks a clear change in the Blogspace, when burstiness and connectivity increased significantly.

A small summary of several studies on web dynamics is presented in Table 3.

The diversity of methodologies adopted to model and analyze the dynamics of the web has led to very different results. Also, the different sampling techniques used for producing collections have a significant impact on the final results [75]. Result differences can also be explained by the algorithms used to,

[2] In this context, half-life is a measure that indicates how much time is needed so that the size of a collection reaches half size. For example, if the half-life of a collection of URLs is approximately 1 year from its publication date, it means that 1 year after publication 50% of these URLs are not accessible.

Work	Period	Total Crawls	Total URL (average)
Koehler [104]	1996-12/2001-02	217	361
Cho [46, 47]	1999-02/1999-06	128	720.000
Baeza-Yates [25]	2000-01/2001-06	2	732.500
Chien [44]	2000-05/2000-11	2	61,000,000
Ntoulas [122]	2002-10/2003-11	51	4,400,000
Gomes [67]	2002-11/2005-07	8	6,200,000
Fetterly [57]	2002-11/2002-12	11	151,000,000

Table 3: Studies on Web Dynamics

for example, compute page change (e.g. hash values, shingling).

4 Web Information Retrieval

The growth of the web as a popular communication medium has fostered the development of the field of Web Information Retrieval (WebIR). Also, web search engines are the "killer application" for the web. Users spend most of their time using search engines. WebIR can be defined as the application of theories and methodologies from IR to the World Wide Web. However, compared with classic IR, WebIR face several different challenges. Below are the main differences between IR and WebIR.

Size The size of the web is estimated to be 11.5 billion documents in 2005. After a year, about 50% of the content on the web is new.

Structure Links between documents exhibit unique patterns on the web. There are millions of small communities scattered through the web.

Dynamics The web exhibits a very dynamic behavior. Significant changes to the link structure occur in small periods of time (e.g. week). Also, URL and content have a very low half-life.

Heterogeneity The web is a very heterogeneous environment. Multiple types of document formats coexist in this environment, including HTML, PDF, images, Flash. The web also hosts documents written in a variety of languages.

Duplication Several studies indicate that nearly 30% of the web's content is duplicated, mainly due to mirroring.

Users Search engines deal with all types of users, generally performing short ill-formed queries. Web information seeking behaviors also have specific characteristics. For example, users rarely pass the first screen of results and rarely rewrite their original query.

In the next section, a general overview of the main components of a WebIR system is presented. In Section 4.2, typical WebIR tasks are presented and defined. In Section 4.3 a special emphasis is given to web crawling. In the following two sections (4.4 and 4.5) content and link analysis are presented. Greated attention is given to link analysis, thus two of the main algorithms are explained.

4.1 WebIR Components

To address the challenges found in WebIR, web search systems need very specialized architectures [38, 131]. Overall, search engines have to address all these aspects and combine them in a unique ranking. Below is a brief description of the main components of such systems.

Crawler Includes the crawlers that fetch web pages. Typically multiple and distributed crawlers operate simultaneously. Current crawlers continuously harvest the web, scheduling operations based on web sites profiles.

Repository Fetched web documents are stored in a specialized database, allowing high concurrent access and fast reads. Full HTML texts are stored here.

Indexes An indexing engine builds several indices optimized for very fast reads. Several types of indices might exist, including inverted indices, forward indices, hit lists, lexicons. Documents are parsed for content and link analysis. Previously unknown links are feed to the crawler.

Ranking For each query, ranks the results combining multiple criteria. A rank value is attributed to each document.

Presentation Sorts and presents the ranked documents. Short snippets of content from each document are selected and included in this final step.

All of these aspects have contributed to the emergence of WebIR as a very active field of research. Multiple areas of expertise are combined in one unified field. The following sections present a broad overview of this field. The next section presents a summary of the typical WebIR tasks. The concept of Web Crawling is introduced in Section 4.3. Section 4.4 present the main concepts in Content Analysis. Link Analysis is presented, along with the main algorithms in the field, in Section 4.5. Finally, in Section 4.6, a detailed survey on Temporal Link Analysis is made.

4.2 WebIR Tasks

Web Information Retrieval research is typically organized in tasks with specific goals to be achieved. This strategy has contributed to the comparability of research results (see TREC in Section 10) and has set a coherent direction for the field. Existing tasks have changed frequently over the years due to the emergence of new fields. Below is a summary of the main tasks and also of the new or emerging ones.

Ad-Hoc Rank documents using non-constrained queries in a fixed collection. This is the standard retrieval task in IR.

Filtering Select documents using a fixed query in a dynamic collection. For example, "retrieve all documents related to 'Research in Portugal' from a continuous feed".

Topic Distillation Find short lists of good entry points to a broad topic. For example, "find relevant pages on the topic of Portugal Geography".

Homepage Finding Find the URL of a named entity. For example, "find the URL of the European Commission homepage".

Adversarial WebIR Develop methods to identify and address the problem of web spam, namely link spamming, that affect the ranking of results.

Summarization Produce a relevant summary of a single or multiple documents.

Visualization Develop methods to present and interact with results.

Question Answering Retrieve small snippets of text that contained an answer for open-domain or closed-domain questions.

TREC has introduced a new Blog Track in 2006 with the purpose of exploring "information seeking behavior in the blogosphere".

Sahami et al. [126] identify several open research problems and applications, including inferring meaning in text, link spam detection, adversarial classification and automated evaluation of search results. According to these authors, information retrieval on the web is still a fertile ground for research.

4.3 Web Crawling

A web crawler is a software program that browses and stores web pages in a methodical and automated way [103]. Typically, web crawlers start with a set of well known URL (seeds) and then fetch and parse each page iteratively, storing new URL for later processing. There are several problems to be addressed in the design, including the selection algorithm (e.g. breadth-first), re-visit policy, dealing with the deep web, URL identification.

Web crawling is an active topic of research in the field of WebIR [45, 43, 48, 85, 46]. Research has mostly dealt with the efficient use of resources and the scheduling of operations. There are several open source web crawlers available.

Web crawling strategies can have a high impact on the final collections. Gurrin et al. [75] analyzed the poor results achieved when applying link analysis to the TREC's web datasets. Their research showed that link density and distribution in these collections was not representative of the web's.

Ntoulas et al. [122] research presents contrary evidence to the notion that update frequency is related to the degree of content change. The degree of content shift does not appear to be directly related to the frequency of content update. Most content changes, although frequent, have very little impact on the page's content. are very small, namely updating visits counters, today's date or it is common to have very small changes introduced to web pages.

4.4 Content Analysis

First web search systems were based on content analysis and used a simple collection of word to rank documents [72]. Term weighting methods [128] like TF.IDF measures were used.

