Revista de Estudos Politécnicos
Polytechnical Studies Review
2008, Vol VI, nº 9
ISSN: 1645-9911

Satisfying Information Needs on the Web: a Survey of Web Information Retrieval*

Nuno Filipe Escudeiro[1•], Alípio Mário Jorge[2•]
nfe@isep.ipp.pt, amjorge@fep.up.pt

(recebido em 20 de Março de 2008; aceite em 22 de Abril de 2008)

Resumo. Desde muito cedo que a espécie Humana sentiu a necessidade de manter registos da sua actividade, para que possam ser facilmente consultados futuramente. A nossa própria evolução depende, em larga medida, deste processo iterativo em que cada iteração se baseia nestes registos. O aparecimento da web e o seu sucesso incrementaram significativamente a disponibilidade da informação que rapidamente se tornou ubíqua. No entanto, a ausência de controlo editorial origina uma grande heterogeneidade sob vários aspectos. As técnicas tradicionais em recuperação de informação provam ser insuficientes para este novo meio. A recuperação de informação na web é a evolução natural da área de recuperação de informação para o meio web. Neste artigo apresentamos uma análise retrospectiva e, esperamos, abrangente desta área do conhecimento Humano.

Palavras-chave: Recuperação de informação na web, motores de pesquisa.

Abstract. Human kind felt, since early ages, the need to keep records of its achievements that could persist through time and that could be easily retrieved for later reference. Our own evolution depends largely on this iterative process, where each iteration is based on these records. The advent of the web and its

* Supported by the POSC/EIA/58367/2004/Site-o-Matic Project (Fundação Ciência e Tecnologia), FEDER e Programa de Financiamento Plurianual de Unidades de I & D.

1 DEI-ISEP – Deptº de Engenharia Informática, Instituto Superior de Engenharia do Porto ; http://www.dei.isep.ipp.pt

2 2FEP-UP – Faculdade de Economia, Universidade do Porto; http://www.fep.up.pt
• LIAAD, INESC Porto LA – Laboratório de Inteligência Artificial e Análise de Dados; http://www.liaad.up.pt

attractiveness highly increased the availability of information which rapidly becomes ubiquituous. However, the lack of editorial control originates high heterogeneity in several ways. The traditional information retrieval techniques face new, challenging problems and prove to be inefficient to deal with web characteristics. In this paper we present a comprehensive and retrospective overview of web information retrieval.

Keywords: Web information retrieval, search engines.

1. Introduction

The World Wide Web or simply the *web* may be seen as a huge collection of documents freely produced and published by a very large number of people, without any solid editorial control. This is probably the most democratic – and anarchic –widespread mean for anyone to express feelings, comments, convictions and ideas, independently of ethnics, sex, religion or any other characteristic of human societies. The web constitutes a comprehensive, dynamic, permanently up-to-date repository of information regarding most of the areas of human knowledge (Hu, 2002) and supporting an increasingly important part of commercial, artistic, scientific and personal transactions, which gives rise to a very strong interest from individuals, as well as from institutions, at a universal scale.

However, the web also exhibits some characteristics that are adverse to the process of collecting information from it in order to satisfy specific needs: the large volume of data it contains; its dynamic nature; being mainly constituted by unstructured or semi-structured data; content and format heterogeneity and irregular data quality are some of these adverse characteristics.End-users also introduce some additional difficulties in the retrieval process: information needs are often imprecisely defined, generating a semantic gap between user needs and their specification.

The satisfaction of a specific information need on the web is supported by search engines and other tools aimed at helping users gather information from the web. The user is usually not assisted in the subsequent tasks of organizing, analyzing and exploring the answers produced. These answers are usually flat lists of large sets of web pages which demand significant user effort to be explored. Satisfying information needs on the web is usually seen as an ephemeral one-step process of

information search (the traditional search engine paradigm). Given these characteristics, it is highly demanding to satisfy private or institutional information needs on the web.

The web itself, and the interests it promotes, are growing and changing rapidly, at a global scale, both as mean of divulgation and dissemination and also as a source of generic and specialized information. Web users have already realized the potential of this huge information source and use it for many purposes, mainly in order to satisfy specific information needs. Simultaneously the web provides a ubiquitous environment for executing many activities, regardless of place and time.

This paper presents the evolution of information retrieval with particular emphasis on the web environment. The rest of the paper is organied as follows: section 2 gives a retrospective view of web information retrieval referring to the most relevant milestones; section 3 defines and describes the web and discusses the distinction between traditional information retrieval and web information retrieval; section 4 presents related areas and base theory for web information retrieval; section 5 discusses research trends in the area and section 6 concludes the paper.

2. A Retrospective View of Web Information Retrieval

Although the need to collect, organize and explore information resources appears very early in human societies, the advent of the web and digital libraries, together with advances in communications and computer technology, given the current challenges, raises the interest on the Information Retrieval field (IR) from scientific research as well as from commercial and public domains.

Human memory and cognitive capabilities are not able of preserving all knowledge that human kind produces; we need to store substantial amounts of information representing this knowledge. These characteristics led human kind to store and organize information records in a way that facilitates later retrieval and usage for over 4000 years (Baeza-Yates et al., 1999). Human culture is preserved through these records.

The common way of referring to information records and allowing for faster access is through a set of predefined concepts with which are associated related records. This structure is called the *index*. Since early IR, indexes represented taxonomies, hierarchical structures of related concepts. The Dewey Decimal System is a system of library classification developed in 1876 that has been greatly used and expanded (Dewey, 2004). These taxonomies, created manually as document indexing, are reasonable for small document collections but are not scalable.

In 1945, Vannevar Bush claims that scientific efforts should shift from increasing physical abilities, which had been the main focus during World War II, to making all previously collected human knowledge more accessible (Bush, 1945). To help this task he envisions and describes Memex, a system that let users organize information and relate documents to each other through a mesh of associative trails running through them. Memex description suggests many technological features that we use today (hypertext, personal computers, the Internet, the web, online encyclopaedias such as Wikipedia) and constitutes a main milestone in the IR area.

The end of World War II released a set of restrictions on scientific and technical information and generated pressure to make it available to the community. Several efforts were made to improve the conventional types of index (card catalogs) and indexing techniques (taxonomies and alphabetical subject headings). In 1950 Mortimer Taube, a government librarian in the USA, realized that a topic list with 40000 subject headings was composed of only 7000 different words. He proposed Uniterm (Cleverdon et al., 1954), a system that uses these words as index terms and combines them at search time performing slightly better than the conventional system, with great reductions of indexing effort (Cleverdon, 1991).

In 1954 the term IR is popularized in a technical report by Cleverdon (Cleverdon et al., 1954).

During the 60s, IR experienced a period of great evolution, motivated for several reasons. In 1960, the ex-USSR launches Sputnik I, the first earth artificial satellite, which generated particular efforts, from the USA, to improve scientific information spreading; we were living in the Cold War and both opponents struggle for leadership, especially in the scientific domain. At the same time, the development of computerized database technology (Codd, 1970) creates conditions for more efficient IR systems contributing to evolutions in the field. However, although some

developments arise on the automatic categorization of documents (Lim et al., 2001), indexing still remains mainly manual and only the search becomes mechanical.

It is also during this decade that Cleverdon develops the mathematics for *recall* (fraction of relevant documents that are retrieved) and *precision* (fraction of retrieved documents that are relevant) and greatly develops the area of IR systems' evaluation, a main problem in the field. The Cranfield tests apply this technology to evaluate IR systems on the first test collections compiled with evaluation purpose (Cleverdon, 1962; Cleverdon et al., 1963).

