# New Methods and Tools for the World Wide Web Search

Vlatko Ceric

Faculty of Economics, University of Zagreb, Croatia

Explosive growth of the World Wide Web as well as its heterogeneity call for powerful and easy to use search tools capable to provide the user with a moderate number of relevant answers. This paper presents analysis of key aspects of recently developed Web search methods and tools: visual representation of subject trees, interactive user interfaces, linguistic approaches, image search, ranking and grouping of search results, database search, and scientific information retrieval. Current trends in Web search include topics such as exploiting Web hyperlinking structure, natural language processing, software agents, influence of XML markup language on search efficiency, and WAP search engines.

*Keywords:* Web search, search engines, subject trees, scientific information retrieval, database search, ranking, XML, WAP search engines

## 1. Introduction

World Wide Web had an astonishing growth in the last few years. Number of Web servers increased from 1.7 million in December 1997, to 3.7 million in December 1998, to 9.6 million in December 1999, and then to 21.2 million in September 2000 (URL 1). The same source estimates that the number of Internet hosts increased from 29.7 million in January 1998, to 43.2 million in January 1999, to 72.4 million in January 2000, and then to 93 million in July 2000. The number of people online is assessed to be over 370 million as of September 2000 (URL 2).

One of the most authoritative studies of Web size (Lawrence and Giles, 1999) estimated that in February 1999 the publicly indexable Web contained approximately 800 million pages, the quantity of information equivalent to 6 terabytes of pure text, as well as some 180 million images equivalent to about 3 terabytes. As a comparison, Library of Congress, the biggest library today, contains about 20 terabytes of text. No single search engine was indexing more than about 16% of Web sites, and all major search engines combined were covering only about 42% of the Web.

The latest study of the Web size done by Cyveillance company (URL 3) estimated that in July 2000 Web contained about 2.1 billion unique publicly available Web pages. The study also found that Web is growing by about 7.3 million Web pages a day. About 84% of all pages are based in the USA.

It is estimated that about 30% Web pages are copied or mirrored, or are very similar (Shivakumar and Garcia-Molina, 1998). Analysis of 200 million Web pages (Broder et al., 1999) has shown that these pages use 1.5 billion hyperlinks and that the Web is made up of four components. A central core (about 56 million pages) is well connected, an 'in' cluster (44 million pages) contain pages that point to the central core but cannot be reached from it, an 'out' cluster (44 million pages) contains pages that can be reached from the core but do not link to it, 'tubes' (44 million pages) connect to either 'in' or 'out' clusters but not to the core, while some 17 million pages are completely unconnected. Web graph structure is interesting in itself, but it can also be used for designing crawl strategies on the Web (Cho and Garcia-Molina, 2000).

Web sites have extremely variable information value since they are written by individuals with any conceivable background, education, culture, interest, and motivation, and in most cases

these materials are neither edited nor reviewed. Web pages size may range from a dozen words to hundreds screens, and often contain grammatical mistakes and typos. Presented material can be obsolete, false or unreliable. Web content itself is unstructured, and multiple formats, different languages, dialects and alphabets are used (Baeza-Yates and Ribeiro-Neto, 1999).

Despite all its weakness, including a lot of poor, obsolete and unreliable material, Web includes a vast quantity of excellent material of all sorts, from practical information to professional and scientific material that cannot be found without appropriate tools. Web search is also an exciting activity in itself, since search is one of the fundamental human activities. The more ordered and structured material is, it is less exciting — the Web, with its heterogeneous and unstructured material, is therefore the perfect search area.

Without software tools for searching and cataloguing of the huge and heterogeneous Web information space it would be absolutely impossible to find relevant information residing on the Web (Ceric, 1997; Baeza-Yates and Ribeiro-Neto, 1999). The need for retrieval of information on the Web is so large that the major search engines receive about 15-20 thousand queries per minute.

What users need are powerful, easy to use search tools capable to provide a moderate number of appropriately ranked relevant answers. In the last few years numerous advancements were made in various areas of the World Wide Web search. This paper presents analysis of major aspects of recently developed Web search methods and tools, as well as of the most important trends in Web search methods.