Second generation systems explored the HTML structure of the documents [52]. For example, a term enclosed in a heading or title HTML tag has a higher value than the same term in a paragraph.

4.5 Link Analysis

The analysis of the hyperlink structure of the web has led to significant improvements in web information retrieval [79]. A significant number of algorithms for ranking web results exploit the linkage information inherent in the structure of the web. The most significant endeavors using this concept are the PageRank [38] and the HITS [101] algorithms. In this section these algorithms are explained. There are several alternative link-based algorithms for ranking web results: SALSA [109], BHITS [37], PHITS [50] and TrustRank [76]. Most of these are improvements or variations of PageRank and HITS.

Approaches based on the link structure of the web consider the web as a directed graph $G(V, E)$, consisting of a set of nodes V and a set of edges E. N is the total number of nodes. Intuitively, each node models a web page and each link between two pages models a specific directed edge. For each vertex V_i, let $In(V_i)$ be the set of vertices that point to it (inlinks) and let $Out(V_i)$ be the set of vertices pointed to by V_i (outlinks).

4.5.1 PageRank

PageRank, proposed by Brin et al. [38] in 1998, is one of the most significant algorithms based on link analysis. It is used by the Google search engine to rank web results [69]. The algorithm produces a final rank for each web page, its PageRank value (PR). PageRank rank is more a paradigm than a specific algorithm since there are multiple variations on the same concept [55]. Here, the generic concept is presented.

PageRank is largely inspired in bibliometrics literature [59], where documents are ranked higher according to the number of references to it. With PageRank, each page's value is distributed uniformly to its outlinks. This concept is formalized in Equation 1. The parameter d is a damping factor which can be set to between 0 and 1[3]. Since the sum of all PageRank values must be 1, each page has an initial value of $\frac{1}{N}$.

$$PR(V_i) = (1 - d) + d * \sum_{V_j \in In(V_i)} \frac{PR(V_j)}{|Out(V_j)|} \quad (1)$$

Consider, for example, a small subset of four pages: A, B, C and D. The initial PageRank value of each page will be $1/4 = 0.25$. Suppose that A, B and C point to D. Hence, $PR(D) = \frac{0.25}{1} + \frac{0.25}{1} + \frac{0.25}{1} = 0.75$. Alternatively, imagine that A also links to B and that C points to A. In this case, $PR(A) = \frac{PR(C)}{2} = 0.125$, $PR(B) = \frac{PR(A)}{2} = 0.0625$. Finally, $PR(D) = \frac{PR(A)}{2} + \frac{PR(B)}{1} + \frac{PR(C)}{2} = 0.0625 + 0.0625 + 0.125 = 0.25$. For simplification, the damping factor was not included in this example. This factor establishes a minimum value for each node, preventing PageRank values of zero for isolated nodes [4].

The recursive definition of this algorithm results in an implementation that iterates until convergence below a given threshold is achieved. Typically, convergence is reached after few iterations. There have been several papers that

[3] The original paper suggests setting this value to 0.85.

[4] Equation 1 was presented by Brin et al. [38], but to support the original statement that "the sum of all PageRanks is one", the dampening factor should be divided by N ($\frac{(1-d)}{N}$).

address the problem of efficiently implement PageRank [55, 32]. The PageRank algorithm is run offline, not at query time.

From the conceptual point of view, PageRank is defined as the time that a user spends on a given page while performing a random walk through the web graph. Now, the d parameter can be seen as a "teleportation factor", needed to avoid dead ends in this random walk and to reach isolated sections of the graph (islands). PageRank can also be seen as the probability of the random walker being on a page.

The PageRank approach is biased against new pages [25]. Both new pages and very old pages exhibit a very low PageRank value, close to the minimum. Also, the peak of PageRank is in three months old pages. This happens because older pages have accumulated more inlinks over time. New high-quality pages might have few inlinks due to several reasons (e.g. haven't been found by other webmasters, webmasters haven't updated their pages or the search engine hasn't performed a new crawl). Very old pages are likely to be abandoned and thus not linked anymore by other pages.

One of the main problems with the PageRank paradigm is its vulnerability to direct manipulation. This practice is widely known as link spamming and its detection is an open research problem [126]. Different implementations of PageRank have tried overcome this limitation. Eiron et al. [55] suggest that new methods of ranking, that are motivated by the hierarchical structure of the web (HostRank, DirRank), may be more resistant to direct manipulation. Link spam detection is an active area of research.

4.5.2 HITS

The Hyperlinked Induced Topic Selection (HITS) algorithm was proposed by Kleinberg [101] in 1999. This algorithm produces two values for each page, an *authority* value and a *hub* value. For a given topic, an authority is a page with a large number of incoming links from hubs, while a hub is a page with a large number of outgoing links to authorities. The HITS algorithm first collects a base document set for each query. Then it recursively calculates the hub and authority values for each document until convergence is achieved. After setting up initial values for each node, Equations 2 and 3 are used to calculate each node's authority value and hub value.

$$HITS_{Authority}(V_i) = \sum_{V_j \in In(V_i)} HITS_{Hub}(V_j) \tag{2}$$

$$HITS_{Hub}(V_i) = \sum_{V_j \in Out(V_i)} HITS_{Authority}(V_j) \tag{3}$$

The HITS algorithm differs from PageRank in three significant aspects, it is executed at query time, it computes two different values for each page and it is applied to a subgraph (typically 1000-5000 nodes [56]) of the web.

The Teoma search engine, subsequently acquired by Ask Jeeves [91], used a similar algorithm to rank results.

4.6 Temporal Link Analysis

Tradicional WebIR research makes use of static snapshots of the web. Only recently researchers have begun to address questions related to the dynamics of the web using multiple snapshots.

In a recent paper, Amitay et al. [17] present a broad work on the subject of temporal link analysis. In this paper, the authors use the HTTP header field "last modified" to approximate the age of the page's content. Using this information to timestamp web resources, several interesting applications are explored. Authors clearly show that real life events can be exposed mainly due to "fossilized" content. Also worth noting is the concept of "Timely Authorities", opposed to simple link based "Authorities". This idea is illustrated with the adaptation of the HITS and SALSA algorithms, adjusting vertices weights to include a time dependent bonus. The evaluation of this experiment, along with the adaptation of the PageRank algorithm, is suggested as future work.

Several recent works have looked at blogs as a good ecosystem to explore the web's dynamic nature. The fact that blogs generally have timestamps included with posts makes this a very useful collection in the analysis of the evolution of the web.

Kumar et. al [105] propose the concept of *time graphs* for the study of graphs that evolve in continuous time. After building a blog graph (the time graph for Blogspace), Kleinberg's [99] approach is used to detect bursty behavior. The results confirm that blogs exhibit a striking temporal characteristic and that 2001 marks a clear change in the Blogspace, when both burstiness and connectivity increased significantly.