The *hypertext* term is coined in 1965 by Ted Nelson who develops Xanadu, a hypertext system (Nelson, 1965).

In 1968, evidence arises that the value of manual indexing is low and that free-text indexing may be as effective and much cheaper. In 1971, Jardine and Rijsbergen establish the *cluster hypothesis* (Jardine et al., 1971) which states that if a document satisfies a particular information need, similar documents will probably also satisfy it. This hypothesis has recently been confirmed for distributed IR (Crestani et al., 2006).

The vector space model for representing text documents, still the most common model in IR, is proposed by Salton in 1975 (Salton et al., 1975). Salton is responsible for the SMART project (Salton et al., 1965), an automatic text processing system, a standard upon which modern retrieval systems are based.

Three years later, in 1978, takes place the first ACM SIGIR conference.

Rijsbergen published *Information Retrieval* (Rijsbergen, 1979), a reference book in IR, with a heavy emphasis on probabilistic models, in 1979. In 1983, Salton publishes another reference book, *Introduction to Modern Information Retrieval* (Salton et al., 1983), focused on vector space models.

During the 70s, IR was viewed as finding the right information in text databases.

In 1989 Tim Berners-Lee proposes a system to manage general information on research projects at CERN (Berners-Lee, 1989) which became the genesis of the

World Wide Web. The evolution of the web had great impact on our daily lives and also generated new challenges and a renewed pressure on IR research.

In the early 90s IR moves to full text. Document indexes are built from full content rather than just abstracts and titles. This decade has been very proficient to IR.

In 1992 the first TREC conference takes place; during this same year the HTML 1.0 formal definition is established. The first web browser, Mosaic which later will became Netscape, came alive during 1993.

In 1994, MIT and CERN agreed to set up the World Wide Web Consortium (W3C), an institution aimed at developing and setting standards for the web and web technologies. This same year, the first full text web search engine, named WebCrawler, created by Brian Pinkerton at the University of Washington, goes online and is immediately followed by Lycos. It is the dawning of the first generation of web search engines that explore mainly on-page data. In 1995 AltaVista comes alive and Yahoo! is incorporated as a company.

With the increase in the number of web users and contents it is realized that the web is spontaneously becoming a large repository of mankind knowledge and culture. The *Internet Archive* project intents to keep records of web content along time to prevent all this knowledge from being lost (Kahle, 1997). In 2006 this project had already archived around 2 PB of information, increasing at a rate of 20 TB per month.

In 1998 two seminal papers, describing the algorithms PageRank (Brin et al., 1998) and HITS (Kleinberg, 1998), are at the genesis of the second generation of web search engines. Both these algorithms explore web structure and determine page quality by analyzing hyperlinks between pages.

The web structure along with web usage patterns are regarded as sources of evidence that might improve web IR. These are new characteristics that were not present at traditional IR. In 2000, Broder proposes a model for the web topology (Broder et al., 2000). In this model, inferred from the mesh of hyperlinks between web pages, the web is viewed as composed of a central core, made of a set of strongly interconnected pages; a set of pages with links leading to the central core,

another set of pages pointed to by pages in the central core and a fourth set composed by relatively isolated pages, that are attached to one of the previous three. Web page distribution among these four subsets is approximately uniform.

The web lacks editorial control which is one of its benefits but also a drawback. It is hard to develop systems to automatically process web content due to its heterogeneity and lack of structure. On the other hand, it is necessary to process web content automatically, due to its volume and dynamics. In 2001, Berners-Lee, proposes the *semantic web* (Berners-Lee et al., 2001). The semantic web is an attempt to develop a framework that facilitates the automatic processing of web information.

Lately, web IR systems try to explore multiple sources of evidence aiming to answer the specific user-need behind the query. Web search engines are now at their third generation, attempting to merge evidence from various sources, such as: phrase-based indexing (Hammouda et al., 2004), linguistic approaches (Apostolico et al., 2006), context (Crestani et al., 2007; Haveliwala, 2005; Ifrim et al., 2005), temporal analysis (Beitzel, 2006) and semantic models (Siddiqui, 2006).

The evolution of IR systems may be organized in four distinct periods, with significant differences among the methods that were applied and the sources used during each one. During an initial period, up to the 50s, the indexing and searching processes were handled manually. Indexes were based on taxonomies or alphabetical lists of previously specified concepts. During this phase, IR systems were mainly used by librarians and scientists.

During a second period, between around 1950 and the advent of web in the early 90s, the pressure on the field and the evolution on computer and database technology allowed for significant improvements. We went from manual to automated annotation of documents; however indexes were still built from restricted descriptions of documents (mainly abstracts and document titles). IR was viewed as finding the right information in text databases. Operating IR systems frequently required specific learning. IR systems utilization was expensive and available only to restricted groups.

During a third period, covering the 90s, the process of indexing and searching becomes fully automated. Full text indexes are built; web mining evolves and explores not only content but also structure and usage. IR systems become unrestricted, cheap, widely available and widely used.

From around 2000 on, the fourth and actual period, other sources of evidence are explored trying to improve systems' performance.

Searching and browsing are the two basic IR paradigms on the web (Baeza-Yates et al., 1999). Three approaches to IR seem to have emerged (Broder et al., 2005):

- the *search-centric approach* argues that free search has become so good and the search user-interface so common, that users can satisfy all their needs through simple queries. Search engines follow this approach;
- the *taxonomy navigation approach* claims that users have difficulties expressing their information needs; organizing information on a hierarchical structure might help finding relevant information. Directory search systems follow this approach;
- the *meta-data centric approach* advocates the use of meta-data for narrowing large sets of results (multi faceted search); third generation search engines are trying to improve the quality of their answers by merging several sources of evidence.

IR systems also have to solve problems related to their sources and how to build their databases/indexes. Several crawling algorithms have been explored, in order to overcome problems of scale arising from web dimension, such as *focused crawling* (Chakrabarti et al., 1999b), *intelligent crawling* (Aggarwal et al., 2001) and *collaborative crawling* (Aggarwal et al., 2004) that explores user behaviour registered in server logs.

Other approaches have also been proposed: *meta-search* explores the small overlap among search engines' indexes sending the same query to a set of search engines and merging their answers – a few specific problems arise from this approximation (Wang et al., 2003); *dynamic search engines* try do deal with web dynamics, such search engines do not have any permanent index but instead crawl for their answers at query time (Hersovici et al., 1998); *interactive search* (Bruza et al., 2000) wraps

a general purpose search engine into an interface that allows users to navigate towards their goal through a query-by-navigation process.

At present, IR research seems to be focused on retrieval of high quality, integration of several sources of evidence and multimedia retrieval.

3. Information Retrieval and Web Information Retrieval

Classical IR has many applications on the web. However, conventional techniques must be adapted to the specific characteristics of the web that raise new problems. Web search engines are the most widely used systems in the web. Recent surveys (http://searchenginewatch.com, accessed on 2007) estimate that over 75% of web users explore the web through search services and spend more than 70% of their time searching online, generating millions of requests daily; Google receives around 2×10^9 queries a day. Theses systems are permanently seeking for innovative ways to achieve the main IR goal: maximize user satisfaction.

3.1 Information Retrieval

IR concerns the study of systems and techniques for representing, organizing and searching information items. Information items are regarded as unstructured or semi-structured objects which lead to significant differences from data retrieval (that concerns structured data) and rises new problems. The IR field is strongly related to information science, library science and computer science.

3.2 The Web

The web is a public service constituted by a set of applications aimed at extracting documents from computers accessible in Internet – the Internet is a network of computer networks. We may also describe the web as an information repository distributed over millions of computers interconnected through Internet (Baldi et al., 2003). The W3C defines web in a broad way: “the World Wide Web is the universe of network-accessible information, an embodiment of human knowledge”.