## 2. Subject trees

Both subject trees (directories) and search engines cope with a serious problem of how to follow a tremendous growth of the Web. Since Web links are incorporated into subject trees by human indexing teams, the size of the team determines the coverage of Web contents by the subject tree. Biggest subject trees have the following approximate sizes (as of mid 2000): Yahoo! has the staff of about 150 editors and includes about 1.5 million Web links, LookSmart has about 200 editors and consist of about 2 million Web links, while NBCi (Snap) has about 50 editors and contains approximately 1 million Web links. One of the most valuable mid-sized subject trees Britannica contains about 130,000 Web links.

Since Web today contains roughly two billion Web pages, biggest subject trees cover only the order of magnitude of one promile of the Web. Although this is a very small coverage, the value of subject trees is primarily in the quality of Web links selected by its editors.

Two interesting new subject trees are Open Directory and About.com. *Open Directory* (URL 4) is based on the concept in which Web community, rather than professional staff, select Web links for a subject tree. Every one of approximately 28,000 editors that build and maintain the directory is an expert in the categories she takes care of. Open Directory contains about 2 million links categorized in more than 300,000 subcategories. Open Directory makes its data available to anyone, free of charge — some of the major search engines using Open Directory are HotBot, Lycos and Netscape. *About.com* (URL 5) is a mid-sized subject tree that maintains about 700 targeted environments (i. e. wide subcategories), each overseen by a professional guide. About.com guides are company-certified subject specialists living and working in over 20 countries and selected by a rigorous certification program. About.com environments have well-organized and rich content that includes new links, articles, newsletter, forums, chat, etc.

*Brittanica.com* (URL 6) is one of the first subject trees that introduced other contents in addition to Web links. This directory provides the complete and updated Encyclopaedia Britannica, as well as selected articles from 70 world's top magazines (like The Economist, Discover and Newsweek) and database of related books.

Subject trees have become so big that traditional user interfaces via hierarchic trees became inadequate for dealing with information since they don't enable the view of the whole information structure, while navigation is done by slow fold-and-unfold process. Such situation led to development of several user interfaces, e. g. Hyperbolic Tree, Mapccucino and ThemeScape, with

novel representation of the subject tree structure. *Hyperbolic Tree* (URL 7), developed by the Xerox Palo Alto Research Center, is a user interface designed for managing and exploring large hierarchies of data such as Web subject trees, product catalogs or document collections. It allows an overview of the whole data set, as well as drilling down and accessing the primary information (e. g. specific Web pages). Hyperbolic Tree is an intuitive and highly interactive tool that allows focusing on the part of data represented in context of the whole unfocused data set. User can easily change the focus via animated moving of hyperbolic tree elements by means of click-and-drag with the mouse.

*Mapuccino* (URL 8) is an IBM product that also dynamically construct visual maps of Web sites using their linking structure. *ThemeScape* (URL 9) organizes a topographical map of textual documents based on the similarity of their content. This software reads documents from the data set and examines their contents using sophisticated statistical and natural language filtering algorithms. Related topics found in the documents are associated, thus forming different concepts. These concepts are assigned to the spatial structure that is transformed into topological maps in such a way that the greater the similarity between two documents the closer together they appear on the map. Therefore, peaks on the map denote a concentration of several documents about the same topic. A resulting topographical map clearly shows which topics are covered in the data set, how much emphasis is given to the specific topic, and how different topics relate to one another. Search across the documents results in highlighting of relevant documents right on the map, instead of showing the list of discovered documents.

## 3. Search engines

Fast growth of the Web led to expansion of search engines data bases. Some of the major search engines have the following size of databases (as of June, 2000): Google database has 560 million Web pages indexed, WebTop.com has 500 million pages, AltaVista and FAST have about 350 million pages, while Northern Light has 265 million pages. The most recent figure for the FAST database size is 575 million pages as of October, 2000.