A recent work from Nakajima et al. [118] explores citations and co-citations on multiple crawls of blog entries to build blog threads. Using this temporal graph, different heuristics are proposed to find thread agitators and thread summarizers. Manual evaluation of the top results indicates that this approach might be a good strategy to find influent bloggers.

Berberich et al. [28, 29] propose a new algorithm (T-Rank), that extends the PageRank technique to explore the user's time windows in the ranking of results. Temporal annotations are added to the nodes and edges of the web graph. Two experiments were performed to access the quality of the technique. One used the DBLP dataset, mapping authors as nodes and citations as edges, another used a crawl to Amazon's products pages, products were mapped as nodes and recommendations as edges. Meaningful rankings were produced in both cases. Two directions were identified for future work: experiment on a large scale web dataset and extend the algorithm to assess emerging authorities.

In a different line of research, Berberich et al. [27] devised a method that builds upon PageRank's scores over time. Since these results are not directly comparable, a normalization method is proposed based on a relevant growth model of importance. Since no adequate web dataset was available, the DBLP bibliographic dataset was used in the analysis. Experiments show interesting results. While with PageRank the same item was always the top-results. Using BuzzRank, different items were retrieved in each year during the period from 89 to 99. Each snapshot was modeled as an independent graph.

In Gruhl et al. [71] previous research from a variety of fields that address the problem of propagation through networks (thermodynamics, epidemiology and marketing) was used in the context of the Blogspace. This work addressed

snapshot models, which focus on short term behavior (weeks or months), leaving long-term analysis (horizon models) as an open research problem. The propagation of information through the Blogspace was modeled using a corpus of 401,021 blog posts. Understanding what causes resonance, a sharply reaction from the community caused from little or no external input, was identified as an interesting problem for future research.

5 Related Topics

5.1 Topic Detection and Tracking

Research into Topic Detection and Tracking (TDT) began in 1996 [14] and the purpose of the study was to determine the effectiveness of state-of-the-art technologies toward addressing several event-based information organization tasks. Examples of TDT applications include story segmentation, topic tracking, topic detection, first story detection and link detection.

Kleinberg's text on the temporal dynamics of on-line information streams [100] is a recent survey on this subject. Kleinberg states that recent "developments have led to a shift in our working metaphor for Internet and web information, from a relatively static one [...] to a much more dynamic one", and that "the 'time axis' of information is increasingly visible". In this work, the basic techniques are grouped according to six strategies: topic detection and tracking, information visualization, timelines and threshold-based methods, state-based methods, trend-based methods and two-point trends. Also discussed in this work are recent applications of these techniques, namely: weblogs, search engine queries and usage data.

Kleinberg also refers that analyzing the temporal properties of information streams is part of the broader area of sequential pattern mining (data mining) and can be viewed as an application of time-series analysis (statistics).

Allan et al. [15] reduced the task of First Story Detection (FSD) to Topic Tracking and showed that effective FSD is either impossible or requires substantially different approaches. It was suggested that improvements are likely to come from exploring task-specific information about how news topics and events are related and defined.

5.2 Information Storage

In this section, a brief survey of current trends in information storage is presented. This survey is focused on solutions for storing large collections of data. Overall, information storage has been shifting from custom made hardware to the adoption of commodity hardware. Replacing off-the-shelf hardware is much cheaper than repairing or buying specialized hardware. Thus, there are several file systems that have abstracted high-performance API over simple commodity machines.

For example, the Internet Archive hosts one of the largest data repositories in the word (at least 2 petabytes [5] and growing 20 TB each month [88]). Its development was made using commodity hardware and open source software [90]. Currently, in a joint project with the University of Cornell, this data is being

[5] 1 petabyte = 2^{50} bytes = $1,000,000,000,000,000$ bytes.

made accessible to researchers worldwide [22]. Recently, a new machine was designed to safely store and process one petabyte - Petabox [89].

The Google File System (GFS) [61] was developed for supporting large-scale data processing workloads on commodity hardware (e.g. SATA disks). Traditional design principles in file system design are reviewed in the light of today's technological environments and Google's specific needs. Hardware failures are treated as the norm rather than the exception. To achieve high read and write throughput, file system control is separated from data transfer. After querying the central control server (exchanging a very small amount of data), further interactions are performed directly with the data servers. The GFS development team has placed a strong emphasis on costs control, thus it has a very low cost per GB.

Gomes et al. [64] describe a system (Webstore), used at the Tumba! search engine, that works at the application level and detects duplicates on the fly (during the crawl), before any I/O operation. Webstore largely outperforms a previous NFS based system on several tests (read, write, delete). This system does not stores deltas (differences between versions of documents) since these add an extra layer of dependence and introduce a significant overhead cost. Duplicates are detected comparing document's MD5 signatures. Safe modes of operation, developed to avoid collisions, are also presented. Eliminating duplicates produces significant storage space savings.

6 Research Groups

In this section, the top research groups in the field of WebIR are listed and briefly presented. The Stanford University InfoLab [94], formerly named the Database Group, is connected to the birth of two of the biggest web companies on the world - Google and Yahoo!. Hector Garcia-Molina and Jeff Ullman are members of the InfoLab. Recent projects include research on managing and analyzing large volumes of dynamic data (DataMotion), infrastructure and services for managing information (Digital Libraries), data stream management systems (STREAM) and crawling, storage, indexing and querying of large collections of web pages (WebBase).

The Glasgow Information Retrieval Group [62] is led by Keith van Rijsbergen and has been active in the broad area of IR. The group's interests include many areas of WebIR such as link analysis, summarization and interaction techniques. Terrier, a robust large-scale framework for building search engines, was developed by members of this group.

At the University of Washington, the Database Research Group [141] has published research on both databases and the web, including personal information management systems and web services. The UC Berkeley School of Information [31] has several projects on the field of IR, including WebIR. The Bailando Project [30] includes research on search interfaces and information visualization.

The XLBD Group [5], a Portuguese research group based at the University of Lisbon, has published and led various research projects on the subject of WebIR. XLBD's has been actively publishing in several areas related to WebIR: characterization [65, 66], ranking [113, 114] and representation [64, 63]. Tumba! (see Section 7) is the most visible result of this work. The XLBD Group is

headed by Mário J. Silva.

Other notable research groups related to IT and WebIR include AT&T [23], HP [84], IBM [92], Sun [117] and PARC [124].

7 Research Projects

In this section, the top research project in the field of WebIR are listed and briefly presented. Stanford's projects on this area include the WebBase Project [136] and the DataMotion Project [135]. The WebBase project builds upon previous Google activity, investigating various issues in crawling, storage, indexing, and querying of large collections of web pages. The goal of the DataMotion project is to build a new infrastructure for managing and analyzing large volumes of dynamic and diverse data (e.g. changes to web pages).