Due to its comprehensiveness, with contents related to most subjects of human activity, and global public acceptance, either at a personal or institutional level, the web is widely explored as an information source. Web dimension and dynamic nature become serious drawbacks when it comes to retrieving information. Another relevant characteristic of the web is the absence of any global editorial control over its content and format. This contributes largely to web success but also contributes to a high degree of heterogeneity in content, language, structure, correctness and validity.

Although the problems raised by the size of the web, around $11{,}5\times10^9$ pages (Gulli et al., 2005), and its dynamics require special treatment, it seems that the major difficulties concerning the processing of web documents are generated by the lack of editorial rules and the lack of a common ontology, which would allow for unambiguous document specification and interpretation. In the absence of such normative rules, each document has to be treated as unique. In this scenario, document processing cannot be based on any underlying structure. Although HTML already involves some structure its use is not mandatory. Therefore, the higher level of abstraction that may assure compatibility with a generic web document is the common *bag-of-words* (Chakrabarti, 2003). This low abstraction level is not very helpful for automatic processing, requiring significant computational costs.

The web is a vast and popular repository, containing information related to almost all human activities and being used to perform an ever growing set of distinct activities (bank transactions, shopping, chatting, government transactions, weather report and getting geographic directions, just to name a few). Despite the difficulties this medium poses to automatic as well as to non-automatic processing, it has been increasingly explored and has been motivating efforts, from both academic and industry, which aim to facilitate this exploration.

Currently the web is a repository of documents, the majority of them HTML documents, that can be automatically presented to users but that do not have a base model that might be used by computers to acquire semantic information on the objects being manipulated. The *semantic web* is a formal attempt from W3C to transform the web in a huge database that might be easier to process automatically than our current *syntactic web*. However, despite many initiatives on the semantic

web (Lu et al., 2002), the web has its own dynamics and *web citizens* are pushing the web to the social plan. Collaborative systems, radical trust and participation are the main characteristics of web2.0, a new paradigm emerging since 2004 (O'Reily, 2004).

3.3 Web Information Retrieval

Web IR is the application of IR to the web. In classical IR, users specify queries, in some query language, representing their information needs. The system selects the set of documents in its collection that seem the most relevant to the query and presents them to the user. Users may then refine their queries to improve the answer. In the web environment user intents are not static and stable as they usually are in traditional IR. In the web, the information need is associated with a given task (Broder, 2002) that is not known in advance and may be quite different from user to user, even if the query specification is the same. The identification of this task and the mental process of deriving a query from an information need are crucial aspects in web IR.

Web IR is related to *web mining* – the automatic discovery of interesting and valuable information from the web (Chakrabarti, 2003). It is generally accepted that web mining is currently being developed towards three main research directions, related to the type of data they mine: web content mining, web structure mining and web usage mining (Kosala et al., 2000). Recently another type of data – document change, page age and information recency – is generating research interests: it is related to a temporal dimension and allows for analyzing the growth and dynamics – over time – of the Web (Baeza-Yates, 2003; Cho et al., 2000; Lim et al., 2001). This categorization is merely conceptual, these areas are not mutually exclusive and some techniques dedicated to one may use data that is typically associated with others.

Web content mining concerns the discovery of useful information from web page content which is available in many different formats (Baeza-Yates, 2003) – textual, metadata, links, multimedia objects, hidden and dynamic pages and semantic data.

Web structure mining tries to infer knowledge from the link structure on the web (Chakrabarti et al., 1999a). Web documents typically point at related documents

through a link forming a social network. This network can be represented by a directed graph where nodes represent documents and arcs represent the links between them. The analysis of this graph is the main goal of web structure mining (Donato et al., 2000; Kumar et al., 2000). In this field, two algorithms, which rank web pages according to their relevance, have received special attention: PageRank (Brin et al., 1998) and Hyperlink Induced Topic Search, or HITS (Kleinberg, 1998).

Web usage mining tries to explore user behavior on the web by analyzing data originated from user interaction and automatically recorded in web server logs. The applications of web usage mining usually intend to learn user profiles or navigation patterns. Web usage mining is essentially aimed at predicting the next user request based on the analysis of previous requests. Markov models are very common in modeling user requests or user paths within a site (Borges, 2000). Association rules and other standard data mining and OLAP techniques are also explored. (Cooley et al., 1997) presents an overview of the most relevant work in web usage mining.

Web IR is the evolution of traditional IR applied to the web; however, web characteristics raise new challenges that are not present in the traditional IR setting.

In web IR, for instance, one has to pay attention to spammers (Chakrabarti, 2003), people motivated by economic reasons and other, that deliberately add popular query terms to documents unrelated to those terms (text can be hidden from user if font and background have the same color, the so called *font color spamming*); web documents may be replicated, generating near or exact copies, and complete web sites may be mirrored in different addresses; web content validity, correctness and quality are highly variable and difficult to measure; the subject of web documents is also difficult to determine and the labeling process is expensive and time consuming. None of these problems arise in traditional IR.

In typical data mining problems, the examples are described in a tabular form where columns represent the properties, characteristics or features that describe the concept and each line corresponds to an individual instance or example. Features are unambiguous; their semantic is precisely known and every single example is described by exactly the same feature set. In a typical problem from these fields, the number of features is usually some orders of magnitude smaller than the number of examples. Learning tasks in web mining do not share many of these desirable

characteristics: the feature space dimension is usually much larger than the number of available examples (documents), each particular document is described by a subset of all features and the intersection of these subsets is of reduced dimension, resulting in a sparse document/term matrix, which creates additional difficulties.

One relevant drawback of a web query, when compared to a database query, which is due to the lack of structure in web documents, is that it is not usually possible to answer a web query without some degree of uncertainty (Baeza-Yates, 1999). While in the database field a query is unambiguous and its mapping to the examples is precise, in the web environment such precision is not possible since the query specification – a set of keywords – is usually ambiguous and the relevance of a certain document to the user information need, specified through that query, is, therefore, subject to ambiguity and uncertainty. Besides there is a semantic gap between the real need the user wants to fulfill and the keywords he uses to express it.

The different aspects between IR and web IR are grouped in four categories by (Lewandowski, 2005) related to: the documents themselves (including language, document length, spam, hyperlinks); the user behaviour; the web characteristics (such as volume, coverage and duplicates) and the search system in itself (user interface, ranking, search functions).

3.4 Web Search

Web search engines are designed to retrieve the web pages that may answer user queries. Originally it was implicitly assumed that user intent is to satisfy information needs; however, recently this assumption was questioned (Broder, 2002). User intention might be *informational* – the aim is to obtain information on a given topic – which arises in 40 to 50% of queries, *transactional* – the aim is to perform a given transaction, such as shopping or downloading music records – arising in 30 to 35% of queries, or *navigational* – the aim is to reach a given site – arising in 20 to 25% of web queries. The identification of user intent for a given query may help improving the relevance of the answer and IR systems quality.

As the web evolves, both in volume, heterogeneity and in the set of activities that users perform on it, search engines utilization increases and the sources of evidence

they explore also evolve. We may classify web search engines according to the set of features they explore (Broder, 2002). First generation web search engines, starting in 1994 with WebCrawler and Lycos, explore on-page data (content and formatting). They support mostly informational queries. The second generation, emerging in 1998, with Google (PageRank), uses off-page web specific data (link analysis, anchor text and click streams data) and supports both informational and navigational queries. The third generation, appearing during the first years of 2000 attempts to merge multiple sources of evidence and aims to support all kinds of queries.