*FAST* (URL 10) is a unique new search engine developed by a group of computer scientists from the Norwegian University of Science and Technology in Trondheim having a long expertise in both search and image/video compression technologies. This search engine is based on the massively parallel architecture running on Dell PowerEdge X86 servers that give the impressive speed and scalability to this search engine. Designers of the FAST Web search engine intend to build the world largest search engine that intends to index the entire Web contents till the end of 2000, a goal that will be almost impossible to fulfill because of the explosive Web size growth. Its spider (crawler, robot) provides data for indexing by scanning the Internet with the throughput of up to 80 million documents per day, thus ensuring a fresh catalog. For the purpose of the FAST Web search the document index is broken into discrete segments of 5 million documents each, with each segment stored on the independent search nodes. Queries are received by multiple dispatch nodes which broadcast them to every search node simultaneously. Search results are assembled and ranked according to the relevancy score. Parallel processing enables that search through the whole index takes less than one second.

*Northern Light* (URL 11) search engine introduced supplementary contents for searching. Besides Web links, this search engine offers the Special Collection, an online business library that consists of over 6,000 full-text journals, books, magazines, newswires and reference sources. Special Collection currently has about 16 million documents, and adds approximately 250 new sources each month. Most of the Special Collection sources include articles back to January 1995. Price of the Special Collection documents range from $1.00 to $4.00 per article, while the summary of the article is free.

One of the most important things for efficient use of the Web search is proper ranking of the resulting Web links. Two new approaches developed in this field were done by Northern Light and Google search engines. *Northern Light* introduced automatic grouping of resulting Web links into meaningful categories (so called Custom Search Folders), that contain only information relevant to that folder. These folders group

results by subject, type (e. g. press releases or product reviews), source (e. g. commercial sites or magazines) and language. Users can concentrate on links from the appropriate folder and thus narrow the results considerably and save time. Folders are not preset, but are rather dynamically created for each search. *Google* (URL 12) search engine uses the PageRank algorithm based on the importance (popularity) of Web pages. Importance of a Web page is discovered through analysis of its link structure, and doesn't depend on the specific search request. PageRank is a characteristic of the Web page itself — it is higher if more Web pages link to this page, as well as if these Web pages have high PageRank. Therefore, important Web pages help to make other Web pages important. One consequence of this approach is that search result may include links to Web pages that were not found by Google spider, but were linked by some Web page accessed by the spider.

Two innovative approaches to Web search are visual approach for searching the Web, and search by meaning. *Ditto.com* (URL 13) is the service that provides visual mechanism for searching the Web. Search begins in a traditional manner, by entering one or more keywords. Ditto.com responds by presenting pictures from discovered Web pages (plus pictures from its own media collection). Presented pictures from Web pages act as a visual clue for determining the relevance of the Web page. Click on the presented picture from particular Web page lead us to that Web page.

Simpli.com and Oingo are search engines dealing with the problem of lexical ambiguity of the traditional search with one or more search terms put out of context. Since words can have many different meanings (polysemy), traditional search engines respond by selecting all Web pages that fit each possible meaning of the search terms. This leads to a large list of resulting links that includes many links not related to the desired meaning of the search. Because there are many words with the same or similar meaning (synonymy) it can happen that keyword-based search cannot find relevant pages because it uses one synonym while authors of Web pages may use some other synonym. *Simpli.com* (URL 14) is using principles of linguistic and cognitive science as well as the interaction with users to place search terms in context. When the user enters the search term Simpli.com activates its knowledge base and automatically generates a list of possible contexts of this search term. User then interacts with the search engine and chooses the appropriate meaning (concept). After that the search engine consults its database to choose related words based on the search term and the chosen concept, and automatically expands the query with these words. *Oingo* (URL 15) takes the similar approach and enables using more than one keyword for search. Oingo offers its user a list with possible meanings for all of the terms used in a query. The user then chooses the most appropriate meaning for each query term, and these meanings are used for the search.