PageTurner [116], a large-scale study of the evolution of web pages supported by Microsoft, performed a series of large-scale web crawls that tracked the evolution of a set of 150 million web pages over the span of eleven weeks.

The Internet Archive (IA) [87] is "a non-profit [organization] that was founded to build an Internet library with the purpose of offering permanent access for researchers, historians, and scholars to historical collections that exist in digital format". The IA includes collections in text, audio, moving images, software and archived web pages. The web collection was started in 1996 and includes more than 55 billion documents. Access to this collection is available for researchers. The Chronica Project [49], used IA's web archive to created a temporal search engine.

The Spatially-Aware Information Retrieval on the Internet (SPIRIT) [134] project was funded through the European Community 5th Framework Programme, and had been a collaborative effort with six European partners. It has been engaged in the design and implementation of a search engine to find documents and datasets on the web relating to places or regions referred to in a query. The SPIRIT dataset is available for research purposes [95].

The Dynamically Evolving, Large-scale Information Systems (DELIS) [54] is an Integrated European Project founded by Sixth Framework Programm. Among the project's main goals is the development of self-regulating and self-repairing mechanisms that are decentralized, scalable, and adapt to changes in their environments (like the web).

In Portugal, the XLBD research group is responsible for Tumba!, a "search engine specially crafted to archive and provide search services to a community web formed bt those interested in subjects related to Portugal and the Portuguese people" [130]. Tumba! was launched in 2001 and is the result of applied research being developed mostly within the XLBD Group. Tumba! has the biggest archive of web pages for the `.pt` domain and uses a combination of document and link analysis to compute a global rank [51]. Versions of each crawled page are stored, using a versioned database [63, 64, 131]. Currently, a total of 57 million documents are stored (1,3 TB) and available to the public through Tomba [4]. Recent developments have been focused on exploring the geographical context of users and web pages.

There are other resources that focus on the specific subject of web search engines. These are not research oriented but are a useful source of up to date information. Search Engine Watch [93] provides "tips and information about

searching the web, analysis of the search engine industry and help to site owners trying to improve their ability to be found in search engines". Search Engine Showdown [121], maintained by Greg R. Notess, is "the users' guide to Web searching, compares and evaluates Internet search engines from the searcher's perspective".

8 Commercial Companies

In this section, the top commercial players in the field of WebIR are presented. Google [2] is a reference in the IR business, particularly in WebIR. It mission is "to organize the world's information and make it universally accessible and useful". It was founded in 1998 by Larry Page and Sergey Brin as a result of their research in link analysis on the web. Currently it has more than 5,500 employees and revenue of 6.14 billion dollars in 2005. Google Labs [3] is an active research group on the area of IR. The Google search engine uses the PageRank paradigm to rank results. Nevertheless, it well known that, in reality, "search engines use a large number of factors to rank results relative to a query, a user and the context in which it is performed" [55].

Yahoo! [145] was founded in 1994 by two Stanford PhD students has a hobby project. It soon became one of the biggest web portals and a global brand. Services offered by Yahoo! include search, communication (e.g. mail and messaging services), content (vertical portals), mobile services and advertising. Yahoo! Research has locations worldwide and focus on machine learning, search, microeconomics and media experiences. Recently, Ricardo Baeza-Yates was appointed Director of Yahoo! Research Barcelona.

Alexa [13] is an Amazon.com company that was founded in 1996. The Alexa's Toolbar has an installed user base of millions and is used to gather user data. This data is combined with Alexa's web crawls to offer services that include access to Alexa's repositories and producing web intelligence based on Alexa's massive amounts of data. Alexa has a strong connection with the Internet Archive project (see Section 7).

Microsoft [115] is a computer technology corporation founded in 1975. In late 2005, Microsoft has announced a new version of its MSN search service. The vision of the web has a software platform has contributed to Microsoft's increasing investment on this area. Microsoft has several research labs worldwide. Microsoft Research Cambridge [42] is specially focused on the field of information retrieval, with emphasis on retrieval models and optimization and learning.

9 People

In this section, top authorities in the field of WebIR and IR are briefly presented. Landmark figures in the field of Information Retrieval include Vannevar Bush, Eugene Garfield, Gerald Salton, Hans Peter Luhn, Karen Spärck Jones, Stephen Robertson and Keith van Rijsbergen.

Vannevar Bush published a seminal work [41] in 1945 envisioning the future of information access and distribution. The MEMEX (an augmenting memory device), presented in this work, is seen as a pioneering concept for hypertext

and the World Wide Web. Inspired by this work, Eugene Garfield developed a comprehensive citation index showing the propagation of scientific thinking. It was one of the founders of bibliometrics [60, 59], which later inspired link analysis exploration on the web (see Section 4.5).

Hans Peter Luhn was one of the first researchers to work on problems of information retrieval using data processing equipment. One of is major works [110] was on the development of ideas that led to the concept of selective dissemination of information (SDI).

Geral Salton was the leading authority in the field of Information Retrieval during the 70s and 80s. His work had a major impact several topics, including the vector space model, term weighting, relevance feedback and automatic text processing. He is responsible for the SMART project, an automatic text processing system, a standard upon which modern retrieval systems are based.

Karen Spärck Jones is an emeritus professor at the University of Cambridge. She co-authored several seminal papers on the fields of information science, natural language processing and information retrieval. She had a major contribution in the design and implementation of TREC (see Section 2.3), a ground for results evaluation in IR. In 1997 she edited, with Peter Willet, a reference book in the IR field - Readings in Information Retrieval [97]. Working with Stephen Robertson, she has proposed a probabilistic model [96] of information retrieval.

Keith van Rijsbergen leads the Glasgow Information Retrieval Group. His work influenced the development of probabilistic models for Information Retrieval [139]. He authored a classic book in this field, Information Retrieval [138].

Beside these historic figures, there is an increasing number of researchers specifically devoted to WebIR. Below is a selection of some of these researchers and a brief comment on their recent work.

Monika Henziger is a Research Director at Google, working in the area of WebIR and efficient algorithms and data structures [8]. Her recent research has focused on knowledge extraction from the web and link analysis [36, 37, 81, 79].

Einat Amitay, working at the IBM Research Haifa Lab, has been actively publishing on the subject of WebIR. Her recent research has focused on temporal link analysis [17], geotagging web content [20], IR evaluation [18], word sense disambiguation [21] and web document's structure analysis [19, 16]).

Hector Garcia-Molina is a professor at Standford University and a member of InfoLab. His work on the field of WebIR has focused on web crawling [46, 48] and documents change detection [129, 47].