4. Related Areas and Theory

IR is a multidisciplinary field exploring several scientific areas. Among them we find *library and information sciences* – concerned with the collection, classification, manipulation, storage, preservation, retrieval and dissemination of information – since early days.

In the late 60s the *database technology* improved greatly and, from there on has a major contribution to the field.

IR requires processing non-structured textual data; thus *linguistics*, *natural language processing* and *artificial intelligence* also contribute greatly to the field.

The advent of the web turns information into a ubiquitous and heterogeneous resource and increases the volume of publicly available information to huge dimensions. In this setting the *statistical* and *machine learning* fields become fundamental to IR.

The web itself and the way people build and use its content are social phenomena in nature, and exploring it may benefit from *social network analysis*, which is another relevant area contributing to the IR field.

Merging contributions from all these and many other scientific areas generates solutions for many problems that arise in the IR field, such as: the categorization and clustering of documents, relevance evaluation, cross-language retrieval,

distributed IR, question answering, spam detection, collaborative filtering, content management and adaptive web sites, recommendation, results presentation.

Web IR requires the automation of web page processing; otherwise it is not scalable for the web dimension and dynamics.

For the sake of IR we may group web page processing issues in three stages:

- *Pre-processing*: converting a web page into a manageable representation; includes a data preparation phase (removes irrelevant data and adds metadata) and a data representation phase (maps these data structures to adequate models);
- *Learning*: inferring knowledge from the available data with the aim of finding and organizing web pages in a way that satisfies user needs;
- *Presentation*: presenting relevant pages to the user in a way that allows for an efficient exploration. Although this is not usually a concern of web search engines, it seems that using visual tools to organize and present large document collections might improve user satisfaction.

4.1 Pre-processing

The pre-processing stage concerns the tasks that are required to obtain a suitable web document representation, valid for the subsequent automatic learning phase. This stage includes data preparation and data representation.

The *data preparation phase* includes several steps that attempt to eliminate non-informative features. This phase usually includes (Baeza-Yates et al., 1999):

lexical analysis – eliminating punctuation, accents, extra spacing, converting to lower/upper case;

stop-word removal – removing irrelevant terms; requires a list of words to eliminate (stop-words);

stemming – reducing words to their semantic root; Porter algorithm (Porter, 1980) is probably the most well known stemming algorithm, there are specific algorithms for the Portuguese language (Orengo et al., 2001);

indexing – defining index terms, the features that will be used for document modeling, and computing their weights.

The application of these pre-processing tasks must be careful because the predictive power of words is highly dependent on the topic of interest (Chakrabarti et al., 1998a). Another essential aspect to consider is the language in which the document is written, which determines the list of stop-words and the stemming algorithm to use. Besides, stop-word removal always reduces information contained in the document; to avoid this loss some search services, like CiteSeer (Lawrence et al., 1999), do not remove any words from the documents to be indexed. Web documents written in HTML still require removing HTML tags.

In the web environment some other problems have to be dealt with, as, for instance, document duplicates that may exist and which should be detected by the system. While exact match between two documents is easy to detect the near duplicates problem is harder to deal with. Duplicate web pages may have slight differences between them – a date, the home URL or editor, to name a few, which raise difficulties to automated duplicate detection.

It is possible to estimate with a reasonable confidence the characteristic distribution of a set of features when the number of training examples is substantially higher than the number of distinct features, which is not the case in web page classification. In text mining the number of features is typically much larger than the number of available training examples and, if care is not taken undesirable overfitting may arise. Feature selection is desirable not only to avoid overfitting but also to keep the same level of accuracy while reducing feature space dimension, and consequently reducing the necessary computational effort. Feature reduction or feature selection techniques may be heuristic – governed by linguistic principles or specific rules from the universe of discourse – or statistical.

Feature selection techniques, in text classification, are usually applied following the process (Chakrabarti, 2003):

1. compute, for each feature, a measure that allows to discriminate categories;
2. list features in decreasing order of that measure and
3. keep the subset of the features with the highest discriminative power.

The *curse of dimensionality* (Koller et al., 1996) requires reducing the feature space dimensionality in order to improve any reasoning over this space. The high feature

space dimensionality can be reduced with techniques that might be categorized as feature selection or re-parameterisation techniques (Aas et al., 1999).

Feature selection attempts to remove non-informative words from documents in order to improve categorization effectiveness and reduce computational complexity while *re-parameterisation* is the process of constructing new features, as combinations or transformations of the original ones. (Yang et al., 1997) describes common feature selection techniques.

It should be stressed that words in the document are not the only features in web pages. Hypertext documents contain other types of features, such as links to and from other web pages and HTML tags, which may be explored, in isolation or in conjunction with others, constituting abstract features (Halkidi et al., 2003) which may hold significant predictive power.

After the data preparation phase, each document is reduced to its representative features and the next step, *data representation*, is to encode this view of the document into a specific format, ready for the learning stage.

Classic IR models (Boolean, vector and probabilistic) assume that each document is described by a set of representative keywords called index terms (Baeza-Yates et al., 1999). *Index terms* are words, or phrases, appearing in the document, whose semantics helps in remembering the document's main themes. Index terms are usually assumed to be independent.

From the *Boolean model* perspective, a query is a Boolean expression using the three connectives "and", "or" and "not". The weights associated with index terms are binary. There is no notion of partial match; each document is either relevant or not relevant at all to a given query; the lack of a relevance notion makes this a data retrieval instead of a real IR model. Simplicity and the clean formalism are the main advantages of this model while its main disadvantages come from the fact that exact matching may lead to retrieval of too many – poor precision – or too few – poor recall – documents.

The *Vector model*, probably the most commonly used, assigns real non-negative weights to index terms in documents and queries. In this model, documents are represented by vectors in a multi-dimensional Euclidean space. Each dimension in this space corresponds to a relevant term/word contained in the document collection.

The degree of similarity of documents with regard to queries is evaluated as the correlation between the vectors representing the document and the query which can be, and usually is, quantified by the cosine of the angle between the two vectors.

In the vector model, index term weights are usually obtained as a function of two factors:

- the *term frequency* factor, TF, a measure of intra-cluster similarity; computed as the number of times that the term occurs in document, normalized in a way as to make it independent of document length and
- an *inverse document frequency*, IDF, a measure of inter-cluster dissimilarity; weights each term according to its discriminative power in the entire collection.

This model main advantages are related to improvements in retrieval performance due to term weighting; partial matching that allows retrieval of documents that approximate the query conditions. The index term independency assumption is probably its main disadvantage.

Some proposals, distinct from the traditional vector space model, try to explore sequences of characters, using kernel functions to measure similarity between documents; the *string kernel* model. Text can further be represented as sequences of words, which are linguistically more meaningful than characters (Nicola et al., 2003).

Probabilistic models compute the similarity between documents and queries as the odds of a document being relevant to a query. Index term weights are binary. This model ranks documents in decreasing order of their probability of being relevant, which is an advantage. Its main disadvantages are: the need to guess the initial separation of documents into relevant and non-relevant; weights are binary; index terms are assumed to be independent.

Classic models for text retrieval view a document in its most primitive form: a text is a *bag-of-words*; the eventual text structure is not explored. Retrieval models, which combine information on text content with information on the document structure, are called *structured text retrieval* models. Web pages, which are mainly hypertext documents contain a set of features, not found in plain text documents, which can reveal to be highly informative. Hypertext features on web pages include: HTML tags, URLs, IP addresses, server names contained in URL, sub-strings contained in URLs, links from the current page to other pages – *out-links* – and links from other pages to the current page – *in-links*.