*AskJeeves* (URL 16) search engine has two interesting features. One of them is that questions can be stated in natural English. AskJeeves uses natural language processing technology to analyze the meaning of the words in question, and the meaning in the grammar of the question. This search engine also keeps a database of questions and answers to these questions and expands it all the time, hence becoming smarter during time. Currently it has more than 7 million answers that contain information about the most frequently asked questions and the answers to these questions discovered on the Web. When a user asks a new question AskJeeves analyses its meaning and provides to the user a list of specific questions. When the user clicks on such specific question, the answer-processing engine retrieves the answers template that contains links to available answers. For example, the original question "Where can I buy the 56K external modem?" triggers the answers template for this question that contains answers to the following specific questions: "Where can I buy modems online?", "Where can I buy and sell modems via online auction?", "Where can I compare prices for computer peripherals?", etc.

*AltaVista* (URL 17) search engine includes two important features, search for images and translation between languages. AltaVista enables search for over 25 million images, audio and video files from both the Web and from several private collections. Images include photographs, graphics, clip art, and galleries. Search is done by specifying appropriate keywords, and AltaVista responds by presenting pictures from Web pages and/or private collections. When a

user finds the picture that fits his interest, it is in some cases possible to activate the search for other pictures with similar features, and thus increase the proportion of pictures belonging to the field of his or her interest. Similarity is based on visual characteristics such as dominant colors, shapes and textures.

AltaVista also enables automatic translation of the contents of discovered Web pages between several languages. Automatic translation of any peace of text between dozen pairs of languages can also be performed. This can even be done independently of the search using AltaVista translation service (URL 18) directly. This translation service enables search for foreign languages by (a) translation of English query to some target language, (b) accomplishing search with the translated terms, and (c) translating the resulting Web pages back to English. Translation is in fact done by AltaVista's partner *SYSTRAN* (URL 19), a leading translation software institution. SYSTRAN's software is widely used by US government and European Union (for translation of its internal communication). It was also used in the US-USSR Apollo-Soyouz space project during 1972-1975.

## 4. Database search

Growing number of databases from various areas are available by means of the Web. However, search engines often cannot distinguish between the simple Web page and the entrance to one or more huge databases. Information contained in databases are hidden and cannot be retrieved by search engines, hence the name "Invisible Web" for databases accessible from the Web. Databases typically include structured data about specific topics like companies, restaurants, or locations of ATM machines. Most database resources are free, while some are fee-based.

Recent study accomplished by the *BrightPlanet* company (URL 20) estimated that the inaccessible part of the Web is several hundred times larger than its visible part (accessible by search engines), and contains about 550 billion individual documents. More than 100,000 invisible Web sites exist, while 60 largest invisible Web sites contain about 10% of total size of invisible Web. More than half of the invisible Web content resides in topic specific databases, and about 95% of the invisible Web is publicly accessible information not subject to fees or subscriptions.

BrightPlanet also developed the *LexiBot* search tool (URL 21). This tool looks at the search request, then selects the most relevant 600 different invisible Web resources and forwards the query to them. The results are returned in the form similar to the one done by meta search engines. BrightPlanet plans to expand the query to about 20,000 invisible Web sites already collected, and finally to all 100,000 significant invisible Web sites.

Some of the best known Web directories containing addresses of Web sites used as entrance for different databases are InvisibleWeb, Lycos Searchable Databases and INFOMINE. *InvisibleWeb* (URL 22) is a directory with over 10,000 databases, archives, and search engines containing information that traditional search engines cannot access. This is a hierarchically structured directory that can also be searched by key words. An example search using the term "ATM" (done in October 1999) found 8 Web sites like Visa ATM locator, 4Banking.com, and MasterCard ATM locator. Each of these Web sites enable entrance to the database of ATM locators that can be retrieved geographically, with the final result of finding addresses of ATM locations in the specific town. Comparative search with AltaVista on the phrase "ATM" gave about 525,000 links, while search on "ATM" in the title gave some 22,000 links. The more precise search using the phrase "ATM locator" gave 3,400 links while search on "ATM locator" in the title gave some 350 links. Moreover, most of these links are not the entrances to databases. As can be seen, in search for specific information the advantage of databases directory over traditional search engine is evident.