Marc Najork and Dennis Fetterly, both at Microsoft Research, have worked on the field of WebIR, namely on web crawling, algorithms and dynamics [85, 83, 57].

Ricardo Baeza-Yates and Berthier Ribeiro-Neto co-authored a seminal book in this field - Modern Information Retrieval [24]. Ricardo Baeza-Yates is now Director of Yahoo! Research Barcelona. His current research interests are on web mining and user interfaces.

Jon Kleinberg is a professor at the Department of Computer Science at Cornell University. In 1998 he proposed the HITS algorithm [101] (see Section 4.5). In 2005 he wrote a chapter on Temporal Dynamics of On-Line Information Streams [100] included in a book on Data Stream Management from Springer.

10 Main Conferences

In this section, the top conferences that typically include specific tracks on WebIR are presented. The TREC Conference [120] series, co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense, was started in 1992 and aims "to encourage research in information retrieval from large text collections". Activities are organized in topical tracks where researchers compete for results. In 2006 there is a new Blog Track whose purpose is to explore information seeking behavior in the Blogspace. Also in the 2006 edition, the Terabyte Tracks set the goal to investigate whether/how the IR community can scale traditional IR test-collection-based evaluation to significantly large collections. Proceedings are edited and available online at TREC's web site.

The ACM SIGIR Conference focuses on research and development in information retrieval. It is the major international forum for the presentation of new research and the demonstration of new systems and techniques in the broad field of information retrieval. The 28th edition, held in Salvador, Brasil and chaired by Ricardo Baeza-Yates and Nivio Ziviani, had a paper acceptance rate of 19% [10]. This edition had a total of 23 sessions from a broad selection of topics, including several sessions on web search, summarization, multimedia and IR theory. SIGIR's proceedings are published by ACM Press.

The European Conference on Information Retrieval (ECIR), currently in its 28th edition [1], is the main European conference on the topic of Information Retrieval. This event started as a colloquium that was held between 1978 and 2001. The proceedings of ECIR are published by Springer-Verlag in their LNCS series. In 2006, a session on web search was chaired by Ricardo Baeza-Yates [107].

The annual WWW Conference, organized by the International WWW Conference Committee, has been the major event since 1994. The 15th edition was held in Edinburgh, Scotland and attracted more than 600 submissions, with an acceptance rate of 11%. A total of 28 sessions were held on a broad range of topics, including search spam, search engineering, search and new search paradigms [12]. Over the five days of the conference, several workshops, tutorials and panels were organized.

The ACM Conference on Hypertext and Hypermedia, currently in it's sixteenth edition, has regularly published papers related to WebIR. In 2005, at Hypertext'05 [125], Monika Henzinger from Google presented a keynote on hyperlink analysis [80] and several papers where presented on the subject of web structure analysis and evolution [137, 53, 86, 58]. Hypertext'04 [143] had a session on "Hypertext through time" [112, 98] and another on "Hypertext versioning" [119, 123, 142].

The ACM, through SIGIR, also sponsors several other conferences and workshops related to IR and the web, namely the ACM Workshop on Information Retrieval in Peer-to-Peer Networks [9], the ACM International Workshop on Web Information and Data Management [11], the ACM Workshop On Geographic Information Retrieval [7] and the ACM Symposium on Document Engineering [6].

11 Main Journals

Information Retrieval is a broad area of research, encompassing several topics. Although most research is published in conference proceedings (see Section 10), there are several journal where both IR and WebIR is frequently published. This section presents a brief list of these journals.

The Journal of the American Society for Information Science (JASIS), published by John Wiley & Sons, is a peer-reviewed journal that accepts papers on broad areas related to information science. Referenced papers published on this journal include [132, 104, 17].

Information Retrieval is a peer-reviewed journal published by Springer. It has a special focus on theory and experiment in information retrieval and its application in the networked information environment. The first number was published in 1999.

ACM Transactions on Internet Technologies (TOIT) is a peer-reviewed journal, published since 2001, that includes research from a broad range of topics (programming, databases, security, distributed systems, data mining). Referenced papers published on this journal include [47, 66]. ACM also publishes SIGIR Forum, an unrefereed newsletter that serves to disseminate short technical papers, book reviews and general information on the field of IR. The Communications of the ACM (ACM) is a reference publication in the field of computer science and has frequently published on the topic of WebIR and IR [133, 74].

IEEE Internet Computing is a refereed journal on the field of Internet technologies and applications published since 1997. Papers related to WebIR have been published occasionally [82].

The Information Processing & Management journal, formerly known as Information Storage and Retrieval, is a peer-reviewed journal published by Elsevier.

Internet Mathematics is a new peer-reviewed journal that began publication in 2003 and has two published volumes.

Springer's Lecture Notes in Computer Science is an important computer science series that reports on state-of-the-art research results from broad range of subjects. It is specially focused on publishing proceedings, post-proceedings and research monographs. Referenced papers published on this journal include [107].

References

[1] Ecir 2006: Home. http://ecir2006.soi.city.ac.uk/.

[2] Google. http://www.google.com/.

[3] Google labs. http://labs.google.com.

[4] Tomba - arquivo da web portuguesa. http://tomba.tumba.pt.

[5] Xlbd group. http://xldb.fc.ul.pt/.

[6] *DocEng '05: Proceedings of the 2005 ACM symposium on Document engineering*, New York, NY, USA, 2005. ACM Press.

[7] *GIR '05: Proceedings of the 2005 workshop on Geographic information retrieval*, New York, NY, USA, 2005. ACM Press.

[8] Monika henzinger, google research scientists and engineers. http://labs.google.com/people/monika/, 2005.

[9] *P2PIR'05: Proceedings of the 2005 ACM workshop on Information retrieval in peer-to-peer networks*, New York, NY, USA, 2005. ACM Press.

[10] *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2005. ACM Press.

[11] *WIDM '05: Proceedings of the 7th annual ACM international workshop on Web information and data management*, New York, NY, USA, 2005. ACM Press.

[12] *WWW '06: Proceedings of the 15th international conference on World Wide Web*, New York, NY, USA, 2006. ACM Press.

[13] Alexa. Alexa web search. http://www.alexa.com, 2006.

[14] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. Technical report, 1998.

[15] James Allan, Victor Lavrenko, and Hubert Jin. First story detection in tdt is hard. In *CIKM '00: Proceedings of the ninth international conference on Information and knowledge management*, pages 374–381, New York, NY, USA, 2000. ACM Press.

[16] Einat Amitay. *What Lays in the Layout: Using anchor-paragraph arrangements to extract descriptions of Web documents*. PhD thesis, Macquarie University, February 2001.

[17] Einat Amitay, David Carmel, Michael Herscovici, Ronny Lempel, and Aya Soffer. Trend detection through temporal link analysis. *J. Am. Soc. Inf. Sci. Technol.*, 55(14):1270–1281, December 2004.