At the end of the 1980s and throughout the 1990s, various structured text retrieval models have appeared in the literature (Charkrabarti, 2003): *non-overlapping list model*, *proximal nodes model*, *simple concordance list models* (Dao, 1998), *path prefixing*, inductive classifiers, such as FOIL, applied to *relational models*.

4.2 Learning

Automated learning techniques, generally adopted from the machine learning and statistics fields, may improve the performance and the functionality of IR systems in a wide variety of forms, ranging from document classification to user behaviour modeling and information extraction.

Document classification is the task of assigning one or more pre-defined categories to documents. In the web environment we are usually interested in sets of several distinct classes; this is known as a *multi-class* problem. Besides, web documents are often related to more than one category, the *multi-label* characteristic. The problem of classification in IR is a multi-class and multi-label problem.

Text classifiers, from text mining field, deal primarily with flat text documents and do not take advantage of other potentially relevant features present in web documents. Web pages contain other features, besides flat text, like hyperlinks, content of neighbours and metadata that might help to improve classifiers' performance.

Several text mining algorithms, derived from the machine learning field, have been applied to the web document classification task. Although the performance of the text classifiers depends heavily on the document collection at hand (Yang, 1999),

some classifiers, particularly Support Vector Machines (SVM) and k-Nearest-Neighbours seem to outperform others in the majority of the domains. (Joachims, 1998) points out a few properties of text – high dimensional feature spaces, few irrelevant features, document vectors are sparse, most text categorization problems are linearly separable – that justify why SVM should perform well for text categorization.

When applied to web pages, classical flat text classifiers treat each page independently, ignoring links between pages and the class distribution of neighbour pages.

Some algorithms explore other sources of evidence, besides plain text. (Yang et al., 2002), define five types of regularities that might be present in a hypertext collection: *no regularity* – the only place that has relevant information about the class of the document is the document itself, *encyclopaedia regularity* – documents with a given class only link to documents with the same class, *co-referencing regularity* – some, or all, of the neighbouring documents belong to the same class but this class is distinct from the document class, *pre-classified regularity* – the regularity is present at the structural level where a single page (hub) points to several pages which belong to the same class – and *metadata regularity* – metadata is available from external sources and can be explored as additional sources of evidence generating new features. The authors then define the types of features that should be used in order to improve the classification task of documents belonging to each of these regularities.

(Chakrabarti et al., 1998b) also test several feature sets, similar to the ones suggested by (Yang et al., 2002): *local text*, *local text concatenated with all neighbours' text*, *local text plus neighbours text prefixed with discriminative tags*, to conclude that naive use of terms in the link neighbourhood of a document can even degrade performance.

In document classification, the class labels are frequently organized in taxonomies where classes are associated through inheritance. Techniques that explore this

structure are commonly referred to by *hierarchical classification* (Chakrabarti et al., 1998a).

Evaluation of systems' performance is a very important issue in IR; TREC is an annual conference especially devoted to this issue. Conventional evaluation is based on test collections. Test collections are composed by a set of documents, a set of queries and a set of relevant documents for each query or the specification of the relevance criteria to apply (Cormack et al., 1998; Rijsbergen, 1979 – chapter 7; Voorhees, 1998).

This evaluation can be based on several measures. Among them *recall* – ratio between the number of documents correctly classified and the total number of documents in the category – and *precision* – ratio between the number of documents correctly classified and the total number of documents classified in the category – assume a very important position.

Usually a classifier exhibits a trade-off between precision and recall, so these measures are negatively correlated: improvement in recall is made at the cost of precision and vice-versa. It is frequent to have text classifiers operating at the break-even point. The *break-even point* of a classification system is the operating point where recall and precision have the same value.

Accuracy – ratio of the number of correctly classified examples by the total number of evaluated examples – and *error* – ratio of the number of incorrectly classified examples by the total number of evaluated examples – are complementary measures of the error probability of the classifier given a certain category.

These measures are defined for binary classifiers. To measure performance in multi-class problems there are two common approaches to aggregate these measures evaluated at each singular category: macro-averaging and micro-averaging. *Macro-averaged* measures are obtained by first computing the individual scores, for each individual category, and then, averaging these scores to obtain the global measure. *Micro-averaging* measures are obtained by first computing the total number of documents correctly and incorrectly classified, for

all of the categories, and then using these values to compute the global performance measure, applying its definition. Macro-averaging gives equal weight to each category while micro-averaging gives equal weight to every document.

In current web IR, it is typical that users specify their information need with a set of keywords. Although this form of interaction is very simple and relies on universally understood specification language and protocol, keywords are weak primitives to specify an information need (with high imprecision and subject to each user's interpretation), resulting in ambiguous, incomplete or excessive specifications and, consequently, poor result sets. These result sets are analysed and validated by the user. The user implicitly analyses the retrieval system performance when he explores the result set. It is reasonable to expect that if a user downloads many of the returned documents, then they probably have some relevance; and, on the other hand, if the user does not download a given document it is probable that this particular document is not very relevant to the user need. The user may even be required to explicitly indicate perceived relevance for a given set of documents. This feedback from the user, either explicit or implicit, might be used in order to improve its performance, particularly as to the relevance of returned documents; this problem is known as *relevance feedback*. Relevance feedback techniques usually produce some query modifications, by adding, removing or changing terms, originating internal queries that are different from the original user query and which are expected to be more representative of user's needs. The common approach is to retrieve an initial set of relevant (and eventually also non-relevant) documents and discover correlated features that can be used to narrow down the original query, improving precision (Chakrabarti, 2003; Glover et al., 2001; Mitra et al., 1997).

4.3 Presentation

In IR systems the user typically interacts with the system in two distinct moments: when specifying information needs and when analysing the results.

Query specification requires the user to describe information needs as a set of keywords and, eventually, some other characteristics to be met, such as the domain and file format. In traditional search engines the user specifies the query in one

single step, while interactive query systems (Bruza et al., 2000) require the user interaction during the initial tuning phase where the user iteratively refines the query by following suggestions from the system.

The analysis of the results is a valuable aspect of search services, however, public search services have not been given it much of importance. Results from a search engine are usually presented as a flat ranked list. Retrieval systems are a black box from the user point of view – the user submits a query and receives an answer without having any knowledge on the decisions, processes and data analysed to provide the answer – which obstructs the analysis of the results. Besides, in a flat list, users do not have a way of analysing a global perspective of the entire collection neither of analysing a single object while still keeping a global image of all the collection. Several difficulties may arise from these facts (Cugini et al., 1996; Olsen et al., 1992).

A visualization tool might help in solving some of these difficulties, providing functionalities that allow for selecting valuable information from large document collections without much effort. The flat list approach is very restrictive when the object collection is large (Carey et al., 2000); this approach subordinates the user to a slow, intensive mental process of reading through the reference list while in a graphical system the user just has to recognize patterns on a visual display, which is a much faster and less demanding process.

Post-retrieval visualization techniques may be categorized, according to their main goal, in the following categories (Zamir et al., 1999): on one hand those that stress inter-document similarities, usually based on document content, and which help the user getting an overview of the full document collection as well as visualizing clusters of topically related documents and, on the other hand, those that aim to display additional information about retrieved documents, which might be predefined document attributes (such as size, date, age, author or source) or user-specified attributes (such as predefined categories).

Although public search services rarely explore post-retrieval visualization techniques, there are many distinct proposals. We briefly review some of them:

(Carey et al., 2000) claim that an *unified approach* – allowing the user to explore the corpus through several different visualization paradigms, which should be integrated, in order to allow for cross comparisons and evaluations – will be valuable. They include Sammon Maps, Tree-Map visualization and Radial visualization in this consolidated framework.

