

# Trends for Web Information Processing over World Wide Web

Dr. Harmunish Taneja  
*M.M. Engineering College,  
M.M. University, Mullana, Ambala*

Dr. Kavita Taneja  
*M.M.I.C.T. & B.M.  
M.M. University, Mullana, Ambala*

**Abstract - Web information processing through modern search engines index zillions of web pages on distributed platforms of thousands of commodity web users. Much of the research has been done on the information processing aspects ranging from crawling, web graph topology, indexing, efficient query processing, caching and ranking. Despite all of the challenges, the expansion of the web has turned information processing over web into a key enabling technology. This paper summarises the major trends and evolution of information processing over World Wide Web. Also, it is emphasised that the object oriented design paradigm when applied to this field may greatly reduce the complexity of processing system while improving reusability and manageability.**

**Keywords: Web Information Processing, Object- Oriented Model, World Wide Web**

## I. INTRODUCTION

Information processing on web takes account of the structure, storage, analysis, searching, and retrieval of information. The primary function of current web search engines in this direction is to efficiently search for the query results at the document level. However, countless structured information about real-world objects is embedded in static web pages and online web databases [1]. Document-level information retrieval can unfortunately lead to irrelevant results in answering miscellaneous queries of diverse users. Section 2 presents the related work. Section 3 summarizes the trends and evolution of information processing in WWW. Object oriented model for web information processing is emphasized in Section 4. Finally the paper is concluded in Section 5.



## II. RELATED WORK




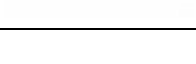





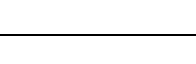
Traditionally the concept of search engine was conceived as quick information retrieval and processing of stored web data and was published way back in 1945 by an American engineer and science administrator, Vannevar Bush. The essay "As We May Think" [2] may have been written as early as 1936. The first concrete search engine was shaped in the 1960s by Gerard Salton at Cornell University as the "SMART information retrieval system" (Salton's Magic Automatic Retriever of Text) [3]. Also, Gerard Salton is marked as the father of modern search technology. But the first on the internet, search engine was called Archie, just Archive with the "v" removed to index FTP archives. The first web search engine was called Wandex released in 1993 and used an index created by the first web crawler, World Wide Web Wanderer, written in Perl by Matthew Gray at Massachusetts Institute of Technology. The need of instant communication was reflected in culture and commerce of common man in mid 1990's. The first full text search engine was launched as WebCrawler in 1994 that indexed entire web pages [4]. Google technology was originally called BackRub, a project Larry Page and Sergey Brin started working on in 1996. Yahoo and Microsoft didn't have their own search engine technology until 2004. Yahoo Search used data from Inktomi, AltaVista and was even powered by Google for some time [5]. Microsoft's MSN Search opted for other search engine results usage and launched their own technology in 2005 (beta in 2004). Also, with higher total number of searches worldwide, Baidu, the Chinese search engine, surpasses Microsoft's Live Search. The first web site put up was <http://info.cern.ch/> on August 6, 1991 [6]. It is estimated that in 1993 web carried only 1% of the information flowing through two way telecommunication, by 2000 this figure had grown to 51% and by 2007 more than 97 % of all telecommunicated information was carried over the internet and not to mention the percentage in 2011 [4].

## III. Web INFORMATION PROCESSING : TRENDS

The growth of WWW from email, text messaging, video calls to blogs, chat rooms, social networking and online shopping sites was accompanied by creation of new search engines supporting vivid information processing capabilities [7] as summarized in Table 1.

Table 1: History of Web Information Processing [8, 21]

Year	Search Engine	Logos'	Remarks
1990	Archie		<ul style="list-style-type: none"> <li>The first search engine by Alan Emtage.</li> <li>Combined the script-based data gatherer with a regular expression matcher for retrieving file names.</li> </ul>
1993	JumpStation		<ul style="list-style-type: none"> <li>Gathered information about the title and header from web pages.</li> <li>Employed linear search.</li> <li>No Ranking system.</li> </ul>
	World Wide Web Worm		<ul style="list-style-type: none"> <li>Indexed web page titles and URL's.</li> <li>No ranking system.</li> </ul>
1994	WebCrawler		<ul style="list-style-type: none"> <li>Meta Search Engine.</li> <li>First crawler to index entire web pages.</li> <li>Provided advertising free interface.</li> </ul>
1995	AltaVista		<ul style="list-style-type: none"> <li>First to allow natural language queries, advanced searching techniques and inbound link checking.</li> <li>Allowed users to add or delete their own URL within 24 hours.</li> </ul>
1996	Ask.com		<ul style="list-style-type: none"> <li>Launched as a natural language search engine.</li> <li>Rank results were based on their popularity.</li> <li>Commendable user interface.</li> </ul>
1997	Northern Light		<ul style="list-style-type: none"> <li>Created by David Seuss in Cambridge, Massachusetts, to sell custom search engine to corporations.</li> <li>Free commercial search engine</li> <li>Organized search results in specific folders labelled by subject.</li> </ul>
1998	Google		<ul style="list-style-type: none"> <li>Developed new approaches to relevance ranking to sort the best results first.</li> <li>Ranked pages using citation notation.</li> <li>Comprehensive web coverage.</li> <li>Provided number of vertical services.</li> <li>Basic layout and unrivalled usability made 'google' a synonym for 'web search.'</li> </ul>
1999	AlltheWeb		<ul style="list-style-type: none"> <li>Good user interface.</li> <li>Rich advanced search features resulting in relevant results.</li> <li>Comprehensive web coverage.</li> <li>Returned result from its own database and Yahoo database.</li> </ul>
1999	Baidu		<ul style="list-style-type: none"> <li>First Chinese Search engine.</li> <li>Surpassed Microsoft in terms of higher number of searches.</li> </ul>
2000	Teoma		<ul style="list-style-type: none"> <li>Employed clustering to organize sites by subject specific popularity.</li> </ul>