[18] Einat Amitay, David Carmel, Ronny Lempel, and Aya Soffer. Scaling ir-system evaluation using term relevance sets. In *SIGIR '04: Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 10–17. ACM Press, 2004.

[19] Einat Amitay, Adam Darlow, David Konopnicki, and Uri Weiss. Queries as anchors: selection by association. In *HYPERTEXT '05: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, pages 193–201, New York, NY, USA, 2005. ACM Press.

[20] Einat Amitay, Nadav Har'el, Ron Sivan, and Aya Soffer. Web-a-where: geotagging web content. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 273–280, New York, NY, USA, 2004. ACM Press.

[21] Einat Amitay, Rani Nelken, Wayne Niblack, Ron Sivan, and Aya Soffer. Multi-resolution disambiguation of term occurrences. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 255–262, New York, NY, USA, 2003. ACM Press.

[22] William Y. Arms, Selcuk Aya, Pavel Dmitriev, Blazej J. Kot, Ruth Mitchell, and Lucia Walle. Building a research library for the history of the web. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 95–102, New York, NY, USA, 2006. ACM Press.

[23] Knowledge V. at&t. At&t labs research. http://public.research.att.com, 2006.

[24] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, May 1999.

[25] Ricardo A. Baeza-Yates, Felipe Saint-Jean, and Carlos Castillo. Web structure, dynamics and page quality. In *SPIRE 2002: Proceedings of the 9th International Symposium on String Processing and Information Retrieval*, pages 117–130, London, UK, 2002. Springer-Verlag.

[26] Ziv Bar-Yossef and Maxim Gurevich. Random sampling from a search engine's index. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 367–376, New York, NY, USA, 2006. ACM Press.

[27] Klaus Berberich, Srikanta Bedathur, Michalis Vazirgiannis, and Gerhard Weikum. Buzzrank ... and the trend is your friend. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, New York, USA, May 2006. University of Southampton, United Kingdom, ACM Press.

[28] Klaus Berberich, Michalis Vazirgiannis, and Gerhard Weikum. T-rank: Time-aware authority ranking. *Lecture Notes in Computer Science*, 3243:131–142, January 2004.

[29] Klaus L. Berberich. Time-aware and trend-based authority ranking. Master's thesis, Universität des Saarlandes, Germany, November 2004.

[30] University Berkeley. Bailando project homepage. http://bailando.sims.berkeley.edu, 2006.

[31] University O. Berkeley. Uc berkeley school of information — research. http://www.sims.berkeley.edu/research, 2006.

[32] Pavel Berkhin. A survey on pagerank computing. *Internet Mathematics*, 2(1):73–120, 2005.

[33] Tim Berners-Lee and Robert Cailliau. Worldwideweb: Proposal for a hypertext project. Technical report, European Laboratory for Particle Physics (CERN), November 1990.

[34] Krishna Bharat and Andrei Broder. A technique for measuring the relative size and overlap of public web search engines. *Comput. Netw. ISDN Syst.*, 30(1-7):379–388, 1998.

[35] Krishna Bharat and Andrei Broder. Mirror, mirror on the web: a study of host pairs with replicated content. In *WWW '99: Proceeding of the eighth international conference on World Wide Web*, pages 1579–1590, New York, NY, USA, 1999. Toronto, Canada, Elsevier North-Holland, Inc.

[36] Krishna Bharat, Bay W. Chang, Monika R. Henzinger, and Matthias Ruhl. Who links to whom: Mining linkage between web sites. In *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM)*, pages 51–58, Washington, DC, USA, 2001. IEEE Computer Society.

[37] Krishna Bharat and Monika R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pages 104–111, Melbourne, AU, 1998.

[38] 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, April 1998.

[39] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. In *Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications networking*, pages 309–320, Amsterdam, The Netherlands, 2000. North-Holland Publishing Co.

[40] Andrei Z. Broder, Steven C. Glassman, and Mark S. Manasse. Syntactic clustering of the web. *Computer Networks and ISDN Systems*, 29(8-13):1157–1166, 1997.

[41] Vannevar Bush. As we may think. *The Atlantic Monthly*, 176(1):101–108, 1945.

[42] Microsoft R. Cambridge. Information retrieval at msrc. http://research.microsoft.com/ir, 2006.

[43] Carlos Castillo. Effective web crawling. *SIGIR Forum*, 39(1):55–56, June 2005.

[44] Steve Chien, Cynthia Dwork, Ravi Kumar, and D. Sivakumar. Towards exploiting link evolution. In *Workshop op Algorithms and Models for the Web Graph*, 2001.

[45] Junghoo Cho. *Crawling the Web: Discovery and Maintenance of Large-Scale Web Data*. PhD thesis, Stanford University, November 2001.

[46] Junghoo Cho and Hector Garcia-Molina. The evolution of the web and implications for an incremental crawler. In *VLDB '00: Proceedings of the 26th International Conference on Very Large Data Bases*, pages 200–209, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[47] Junghoo Cho and Hector Garcia-Molina. Estimating frequency of change. *ACM Trans. Inter. Tech.*, 3(3):256–290, August 2003.

[48] Junghoo Cho, Hector Garcia-Molina, and Lawrence Page. Efficient crawling through url ordering. In *WWW7: Proceedings of the seventh international conference on World Wide Web 7*, pages 161–172, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

[49] Project Chronica. Chronica project: Internet archive temporal search. http://www.cs.usfca.edu/chronica, 2004.

[50] David Cohn and Huan Chang. Learning to probabilistically identify authoritative documents. In *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, pages 167–174, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[51] Miguel Costa and Mário J. Silva. Ranking no motor de busca tumba. In *CRC'01 - 4a Conferência de Redes de Computadores*, Covilhã, Portugal, November 2001.

[52] Michal Cutler, Yungming Shih, and Weiyi Meng. Using the structure of html documents to improve retrieval. In *Proceedings of the USENIX Symposium on Internet Technologies and Systems*, pages 241–252, 1997.

[53] Paul Davis, Alexey Maslov, and Scott Phillips. Analyzing history in hypermedia collections. In *HYPERTEXT '05: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, pages 171–173, New York, NY, USA, 2005. ACM Press.

[54] delis. Delis: Dynamically evolving, large-scale information systems. http://delis.upb.de, 2006.

[55] Nadav Eiron, Kevin S. Mccurley, and John A. Tomlin. Ranking the web frontier. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 309–318, New York, NY, USA, 2004. ACM Press.

[56] Ayman Farahat, Thomas Lofaro, Joel C. Miller, Gregory Rae, and Lesley A. Ward. Authority rankings from hits, pagerank, and salsa: Existence, uniqueness, and effect of initialization. *SIAM Journal on Scientific Computing*, 27(4):1181–1201, 2006.