*Lycos Searchable Databases* (URL 23) is a directory similar to InvisibleWeb, while *INFOMINE* (URL 24) is a huge scholar directory of more than 10,000 databases in biological, agricultural, medical and physical sciences, engineering, computing, mathematics, social sciences and humanities. INFOMINE also includes directories of electronic journals, elec-

tronic books, online library card catalogs and articles.

*Terraserver* (URL 25) is a unique huge database of high-resolution satellite imagery and aerial photography that started as a research project between Aerial Images, Inc., Microsoft, the USGS, and Compaq. Special interest of Microsoft was to test the ability of a database server built as a large scalable network (using Windows NT Server and Microsoft SQL Server) to store terabytes of data for thousands of simultaneous users, and to explore the usage of relational database technology as a general-purpose image repository. Terraserver currently has imagery from more than 60 countries with resolution down to 1 meter, and has a goal to completely cover the earth's surface. It is using imagery from various sources like the Russian Space Agency, the Indian Space Agency, as well as US companies using commercial satellites.

## 5. Scientific information retrieval

Web is increasingly becoming a distributed collection of scientific literature where scientific papers can be rather easily retrieved and quickly accessed (Lawrence and Giles, 1999a). Numerous researchers make their publications available on their homepages, and many traditional journals offer access to the whole text of papers on the Web. Some publishers allow their papers to be placed on the author's Web site, while others permit prepublication of articles on the Web. The main problem for search engines is in location of articles prepared in Postscript or PDF format, predominantly used formats for scientific publications.

Important step forward in enabling efficient and effective service for scientific literature on the Web was done by the NEC Research Institute team in developing the *ResearchIndex* (URL 26). ResearchIndex is a search engine specifically designed for scientists, based on the combination of digital library and citation index. Citation indices index the citations in a paper so that the paper is linked with the cited articles. They enable scientists to find papers that cite a given paper, and also facilitate evaluation of articles and authors. The main problem with traditional citation indices is their high price related to manual effort necessary for indexing.

In order to avoid this problem NEC Research Institute team developed a digital library of scientific publications that create a citation index autonomously using Autonomous Citation Indexing (ACI), a system that doesn't require any manual effort (Lawrence, Giles and Bollacker, 1999). ACI allows literature search using both citation links and the context of citations. It retrieves PDF and postscript files and uses simple rules based on formatting a document to extract the title, abstract, author and references of any scientific paper it finds. ACI enables fast feedback by indexing items such as conference proceedings or technical reports. New papers can be automatically located and indexed as soon as they are published on the Web or announced on mailing lists.

The same team built a prototype digital library called CiteSeer. CiteSteer downloads papers from the Web and converts them to text, parses the papers to extract the citations and the context in which the citation were made, and then stores these information in a database. CiteSeer includes full-text articles and citation indexing and allows location of papers by keyword search or citation links. NEC Research has made the CiteSeer software available at no cost for non-commercial use. A demonstration version of *CiteSeer* (URL 27) indexes over 200.000 computer scientific papers, more than the largest online scientific archives.

Retrieval of scientific data is a particularly complex problem, since it has to deal with a widely distributed, heterogeneous collections of data. Scientific data form huge and fast growing collections of data stored in specialized formats (e. g. astronomic or medical data), and their retrieval requires powerful search tools. For this purpose a consistent set of metadata semantics, as well as standard information retrieval protocol supporting the metadata semantic is required. One technology developed for scientific information retrieval is *Emerge* (URL 28), the system based on XML-based translation engine that can perform metadata mapping and query transformation.

## 6. Trends

Several new approaches to search engines mechanisms were developed around the idea of explo-