2000	Vivisimo		<ul style="list-style-type: none"> <li>• Web meta-search engine.</li> <li>• Dynamically clustered users' search results.</li> </ul>
2003	Objects Search		<ul style="list-style-type: none"> <li>• Allowed information, images, news, and videos search across the Web.</li> </ul>
2004	Yahoo! Search		<ul style="list-style-type: none"> <li>• Hybrid search engine.</li> <li>• Combined results from its own directories and crawler based results from Google.</li> </ul>
	MSN Search		<ul style="list-style-type: none"> <li>• Hybrid search engine.</li> <li>• Combined results from Inktomi and Looksmart databases with own handpicked directory of websites.</li> <li>• Relevant Results.</li> </ul>
2005	Quaero		<ul style="list-style-type: none"> <li>• Quaero is often cited as a European competitor to Google, Yahoo, and Bing.</li> <li>• Not intended to be a text-based search engine but is mainly meant for multimedia search.</li> <li>• Designed to recognize, transcribe, index, and automatic translate audiovisual documents and operate in several languages.</li> </ul>
2006	Trumalia		<ul style="list-style-type: none"> <li>• Big and fast growing index ensures relatively comprehensive and relevant search results.</li> <li>• Displayed the work of talented contemporary artists.</li> <li>• Featured a series of challenging enigmas.</li> </ul>
2009	Bing		<ul style="list-style-type: none"> <li>• Formerly Live Search, Windows Live Search, and MSN Search.</li> <li>• Supported additional interface features, media features (video thumbnail, video and image search)</li> <li>• Supported integration with Hotmail and Facebook and also provide local information.</li> </ul>
2010	Blekkio		<ul style="list-style-type: none"> <li>• Used a set of short community-created conventions for attributing information to provide results for common searches.</li> <li>• Also offered a downloadable search bar.</li> </ul>
2011	Yandex		<ul style="list-style-type: none"> <li>• Launched in Turkey in September 2011, with its services, including web search, maps and email, tailored specifically to the needs of local web users.</li> </ul>
2012	Volunia		<ul style="list-style-type: none"> <li>• An Italian web search engine or social search engine.</li> <li>• It crawls the web, indexes websites and builds the ranking using the comments and opinions of other users.</li> </ul>

#### IV. OBJECT ORIENTED MODELS FOR WEB INFORMATION PROCESSING

It is realized that web users usually search for information of a certain 'object', rather than a web page containing the query terms. Object identification on the Web [9] has been developed in recent years. PopRank [10] is a method which considers both the web popularity of an object and the object relationships for object oriented information processing to compute the popularity score of the web object. PopRank extends the PageRank model by adding a popularity propagation factor (PPF) to each link pointing to an object, and uses different propagation factors for links of different types of relationships. Large combinations of feasible factors are required to get a reasonable

assignment of the propagation factor. The existing web information retrieval (IR) techniques cannot provide satisfactory solution to the web object extraction task highly heterogeneous diverse web sources.

Wrapper deduction [11, 12], web database schema matching [13, 14] made it possible to extract and integrate all the related web information about the same object together as an information unit. PageRank technique [15] calculates the importance of a web page based on the scores of the pages pointing to the page. Hence, importance of web pages pointed by many high quality pages rises. PageRank and HITS algorithms [16] are special cases of the unified link analysis framework but all the links have the same authority propagation factors in the PageRank model; it could not be explicitly applied to object-level ranking problem and Extended Hyper Text Markup Language (XML) elements. Ranked Keyword search over XML Documents (XRANK) [17] rank XML elements using the link structure of the database. Object Rank system [18, 19] applies the random walk model to keyword search in databases modelled as labelled graphs. A unified link analysis framework [20] called "link fusion" considers both inter and intra type link structures among multi type inter related data objects for searching.