[57] Dennis Fetterly, Mark Manasse, Marc Najork, and Janet L. Wiener. A large-scale study of the evolution of web pages. *Softw. Pract. Exper.*, 34(2):213–237, February 2004.

[58] Luis Francisco-Revilla and Frank Shipman. Parsing and interpreting ambiguous structures in spatial hypermedia. In *HYPERTEXT '05: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, pages 107–116, New York, NY, USA, 2005. ACM Press.

[59] E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178(60):471–479, November 1972.

[60] E. Garfield. *Citation Indexing.* ISI Press, 1979.

[61] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In *SOSP '03: Proceedings of the nineteenth ACM symposium on Operating systems principles*, pages 29–43, New York, NY, USA, 2003. ACM Press.

[62] University O. Glasgow. Glasgow information retrieval group. http://ir.dcs.gla.ac.uk, 2006.

[63] Daniel Gomes, João P. Campos, and Mário J. Silva. Versus: A web repository. In *WDAS - Workshop on Distributed Data and Structures 2002*, Paris, France, March 2002.

[64] Daniel Gomes, André L. Santos, and Mário J. Silva. Managing duplicated in a web archive. In *21st Annual ACM Symposium on Applied Computing*, Bourgogne University, Dijon, France, April 2006. ACM Press.

[65] Daniel Gomes and Mário J. Silva. A characterization of the portuguese web. In *Proceedings of 3rd ECDL Workshop on Web Archives*, Trondheim, Norway, August 2003.

[66] Daniel Gomes and Mário J. Silva. Characterizing a national community web. *ACM Transactions on Internet Technologies*, 5(3):508–531, August 2005.

[67] Daniel Gomes and Mário J. Silva. Modelling information persistence on the web. In *6th International Conference on Web Engineering*, New York, NY, USA, July 2006. ACM Press.

[68] Google. Google code: Web authoring statistics. http://code.google.com/webstats/index.html.

[69] Google. Google technology. http://www.google.com/technology/ [Visited 2007/03/16].

[70] Michael Gordon and Praveen Pathak. Finding information on the world wide web: the retrieval effectiveness of search engines. *Information Processing & Management*, 35(2):141–180, March 1999.

[71] Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 491–501, New York, NY, USA, May 2004. ACM Press.

[72] Venkat N. Gudivada, Vijay V. Raghavan, William I. Grosky, and Rajesh Kasanagottu. Information retrieval on the world wide web. *IEEE Internet Computing*, 1(5):58–68, September 1997.

[73] A. Gulli and A. Signorini. The indexable web is more than 11.5 billion pages. In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 902–903, New York, NY, USA, 2005. ACM Press.

[74] Amarnath Gupta and Ramesh Jain. Visual information retrieval. *Commun. ACM*, 40(5):70–79, May 1997.

[75] Cathal Gurrin and Alan F. Smeaton. Replicating web structure in small-scale test collections. *Inf. Retr.*, 7(3-4):239–263, 2004.

[76] Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. Combating web spam with trustrank. In *30th International Conference on Very Large Data Bases*, pages 576–587, August 2004.

[77] Siegfried Handschuh, Steffen Staab, and Raphael Volz. On deep annotation. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 431–438, New York, NY, USA, 2003. ACM Press.

[78] David Hawking, Nick Craswell, Peter Bailey, and Kathleen Griffihs. Measuring search engine quality. *Inf. Retr.*, 4(1):33–59, April 2001.

[79] Monika Henzinger. Link analysis in web information retrieval. 2000.

[80] 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 Press.

[81] Monika R. Henzinger. Web information retrieval - an algorithmic perspective. In *European Symposium on Algorithms*, pages 1–8, 2000.

[82] Monika R. Henzinger. Hyperlink analysis for the web. *IEEE Internet Computing*, 5(1):45–50, 2001.

[83] Monika R. Henzinger, Allan Heydon, Michael Mitzenmacher, and Marc Najork. Measuring index quality using random walks on the web. *Comput. Networks*, 31(11-16):1291–1303, 1999.

[84] Development C. hewlett-Packard. Hp labs - advanced research at hp. http://www.hpl.hp.com, 2006.

[85] Allan Heydon and Marc Najork. Mercator: A scalable, extensible web crawler. *World Wide Web*, 2(4):219–229, 1999.

[86] Ikumi Horie, Kazunori Yamaguchi, and Kenji Kashiwabara. Higher-order rank analysis for web structure. In *HYPERTEXT '05: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, pages 98–106, New York, NY, USA, 2005. ACM Press.

[87] ia. Internet archive. http://www.archive.org, 2006.

[88] ia. Internet archive frequently asked questions. http://www.archive.org/about/faqs.php, 2006.

[89] ia. Internet archive: Petabox. http://www.petabox.org, 2006.

[90] ia. Internet archive: The wayback machine - hardware. http://www.archive.org/web/hardware.php, 2006.

[91] Iac. Ask.com. http://www.ask.com/.

[92] ibm. Ibm research. http://www.research.ibm.com, 2006.

[93] iim. Search engine watch: Tips about internet search & search engine submission. http://searchenginewatch.com, 2006.

[94] Infolab. The stanford university infolab. http://i.stanford.edu, 2006.

[95] Hideo Joho and Mark Sanderson. The spirit collection: an overview of a large web collection. *SIGIR Forum*, 38(2):57–61, December 2004.

[96] Karen S. Jones, Steve Walker, and Stephen E. Robertson. A probabilistic model of information retrieval: Development and status. Technical report, Cambridge University Computer Laboratory, 1998.

[97] Karen S. Jones and Peter Willett. *Readings in information retrieval*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[98] Madhur Khandelwal, Andruid Kerne, and Michael J. Mistrot. Manipulating history in generative hypermedia. In *HYPERTEXT '04: Proceedings of the fifteenth ACM conference on Hypertext and hypermedia*, pages 139–140, New York, NY, USA, 2004. ACM Press.

[99] Jon Kleinberg. Bursty and hierarchical structure in streams. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 91–101, New York, NY, USA, 2002. ACM Press.

[100] Jon Kleinberg. *Temporal Dynamics of On-Line Information Streams*. Springer-Verlag, New York, USA, 2006.

[101] Jon M. K. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[102] Mei Kobayashi and Koichi Takeda. Information retrieval on the web. *ACM Comput. Surv.*, 32(2):144–173, June 2000.

[103] Mei Kobayashi and Koichi Takeda. Information retrieval on the web. *ACM Comput. Surv.*, 32(2):144–173, June 2000.

[104] Wallace Koehler. Web page change and persistence—a four-year longitudinal study. *J. Am. Soc. Inf. Sci. Technol.*, 53(2):162–171, January 2002.