iting rich Web hyperlinking structure for clustering and ranking of Web documents. *Clever* project (Members of the Clever team, 1999), run by researchers from IBM, Cornell University and the University of California at Berkeley, analysis Web hyperlinks and automatically locates two types of pages: "authorities" and "hubs". "Authorities" are the best sources of information on a particular broad search topic, while "hubs" are collection of links to these authorities. A respective authority is a page that is referred to by many good hubs, while a useful hub is a page that points to many valuable authorities. For any query Clever first performs an ordinary text-based search using an engine like AltaVista, and takes a list of 200 resulting Web pages. This set of links is then expanded with Web pages linked to and from those 200 pages — this step is repeated to obtain a collection of about 3,000 pages. Clever system then analyses the interconnections between these documents, giving higher authority scores to those pages that are frequently cited, and higher hub scores to pages that link to those authorities. This procedure is repeated several times with iterative adjusting of authorities and hub scores: authority that has many high-scoring hubs pointing to it earns a higher authority score, while a hub that points to many high-scored authorities gets a higher hub score. The same page can be both the authority and the hub. A side-effect of this iterative processing is that the algorithm separates Web sites into clusters of similar sites. Clever system is also used for automatic compilation of lists of Web resources similar to subject trees — these lists appear to be competitive with handcrafted ones.

One extension of this research is the *Focused Crawling* system (Chakrabarti et al, 1999), a software for topical resource discovery. This system learns the specialization by examples and then explores the Web, guided by a relevance and popularity rating mechanism. The system selects Web pages at the data-acquisition level, so it can crawl to a greater depth (dozens of links away from the starting set of Web pages) when on the interesting trace. Focused Crawling will be used for automated building of high-quality collection of Web documents on specific topic.

An interesting approach to Web search, sometimes called 'surfing engine', is developed by *Alexa* service (URL 29). When an Alexa user navigates to a Web page, Alexa service retrieves data about this page and presents them to the user. These data (or metadata) includes information on who is behind the site the user is navigating to, how often is this site updated, how fast this site responds to requests, what is its popularity ranking, etc. All these information help the user to decide whether the site is what he is looking for. Alexa service also retrieves information from its servers to suggest to the user some other pages (Related Links) that might be of interest to him. To find Related Links Alexa service uses the paths of the collective Alexa community, information about clusters of similar Web sites (found by prior analysis), analysis of texts on individual Web pages, etc. To avoid the "Not Found" message Alexa service checks its archive to see whether it can find an archived version of the page.

Visual information retrieval becomes increasingly important as the amount of image and video information in different computing environments continues to grow at an extremely fast rate. This kind of retrieval is very specific and complex, and even formulating the query is not simple since text queries are unable to adequately express the query requirement, making the query processing inefficient. Readers interested in this field are recommended to read two introductory papers (Gupta and Jain, 1997; Chang et al., 1997).

Traditional search in the batch mode leads to the situations where users may miss recently- added Web pages because of the slow update of search engines data bases (crawlers often revisit the same Web site after one or two months). This led IBM to develop the *Fetuccino* (URL 30) software, an interesting combination of traditional search in batch mode and dynamic search in real time. In the first stage some traditional batch search engine is exploited. In the second stage results obtained in the first stage are refined via dynamic searching and crawling. This is done by feeding the search results from the first phase to the Mapuccino's dynamic mapping system which augments those results by dynamically crawling in directions where relevant information is found.

Natural language processing has an important role in developing advanced search systems because of the fundamental problem of presenting

accurate meaning of search request to the search system. Natural language includes a lot of ambiguities on various levels: morphological, syntactic, semantic, discourse and pragmatic levels, and natural language processing can be used in all stages of information retrieval processing (Feldman, 1999). Some of the research directions in this field are entity extraction, cross language retrieval, automatic information categorization, and summarization. Entity extraction may e. g. enable extraction of names of people, places, and chronological information from text, store them and enable retrieval of these information with natural language search requests. Summarization attempts to automatically reduce document text to its most relevant content based on the user requirements.

An example of research in use of natural language processing on search engines is a system that performs query expansion using an online lexical database WordNet (Moldovan and Mihalcea, 2000). This system first disambiguates the sense of the query words using the WordNet database, so that each keyword in the query is mapped into its corresponding semantic form as defined in WordNet. After that, WordNet is used for finding synonyms with the query semantic concepts that are subsequently used in Internet search. Documents found in the search are subjected to a new search using the operator which extracts only the paragraphs that provide relevant information to the query. The goal of the system is not to retrieve entire documents but to provide the user with answers, so it returns to the user the paragraphs that may contain the answer.