Sammon maps try to represent *n*-dimensional objects in a 2-dimensional space while attempting to preserve the pair-wise distances between objects.

Tree-Map visualization represents clusters as rectangles arranged in such a way as to fit inside a larger rectangle. The area of a rectangle is proportional to the size of the cluster it represents. Similar clusters are grouped together in *super-clusters*.

Radial visualization is a common paradigm where keywords are represented as nodes, which are homogenously displayed, usually around a circumference, in the display space. Documents are then placed as if they were connected to the relevant keyword nodes by forces that attract them proportionally to the weight of the word in the document.

The *VIBE* system (Olsen et al., 1992) explores this radial visualization paradigm by defining *Points Of Interest* as a set of representative keywords working as anchor nodes relatively to which documents are placed in a bi-dimensional space.

(Spangler et al., 2003) represents relevant categories, associated to the query result set, in a *Radial graph*, allowing the user to understand which categories are strongly related to the query and associates this view with a binary tree that allows to refine a query, once the user as chosen a given category.

(Viji, 2002) describes a visualization tool that uses springs to represent documents semantic correlation. The length of a spring connecting two documents is proportional to their correlation. Correlation between two documents is based on textual content and link based information.

The *Document Spiral* paradigm, proposed in (Cugini et al., 1996) displays documents along a spiral in a two dimensional space. Documents at the centre of the spiral are the most relevant; relevance decreases as one travels along the spiral.

Kohonen maps apply neural network principles to self-organize document collections into a *galaxy* – a galaxy is simply a group of clusters represented in a two-dimensional space, one of the simplest forms of cluster mapping (Chewar et al., 2001).

Star coordinates is an exploratory visualization technique for organizing and representing n-dimensional spaces on a two-dimensional surface. Together with interactive functionalities is an effective tool to explore large corpus and to detect clusters (Kandogan, 2001).

5. Research in Information Retrieval

The amount, heterogeneity and dynamics of available information on the web along with the vast number of activities that are available online demand for new approaches to the problem of searching the web. Understanding the task that users are trying to perform becomes a relevant part of the problem together with the perception of their specific need. Search systems still struggle to deliver quality answers and new sources of evidence are being explored.

Open research problems on the IR field are recognized by several experts. We resume some of these perspectives.

(Baeza-Yates et al., 1999) identified a set of research directions, organized by specific areas on the IR field. Among them we selected those that seem more relevant:

- Retrieval of higher quality;
- Combining several evidential sources to improve relevance judgements;
- Querying on the structure of the text besides content. Visual query languages;
- Interactive user interfaces;
- New retrieval strategies, based on the understanding of user behaviour;
- User-centered design; cognitive and behavioural issues;
- Understanding the criteria by which users determine if retrieved information meets their information needs.

This list stresses the importance of improving retrieval quality, combining several evidential sources and focusing the retrieval process on the user.

(Chakrabarti, 2003) refers, among other, the importance of combining several external sources of evidence and personalization of search systems:

- Disassembling pages and page zones into finer structures, such as phrases and sentences;
- Integration of several external sources of evidence, such as lexical networks (WordNet), thesaurus and ontologies;
- Profiles, personalization and collaborative search;
- Modeling of semantics (web pages, user, user needs).

(Croft, 2003) refers to global information access and contextual retrieval:

- Leveraging worldwide structured and unstructured data in any language;
- Combine search technologies, knowledge about query and user context to provide the most appropriate answer for a user's information need.

(Henzinger et al., 2003) refers a more specific list of problems:

- Spam; pro-active approaches to detect and avoid spam (text spam, link spam, cloaking or combined), collectively known as adversarial search;
- Combine link-analysis quality judgments with text-based judgments to improve answer quality;
- Estimating web page quality from web structure (within a page, across different pages);
- Quality evaluation; implicit relevance feedback, click-through data;
- Detect and explore web conventions; understanding the nature of links (commercial, editorial, metadata);
- Vaguely structured data; try to infer semantic information from HTML tags, since layout conveys semantic information.

(Shwarzkopf, 2003) emphasizes the importance of conveniently organizing documents in the answer. Interaction with information does not end with retrieving relevant documents; it also includes making sense of the retrieved information, organizing collected materials according to user needs.

(Sahami, 2004) also refers to high quality search results, dealing with spam and search evaluation:

– Identify which pages are of high quality and relevance to a user's query;
– Linked-based methods for ranking web pages;
– Adversarial classification, detecting spam;
– Evaluating the efficacy of web search engines;
– Determining the relatedness of fragments of text, web contextual kernel;
– Retrieval of images and sounds;
– Harnessing vast quantities of data.

(Apostolico et al., 2006) stress the importance of query expansion, search evaluation and retrieval from XML sources:

– Measures to assess IR system's efficiency and to compare it with others;
– Efficient query expansion;
– Query performance prediction, particularly in the case of query expansion;
– Retrieval model and query language for XML documents.

Among all these perspectives on IR research we detect some common trends:

Retrieval of high quality seems to be still the most relevant aspect to be solved. The majority of research efforts on IR follow this major goal. Recently, as the amount of human activity online increases, the task that the user is trying to perform while using IR systems assumes relevance as an indicator of what is the real need behind the query (Broder, 2002). Understanding and classifying user queries is an important step (Betzel, 2006). Information overload is a web characteristic that requires high quality retrieval; otherwise IR systems will fail their goal because they will not be able to produce answers made of (small) sets of relevant

documents. Besides, this high quality must be achieved without requiring explicit user effort. Semi-supervised classification methods (Li et al., 2003; Nigam et al., 2000; Bennet et al., 1998; Blum et al., 1998), specifically applied on the web environment might improve retrieval quality while reducing user's workload. Conditional Random Fields (Lafferty et al., 2001) may also help improving retrieval quality.

Web IR is also being applied to new specific ways of using the web. TREC introduced a new track in 2006 that deals with retrieval from the blogosphere. (Sahami, 2004) refers to specific methods for ranking UseNet or bulletin board postings. The Web 2.0 paradigm (O'Reilly, 2004), based on active users, reinforces the dynamic nature of the web and originates new challenges and opportunities.

Integration of several sources of evidence is being explored by researchers trying to improve the modeling of users and user needs. Several distinct features are being considered and analyzed: linguistic approaches (Arcot, 2004), context sensitive search (Crestani, 2007; Haveliwala, 2005; Ifrim et a, 2005; Zakos et al., 2006), task or topic-based analysis of queries (Beitzel, 2006), query-dependent PageRank (Richardson et al., 2004), Wikipedia-assisted feedback (Liu et al., 2005), semantic models (Siddiqui et al., 2006) and phrase-based indexing (Hammouda et al., 2004) are some examples of research on this subject. Exploring new ways of integrating distinct feature sets, such as Formal Concept Analysis (Shen, 2005; Wolf, 1993) or Markov Logic (Domingos, 2007: Domingos et al., 2006) may produce interesting results.

Personalization issues are also being explored (Escudeiro et al., 2006). Information needs are user specific and IR systems should provide user specific answers, organized and presented according to particular users or groups of users' specific interests.

Besides these long breath problems there are a few more specific problems generating research interest, such as *adversarial search* that deals with spam, *retrieval from XML* sources and exploring *web conventions*.

6. Conclusions

The web is a vast repository of information with some characteristics that are adverse to IR: large volume of data, mainly unstructured or semi-structured; dynamic nature; content and format heterogeneity and irregular data quality are some of these adverse characteristics. These specific web characteristics require specific treatment.