[105] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. On the bursty evolution of blogspace. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 568–576, New York, USA, 2003. ACM Press.

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

[107] Mounia Lalmas, Andy Macfarlane, Stefan M. Rüger, Anastasios Tombros, Theodora Tsikrika, and Alexei Yavlinsky, editors. *Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12, 2006, Proceedings*, volume 3936 of *Lecture Notes in Computer Science*. Springer-Verlag, 2006.

[108] Steve Lawrence and Lee C. Giles. Accessibility and distribution of information on the web. *Nature*, 400(6740):107–109, July 1999.

[109] R. Lempel and S. Moran. Salsa: the stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst.*, 19(2):131–160, April 2001.

[110] Hans P. Luhn. A business intelligence system. *IBM Journal of Research and Development*, 2(2):159–165, 1958.

[111] John Markwell and David W. Brooks. "link rot" limits the usefulness of web-based educational materials in biochemistry and molecular biology. *Biochemistry and Molecular Biology Education*, 31(1):69–72, January 2003.

[112] Catherine C. Marshall and Gene Golovchinsky. Saving private hypertext: requirements and pragmatic dimensions for preservation. In *HYPERTEXT '04: Proceedings of the fifteenth ACM conference on Hypertext & hypermedia*, pages 130–138. ACM Press, 2004.

[113] Bruno Martins and Mário J. Silva. A graph-ranking algorithm for geo-referencing documents. In *Proceedings of ICDM-05, the 5th IEEE International Conference on Data Mining*, November 2005.

[114] Bruno Martins, Mário J. Silva, and Leonardo Andrade. Indexing and ranking in geo-ir systems. In *Proceedings of the Workshop on Geographic Information Retrieval at CIKM 2005*, October 2005.

[115] Corporation Microsoft. Microsoft corporation. http://www.microsoft.com/, 2006.

[116] Corporation Microsoft. Pageturner: A large-scale study of the evolution of web pages. http://research.microsoft.com/research/sv/PageTurner, 2006.

[117] Sun Microsystems. Sun microsystems laboratories. http://research.sun.com, 1994.

[118] Shinsuke Nakajima, Junichi Tatemura, Yoichiro Hino, Yoshinori Hara, and Katsumi Tanaka. Discovering important bloggers based on analysing blog threads. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, New York, USA, May 2005. ACM Press.

[119] Tien N. Nguyen, Ethan V. Munson, and John T. Boyland. The molhado hypertext versioning system. In *HYPERTEXT '04: Proceedings of the fifteenth ACM conference on Hypertext and hypermedia*, pages 185–194, New York, NY, USA, 2004. ACM Press.

[120] nist. Text retrieval conference (trec) home page. http://trec.nist.gov, 2006.

[121] Greg R. Notess. Search engine showdown: The users' guide to web searching. http://searchengineshowdown.com, 2006.

[122] Alexandros Ntoulas, Junghoo Cho, and Christopher Olston. What's new on the web?: the evolution of the web from a search engine perspective. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 1–12, New York, NY, USA, 2004. ACM Press.

[123] Kai Pan, James E. Whitehead, and Guozheng Ge. Hypertext versioning for embedded link models. In *HYPERTEXT '04: Proceedings of the fifteenth ACM conference on Hypertext and hypermedia*, pages 195–204, New York, NY, USA, 2004. ACM Press.

[124] parc. Palo alto research center. http://www.parc.xerox.com, 2006.

[125] Siegfried Reich and Manolis Tzagarakis, editors. *HYPERTEXT '05: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, New York, NY, USA, 2005. ACM Press.

[126] M. Sahami, V. Mittal, S. Baluja, and H. Rowley. The happy searcher: Challenges in web information retrieval. In Chengqi Zhang, Hans W. Guesgen, and Wai K. Yeap, editors, *PRICAI 2004: Trends in Artificial Intelligence: 8th Pacific Rim International Conference on Artificial Intelligence*, volume 3157, pages 3–12, 2004.

[127] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November 1975.

[128] Gerard Salton and Chris Buckley. Term weighting approaches in automatic text retrieval. Technical report, Ithaca, NY, USA, 1987.

[129] Narayanan Shivakumar and Hector Garcia-Molina. Finding near-replicas of documents on the web. In *WebDB'98: Selected papers from the International Workshop on The World Wide Web and Databases*, pages 204–212, London, UK, 1999. Springer-Verlag.

[130] Mário F. Silva. The case for a portuguese web search engine. In *IADIS International Conference WWW Internet 2003*, November 2003.

[131] Mário J. Silva. Searching and archiving the web with tumba! In *CAPSI 2003 - 4a Conferência da Associação Portuguesa de Sistemas de Informação*, November 2003.

[132] H. Small. Co-citation in scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science (JASIS)*, 24(4):265–269, 1973.

[133] Diomidis Spinellis. The decay and failures of web references. *Commun. ACM*, 46(1):71–77, January 2003.

[134] spirit. Spirit - spatially-aware information retrieval on the internet. http://www.geo-spirit.org/, 2006.

[135] University O. Stanford. Stanford datamotion. http://www-db.stanford.edu/datamotion, 2006.

[136] University O. Stanford. Webbase project. http://dbpubs.stanford.edu:8091/ testbed/doc2/WebBase, 2006.

[137] Masashi Toyoda and Masaru Kitsuregawa. A system for visualizing and analyzing the evolution of the web with a time series of graphs. In *HYPERTEXT '05: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, pages 151–160, New York, NY, USA, 2005. ACM Press.

[138] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, UK, 1979.

[139] C. J. van Rijsbergen. A new theoretical framework for information retrieval. *SIGIR Forum*, 21(1-2):23–29, 1987.

[140] Ellen M. Voorhees. The philosophy of information retrieval evaluation. In *CLEF '01: Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370, London, UK, 2002. Springer-Verlag.

[141] University O. Washington. Uw cse database group. http://data.cs.washington.edu/, 2006.

[142] James E. Whitehead, Guozheng Ge, and Kai Pan. Automatic generation of hypertext system repositories: a model driven approach. In *HYPERTEXT '04: Proceedings of the fifteenth ACM conference on Hypertext and hypermedia*, pages 205–214, New York, NY, USA, 2004. ACM Press.

[143] Jim Whitehead and David De Roure, editors. *HYPERTEXT '04: Proceedings of the fifteenth ACM conference on Hypertext and hypermedia*, New York, NY, USA, 2004. ACM Press.

[144] M. Wu and D. H. Sonnenwald. Reflections on information retrieval evaluation. In *Proceedings of the 1999 EBTI, ECAI, SEER & PNC Joint Meeting*. Academia Sinica, Taipei, Taiwan, 1999.

[145] Inc Yahoo! Yahoo! http://www.yahoo.com/, 2006.