## V. CONCLUSION

The most distinguishing requirement of today's search result through information processing systems for complex heterogeneous web is to dynamically adapt to vigorously changing web contents and demands. Object oriented approach may improve performance for non-conventional web search that handle outsized volumes of web. The new algorithmic methodologies to efficiently and effectively process information are growing research areas. Object oriented information processing collect information for objects relevant for a specific application domain and rank objects in terms of their relevance and attractiveness to respond user queries. But still improvement in ranking and user interface supported by modern information processing tools and web applications is an open research area.

## REFERENCES

- [1] Lawrence S., and Giles C.: "Searching the World Wide Web", in the Journal of Science, Volume 280, Number 5360, pp. 98-100, pp. 8-100, 1998.
- [2] Bush, V.: "As we may think", in Atlantic Monthly, Volume 176, Issue 1, pp. 101-108, 1945.
- [3] Salton G., and McGill M.: "Introduction to Modern Information Retrieval", by McGraw-Hill, New York, 1983.
- [4] <http://www.searchenginehistory.com/>
- [5] Battelle: "The Search: How Google and its Rivals Rewrote the Rules of Business and Transformed Our Culture" 1<sup>st</sup> Edition, Portfolio Hardcover 2005.
- [6] <http://info.cern.ch>
- [7] Xu J. L.: "Internet search engines: Real world information retrieval issues and challenges", in the Proceedings of Conference on Information and Knowledge Management, Kansas City, Missouri, November 1999.
- [8] Laure B. Kohen: "The world of Search Engines", Internet Tutorials 2011 [Online], available at, <http://www.internettutorials.net/world-of-search-engines.asp>, 13 Dec. 2011.
- [9] Sheila Tejada, Craig A. Knoblock, and Steven Minton: "Learning domain-independent string transformation weights for high accuracy object identification", in the Proceedings of the 8<sup>th</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 350-359, 2002.
- [10] Zaiqing Nie, Zhang Y., Wen J., Ma W.: "Object-level ranking: Bringing order to web objects", in the Proceedings of the 14<sup>th</sup> international conference on World Wide Web, pp. 567-574, Chiba, Japan, 2005.
- [11] Ashish N., and Knoblock C.: "Wrapper generation for semi-structured internet sources", in the Proceedings of Workshop on Management of Semi structured Data, Tucson, and SIGMOD Record, Volume 26, Number 4, pp. 8-15, 1997.
- [12] Nickolas Kushmerick, Daniel S. Weld, and Robert B. Doorenbos: "Wrapper induction for information extraction", in the Proceedings of International Joint Conference on Artificial Intelligence, pp. 729-737, 1997.
- [13] Bin He, Kevin Chen, Chuan Chang, and Jiawei Han: "Discovering complex matchings across web query interfaces: a correlation mining approach" in the Proceedings of ACM SIGKDD International Conference Knowledge Discovery and Data Mining, Seattle, Washington, USA, 2004.
- [14] Jiying Wang, Ji-Rong Wen, Frederick H. Lochovsky, and Wei-Ying Ma: "Instance-based schema matching for web databases by domain-specific query probing", in the Proceedings of 30<sup>th</sup> Very Large Data Bases (VLDB) Conference, Toronto, Canada, pp. 408-419, 2004.
- [15] Page L., Brin S., Motwani R., and Winograd T.: "The pagerank citation ranking: Bringing order to the web", in the Technical report, Stanford Digital Libraries, pp. 1-17, 1998.
- [16] Kleinberg J. M.: "Authoritative sources in a hyperlinked environment", in the Journal of the ACM, Volume 46, Number 5, pp. 604

-632, 1999.

- [17] Guo L., Shao F., Botev C., and Shanmugasundaram J.: "Xrank: Ranked keyword search over XML documents", in the Proceedings of ACM SIGMOD International Conference on Management of data, San Diego, CA, pp. 16-27, 2003.
- [18] Lafferty J., McCallum A., and Pereira F.: "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", in the Proceedings of the 11<sup>th</sup> International Conference on Machine Learning (ICML), Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 282-289, 2001.
- [19] Andrey B., Vagelis H., and Yannis P.: "Authority-based keyword queries in databases", in the Proceedings 30<sup>th</sup> Very Large Data Bases (VLDB) Conference, Toronto, Canada, 2004.
- [20] Xi W., Zhang B., Chen Z., Lu Y., Yan S., Ma W.Y., and Fox E.A.: "Link fusion: A unified link analysis framework for multi-type interrelated data objects", in the Proceedings of the 13<sup>th</sup> International Conference on World Wide Web, pp. 319 - 327, 2004.
- [21] [http://en.wikipedia.org/wiki/Web\\_search\\_engine](http://en.wikipedia.org/wiki/Web_search_engine)