Software agents (bots) are used for various search activities (URL 31) like searching remote databases, multiple virtual bookstores, or searching and indexing Internet resources. One of the research projects from the MIT Media Lab is the *Expert Finder* project (URL 32) that helps in finding someone who is an expert on some area. The project is based on the assumption that each user has an agent that knows about his or her areas and levels of expertise. When necessary, the user asks the agent to find another user that is an expert on some area. The agent then goes out on the Internet and exchanges information with other expert finder agents, getting their users' profiles. After that, the agent presents to the user a list of people it thought may be able to help, with best candidates first. User can now exchange messages with the selected experts.

Intensive use of the XML markup language is expected to drastically improve information retrieval on the Web and make it more efficient. XML (eXtensible Markup Language) provides a standard system for browsers and other applications to recognize the type of data in documents. This enables search engines to search only certain fields in the documents, instead of the whole documents. This will lead to much faster and more precise search. However, in order to enable use of XML each industry will have to set up its own standards for document structures, Web site providers will have to tag the pages according to the standard structures, and search engines indexing applications will have to hold tag information as metadata.

Currently there are more than 400 million mobile phone users in the world, and it is estimated by Forrester Research that by 2004, 95% of all mobile users will be Internet enabled. Because of this, in the beginning of 2000 several search engines start offering search under the Wireless Application Protocol (WAP). *FAST* search engine thus launched its *WAP search engine* (URL 33) in February 2000, offering search of the index with more than 100,000 WML (Wireless Markup Language) documents, growing at a rate of several hundred percents per month. Besides general search, FAST is preparing search for location information (e. g. location of nearest hospitals or hotels), real-time alerts (e. g. about congested traffic in the area) and image and video streaming. *Google* search engine offers *WAP search* (URL 34) of both WML and HTML documents, since less than 1% of Web sites are available in WML. When a wireless user requests a HTML page, Google translates the requested HTML document on the fly into WML.

## 7. Conclusion

In the last few years significant improvements were made in a number of areas of the World Wide Web search. This paper first presented the current state of the World Wide Web, the speed of its development, as well as its heterogeneity. After that, the paper discussed some of the most

advanced recent approaches to the Web search, as well as various novel operational Web search elements or systems in the area of subject trees, search engines and database search.

There is certainly room for further advancements of both technology and services. One type of service that would be extremely helpful for professionals are *specialty search services* that would exclusively cover the domain of interest of specific groups of professionals. Coverage should be deep, and information should be fresh and almost without any dead links. *Academic search engine* is another highly important service that should cover academic Web sites. It should carry out a deep and very frequent crawl of these Web sites. Covering a fresh state of research and research publishing could e. g. lead to decrease duplication of research work.

## References


[1] R. BAEZA-YATES, B. RIBEIRO-NETO *Modern Information Retrieval*. Addison-Wesley, Harlow, England, 1999.

[2] A. BRODER ET AL. Graph structure in the web. *Proceedings of the* $9^{th}$ *International World Wide Web Conference* (2000) Amsterdam. (Online at `http://www9.org/w9cdrom/160/160.html`)

[3] S. CHAKRABARTI ET AL. Focused Crawling: a new approach to topic-specific Web resource discovery. *Proceedings of the* $8^{th}$ *International World Wide Web Conference* (1999) Toronto. (Online at `http:// www8.org/w8-papers/5a-search-query/crawling/index.html`)

[4] S. CHANG ET AL. Visual information retrieval from large distributed online repositories. *Communications of the ACM*, **40** (1997), No. 12, 63-71.

[5] J. CHO AND H. GARCIA-MOLINA Synchronizing a database to improve freshness. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data* (2000), Dallas, TX.

[6] V. CERIC (1997), World Wide Web search: techniques, tools and strategies. *Proceedings of the* $19^{th}$ *International Conference on Information Technology Interfaces ITI'2000* (2000), Pula, Croatia.