The user also introduces additional difficulties to the retrieval process, such as the semantic gap, arising from ambiguous query specifications and the fact that the required organization of the answer and the aim the user is seeking are not fed to the IR system.

Despite these difficulties the web is being used as an information source as well as a support for an increasing number of activities by an increasing number of people with rather distinct background, motivations and needs.

All these aspects give us reasons to believe that the web IR field is, and will remain, relevant and challenging to academic and economic areas.

References

Aas, K., Eikvil, L. (1999), *Text Categorization: A Survey*, Norwegian Computing Center

Aggarwal, C.C., Al-Garawi, F., Yu, P. (2001), *Intelligent crawling on the World Wide Web with arbitrary predicates*, Proceedings of the 10th World Wide Web Conference

Aggarwal, C.C. (2004), *On Leveraging User access Patterns for Topic Specific Crawling*, Data mining and Knowledge Discovery, 9, pp 123-145, Kluwer Academic Publishers

Apostolico, A., Baeza-Yates, R., Melucci, M. (2006), *Advances in information retrieval: an introduction to the special issue*, Journal of Information Systems, Elsevier Science Ltd., 31(7), p.569-572

Arcot, H.G.A. (2004) *Perception-based fuzzy information retrieval*. United States -- California: San Jose State University

Baeza-Yates, R. (2003), Information Retrieval in the Web: beyond current search engines, *Elsevier International Journal of Approximate Reasoning*, 34, pp 97-104

Baeza-Yates, R., Ribeiro-Neto, B. (1999), *Modern Information Retrieval*. ACM Press

Baldi, P., Frasconi, P., Smyth, P. (2003), *Modeling the Internet and the Web. Probabilistic Methods and Algorithms*, Wiley

Beitzel, Steven M. (2006) *On understanding and classifying web queries*, PhD dissertation USA, Illinois, Illinois Institute of Technology

Bennet, K.P., Demiriz, A. (1998), Semi-Supervised Support Vector Machines, *Proceeding of Neural Information Processing Systems*

Berners-Lee, T. (1989), *Information Management: a proposal*., CERN

Berners-Lee, T., Hendler, J., Lassila, O. (2001), The Semantic Web. *Scientific American*

Blum, A., Mitchell, T. (1998), Combining labelled and unlabelled data with Co-training, *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pp 92-100

Borges, J.L.C.M. (2000), *A Data Mining Model to Capture User Web Navigation Patterns*, PhD dissertation, University of London

Brin, S., Page, L. (1998), "The anatomy of a large-scale hypertextual web search engine", *Proceedings of the 7th World Wide Web Conference*, pp 107-117

Broder, A. (2002) A taxonomy of web search. *SIGIR Forum*. 36:2. p. 3-10

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J. (2000), *Graph structure in the web* World Wide Web Conference, Amsterdam, Holand

Broder, A., Maarek, Y., Bharat, K., Dumais, S., Papa, S., Pedersen, J., Raghavan, P.(2005), Current Trends in the Integration of Searching and Browsing, Special interest tracks and posters of the 14th World Wide Web Conference , Chiba, Japan, p.793

Bruza, P., McArthur, R., Dennis, S. (2000), *Interactive Internet search: keyword, directory and query reformulation mechanisms compared*, Research and Development in Information Retrieval

Bush, V. (1945), *As We May Think*, The Atlantic Monthly, July

Carey, M., Kriwaczek, F., Ruger, S.M. (2000), A Visualization Interface for Document Searching and Browsing, *Proceedings of the NPIVM 2000*

Chakrabarti, S. (2003), Mining the Web. Discovering Knowledge from Hypertext Data. Morgan Kaufmann Publishers

Chakrabarti, S., Dom, B., Agrawal, R., Raghavan, P. (1998a), Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies, *The VLDB Journal*, 7, pp 163-178

Chakrabarti, S., Dom, B., Indyk, P. (1998b), Enhanced hypertext categorization using hyperlinks, *Proceedings of ACM SIGMOD International Conference on Management of data*, pp 307-318

Chakrabarti, S., Byron, E., Kumar, S., Raghavan, P., Rajagopalan, S., Tomkins, A., Gibson, D., Kleinberg, J. (1999a), Mining Web Link Structure, *IEEE Computer*, 32(8), pp 60-67

Chakrabarti, S., Berg, M., Dom, B. (1999b), *Focused crawling: a new approach to topic-specific resource discovery*, Proceedings of the 8th World Wide Web Conference

Chewar, C.M., Krowne, A., O´Laughlen, M. (2001), *User Object Collections: Visualization Concepts by collection-Insight Need*, CITIDEL project

Cho, J., Garcia-Molina, H. (2000), *Estimating Frequency of Change*, Technical report, Stanford University

Cleverdon, C.W. (1991), *The significance of the Cranfield tests on index languages*, Proceedings of the ACM – SIGIR, p. 3-12

Cleverdon, C.W. (1962), *Comparative Efficiency of Indexing Systems*, Cranfield

Cleverdon, C.W., Aitchison, J. (1963), *Test of the Index of Metallurgical Literature*, Cranfield

Cleverdon, C.W., Thorne, R.G. (1954), *An Experiment with the Uniterm System*, R.A.E. Cranfield, 7

Codd, E.F. (1970), A Relational Model of Data for Large Shared Data Banks, *Communications of the ACM*, Vol. 13, No. 6, June 1970, pp. 377-387

Cooley, R., Mobasher, B., Srivastava, J.(1997), Web Mining: Information and Pattern Discovery on the World Wide Web, *Proceedings of the 9th IEEE International conference on tools with Artificial Intelligence*, pp 558-567

Cormack, G.V., Palmer, C.R, Clarke, C.L.A. (1998), *Efficient Construction of Large Test Collections*, Proceedings of the ACM SIGIR 1998 Conference

Crestani, F., Shengli, W. (2006), Testing the cluster hypothesis in distributed information retrieval, *Information Processing and Management*. 42, p. 1137-1150

Crestani, F., Ruthven (2007), I., Introduction to special issue on contextual information retrieval systems. *Information Retrieval*. 10, p. 111-113

Croft, W.B. (2003), *Information retrieval and computer science: an evolving relationship*, ACM SIGIR Conference, Toronto, Canada, p.2-3

Cugini, J., Piatko, C., Laskowski, S. (1996), Interactive 3D Visualization for Document Retrieval, *Proceedings of the ACM Conference on Information and Knowledge Management*

Dao, T. (1998), *An Indexing Model for Structured Documents to Support Queries on Content, Structure and Attributes*, Proceedings of IEEE ADL Conference, Santa Barbara, California, USA

Dewey, M. (2004), *A Classification and Subject Index for Cataloguing and Arranging the Books and Pamphlets of a Library*, Project Gutenberg EBook

Domingos, P. (2007), What's missing in AI: The Interface Layer, University of Washington, Washington, USA

Domingos, P., Kok, S., Poon, H., Richardson, M., Singla, P. (2006), Unifying Logical and Statistical AI, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, Boston, Massachusetts, USA

Donato, D., Laura, L., Millozi, S. (2000), A beginner's guide to the Webgraph: Properties, Models and Algorithms, *Proceedings of the 41st FOCS*, pp.57-65

Escudeiro, N., Jorge, A., (2006) Semi-automatic Creation and Maintenance of Web Resources with webTopic. *Semantics, Web and Mining*. LNCS, vol. 4289, pp. 82-102, Springer, Heidelberg

Glover, E.J., Flake, G.W., Lawrence, S., Birmingham, P., Kruger, A., Giles, C.L., Pennock, D.M. (2001), Improving Category Specific Web Search by Learning Query Modifications, *Symposium on Applications and the Internet, IEEE Computer Society*, pp 23-31

Gulli, A., Signorini A. (2005), *The Indexable Web is More than 11.5 billion pages*. In: WWW 2005, Chiba, Japan

Halkidi, M., Nguyen, B., Varlamis, I., Vazirgiannis, M. (2003), "Thesus: Organizing Web document collections based on link semantics", *The VLDB Journal*, 12, pp 320-332

Hammouda, K.M., Kamel, M.S. (2004), Efficient Phrase-Based Document Indexing for Web Document Indexing. *IEEE Transactions on Knowledge and Data Engineering*. 16:10, p. 1279-1296

Haveliwala, T.H. (2005), *Context-sensitive Web search*, PhD dissertation, Stanford University, California, USA

Henzinger, M., Motwani, R., Silverstein, C. (2003), *Challenges in Web Search Engines*, 18th International Joint Conference on Artificial Intelligence

Hersovici, M., Jacovi, M., Maarek, Y.S., Pelleg, D., Shtalaim, M., Ur, S. (1998), The Shark-search algorithm. An application: tailored web site mapping, *Computer Networks* 30(1-7), pp 317-326

Hu, W., (2002), World Wide Web Search Technologies, *Architectural Issues of Web-Enables Electronis Business*, edited by Shi Nansi for Idea Group Publishing

Ifrim, G., Theobald, M., Weikum, G. (2005), Learning Word-to-Concept Mappings for Automatic Text Classification, International Conference on Machine Learning

Jardine, N., Rijsbergen, C.J. (1971), The use of hierarchic clustering in information retrieval, *Information Storage and Retrieval*, 7(5), pp. 217-240

Joachims, T. (1998), *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*, Research Report of the unit no. VIII(AI), Computer Science Department of the University of Dortmund

Kandogan, E. (2001), Visualizing Multi-dimensional Clusters, Trends, and Outliers using Star Coordinates, *Proceedings of the KDD Conference*, San Francisco, California, USA

Kahle, B. (1997), Preserving the internet, *Scientific American*. 276:3, p. 82-83

Kleinberg, J. (1998), Authoritative sources in a hyperlinked environment, *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*, pp 668-677

Koller, D., Sahami, M. (1996), Toward Optimal Feature Selection, *Proceedings of the 13th International Conference on Machine Learning*, pp. 284-292, Morgan Kaufmann

Kosala, R., Blockeel, H. (2000), Web Mining Research: A Survey, *SIGKDD Explorations*, Vol. 2, No. 1, pp 1-13

Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., Upfal, E. (2000), The Web as a graph, *Proceedings of the 19th ACM SIGACT-SIGMOD-AIGART Symp. Principles of Database Systems*

Lafferty, J., McCallum, A., Pereira, F. (2001), *Conditional random fields: Probabilistic models for segmenting and labeling sequence data* , 18th International Conference on Machine Learning, 2001

Lawrence, S., Bollacker, K., Giles, C.L. (1999), Indexing and Retrieval of Scientific Literature, *Proceedings of the 8th International Conference on Information and Knowledge Management*, pp 139-146

Lewandowski, D. (2005), Web searching, search engines and Information Retrieval. *Information Services and Use*. 25:3-4/2005, p. 137-147

Li, X., Liu, B. (2003), Learning to classify text with positive and unlabelled data, *Proceeding of IJCAI – 2003*

Lim, L., Wang, M., Padmanabhan, S., Vitter, J.S., Agarwal, R. (2001), Characterizing Web Document Change, *Lecture notes in Computer Science*

Liu, R. L., Lin, W. J.(2005), Incremental mining of information interest for personalized web scanning, Information Systems journal, 30(8), p. 630-648

Lu, S., Dong, M., Fotouhi, F. (2002), The semantic web: opportunities and challenges for next generation web applications, *Information Research*, 7 (4)

Mitra, M., Singhal, A., Buckley, C. (1998), Improving automatic query expansion, *Proceedings of the 21st ACM SIGIR Conference*

Nelson, T. (1965), *A file structure for the complex, the changing, and the indeterminate*, ACM National Conference, 84-100

Nicola, C., Gaussier, E., Goutte, C., Renders, J. M. (2003), "Word-Sequence Kernels", *Journal of Machine Learning Research*, Nº 3, pp 1053-1082

Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.M. (2000), Text classification from labeled and unlabeled documents using EM, *Machine Learning*, 39, pp 103-134

Olsen, K.A., Korfhage, R.R., Sochats, K.M., Spring, M.B., Williams, J.G. (1992), Visualization of a Document Collection: The VIBE System, *Information Processing & Management*, Vol. 29, No. 1, pp 69-81

Orengo, V., Huyck, C. (2001), A Stemming Algorithm for the Portuguese Language, *Proceedings of the 8th SPIRE*

O'Reily (2004), *Web 2.0*

Porter, M.F. (1980), "An algorithm for suffix stripping", *Program*, 14, No. 3, pp 130-137

Richardson, M., Domingos, P. (2004) Combining Link and Content Information in Web Search, Washington University, Washington, USA

Rijsbergen, K. (1979), *Information Retrieval*, Butherworth

Sahami, M. (2004),*The happy searcher: Challenges in the web information retrieval*, Pacific Rim International Conference on Artificial Intelligence, 3157, p.3-12

Salton, G., Lesk, M.E. (1965), The SMART automatic document retrieval systems - an illustration, *Communications of the ACM*, 8:6 (June 1965), p.391-398

Salton, G., McGill, M. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill

Salton, Wong, Yang (1975), A vector space model for automatic indexing. *Communications of the ACM*. 18:11 (1975). p. 613-620

Shen, G. (2005), Formal concepts and applications, PhD dissertation, Case Western Reserve University, Ohio, USA

Shwarzkopf, E. (2003), *Personalized Interaction with Semantic Information Portals*, German Research Center for Artificial Intelligence

Siddiqui, Tanveer, J. (2006), *Intelligent Techniques for Effective Information Retrieval (A Conceptual Graph Based Approach)*, ACM SIGIR Forum. 40:2

Spangler, S., Kreulen, J.T., Lessler, J. (2003), *Generating and Browsing Multiple Taxonomies Over a Document Collection*, Journal of Management Information Systems, 19(4), p. 191-212

Viji, S. (2002), *Term and Document Correlation and Visualization for a set of Documents*, Technical report, Stanford University

Voorhees, E.M. (1998), *Variations in Relevance Judgements and the Measurement of Retrieval Effectiveness*, Proceedings of the ACM SIGIR 1998 Conference

Wang, J., Lochovsky, F. (2003), "Web Search Engines", Journal of ACM Computing Survey (accepted for revision)

Wolf, K.E. (1993), *A First Course in Formal Concept Analysis*, Advances in Statistical Software, 4, p. 429-438

Yang, Y. (1999), An Evaluation of Statistical Approaches t

o Text Categorization, *Journal of Information Retrieval*, vol. 1, nos. 1/2, pp 67-88

Yang, Y., Pederson, J. (1997), "A Comparative Study of Feature Selection in Text Categorization", *International Conference on Machine Learning*

Yang, Y., Slattery, S., Ghani, R. (2002), *A Study of Approaches to Hypertext Categorization*, Kluwer Academic Publishers, pp. 1-25

Zakos, J., Verma, B. (2006), A Novel Context-based Technique for Web Information Retrieval, World Wide Web, 9(4), p. 485-503

Zamir, O., Etzioni, O. (1999), Grouper: A Dynamic clustering Interface to Web Search Results, *Proceedings of the 1999 World Wide Web Conference*