[7] S. FELDMAN NLP meets the jabberwocky: natural language processing in information retrieval. *ONLINE*, **23** (1999), No. 3. (Online at `http://www.onlineinc.com/onlinemag/OL1999/feldman5.html`)

[8] A. GUPTA AND R. JAIN Visual Information Retrieval. *Communications of the ACM*, **40** (1997), No. 5, 71-79.

[9] S. LAWRENCE AND C. L. GILES Accessibility of information on the web. *Nature*, **400** (1999), 107-109.

[10] S. LAWRENCE AND C. L. GILES Searching the Web: general and scientific information access. *IEEE Communications Magazine*, **37** (1999a), 116-122.

[11] S. LAWRENCE, C. L. GILES AND K. BOLLACKER Digital libraries and autonomous citation indexing. *IEEE Computer*, **32** (1999), 67-71.

[12] MEMBERS OF THE CLEVER TEAM Hypersearching the Web. *Scientific American*, June 1999. (Online at `http://www.sciam.com/1999/0699issue/0699raghavan.html`)

[13] D. I. MOLDOVAN AND R. MIHALCEA Improving the search on the Internet by using WordNet and lexical operators. *IEEE Internet Computing*, **4** (2000), No. 1.

[14] N. SHIVAKUMAR AND H. GARCIA-MOLINA Finding near-replicas on documents on the Web. In *Workshop on Web databases* (1998), Valencia, Spain.


## URLs

1. Hobbes' Internet Timeline, `http://www.isoc.org/zakon/Internet/History/HIT.html`.
2. Nua Ltd., `http://www.nua.ie/surveys/how_many_online/`
3. Cyveillance comp., `http://www.cyveillance.com/newsroom/pressr/`
4. Open Directory, `http://dmoz.org`
5. About.com, `http://about.com`
6. Britannica.com, `http://www.britannica.com`
7. Inxight: Hyperbolic Tree, `http://www.inxight.com/products/developer/hyperbolic_tree.html`
8. IBM: Mapuccino, `http://www.ibm.com/java/mapuccino/index.html`
9. Cartia: ThemeScape, `http://www.cartia.com/products/index.html`
10. FAST, `http://alltheweb.com`
11. Northern Light, `http://www.northernlight.com`
12. Google, `http://www.google.com`
13. Ditto.com, `http://www.ditto.com`
14. Simpli.com, `http://www.simpli.com`
15. Oingo, `http://www.oingo.com`
16. AskJeeves, `http://www.ask.com`
17. AltaVista, `http://www.altavista.com`
18. AltaVista translation service, `http://world.altavista.com`
19. SYSTRAN, `http://www.systransoft.com`
20. BrightPlanet, `http://www.completeplanet.com/Tutorials/DeepWeb/`

21. LexiBot, http://www.lexibot.com
22. InvisibleWeb, http://www.invisibleweb.com
23. Lycos Searchable Databases, http://dir.lycos.com/Reference/Searchable_Databases/
24. INFOMINE, http://infomine.ucr.edu/search.phtml
25. Terraserver, http://terraserver.com
26. Emerge, http://emerge.ncsa.uiuc.edu/
27. CiteSeer demonstration version, http://csindex.com
28. ResearchIndex, http://www.researchindex.com
29. Alexa, http://www.alexa.com
30. IBM: Fetuccino, http://www.ibm.com/java/fetuccino/index.html
31. Search agents, http://www.botspot.com/s-search.htm
32. Expert Finder project, http://www.media.mit.edu/~adriana/projects/EF/
33. FAST WAP search, http://wap.fast.no
34. Google WAP search, http://www.google.com/palm

*Received:* October, 2000
*Accepted:* November, 2000

*Contact address:*
Vlatko Ceric
Faculty of Economics
University of Zagreb
Kennedyjev trg 6
10000 Zagreb
Croatia
Phone: +385 1 2383280
Fax: +385 1 2335633
e-mail: vceric@efzg.hr

DR. VLATKO CERIC is a professor at the Faculty of Economics, University of Zagreb. His main research interests are simulation modelling, decision support systems, information retrieval, electronic commerce and operations management. He published over 80 papers and a few books, and was a leader of several research and application-oriented projects.