ISSN (Online): 2395- 6593

AN INTERNATIONAL JOURNAL OF MULTIDISIPLINARY RESEARCH

The Impact of Web Technologies on Information Retrieval Systems: A Review

Akib Ahmed
Department of Library
Government Degree College Kangan
Email: akibahmed1@gmail.com

Abstract
There is little research that focuses on the impact and applications of web technologies on online Information Retrieval Systems. The main contribution of this work/paper is to highlight/explore the technological imprints on online information retrieval mechanism. The author confined the study to the online information retrieval systems, their use, impact, evaluation and response of user community. The empowerment of the web user with web technologies has led to the exponential growth of data, information and knowledge and web has also tailored a novel way to seek information. With the rapid exponential growth of information resources, there is a need to logically categories this information and knowledge so it can be fully utilized by all. Prominent evaluation techniques of information retrieval systems viz-a-viz vector space model and precision and recall ratio facilitated measures to compute retrieval efficiency.

Keywords: Web technologies, Information retrieval, Ontology, Web generations, Precision and Recall, vector space model

Web Technologies:

A web technology has its impact on almost every field and Information Centers are no exception to adapt to the technological advancements time to time. Web technologies have witnessed many changes since its inception. Web 1.0, Web 2.0, Web 3.0, Web 4.0 and Web 5.0 are different versions of Web with each succeeding version adding amazing possibilities in information retrieval mechanism. The task of a full-text information retrieval system is to satisfy a user's information need by identifying the documents in a collection of documents that contain the desired

information. In information retrieval, the main concern is to retrieve a set of documents that is semantically related to a given user query. Semantic similarity is a concept whereby a set of terms within term lists are assigned a metric based on the likeness of their meaning. Measuring the semantic similarity between words is an important component of information retrieval mechanism on the web such as relation extraction, data mining, automatic meta-data extraction etc (Pushpa, 2013). Due to extensive research, the applications of information retrieval became diverse; they include but are not limited to document indexing and archiving, information extraction, knowledge storage, document classification and clustering, question-answering, speech retrieval, and web searching. (Bassil, 2012). With the exponential growth of information, to return relevant results to user query is a common challenge that all retrieval mechanisms face. Major scientific primary publishers, such as Elsevier, Wiley, Springer, etc. all have their own search and retrieval platforms in addition to participating in the search and retrieval systems of others by linking and other agreements. Their articles are likely searchable from their own platform, from various secondary indexes, and by major search engines such as Google with links back to their own repository of articles (Tenopir,2008). (Brophy & Bawden, 2005) also showed that database searching yield a higher percentage of good quality documents (84%) than Google (52%). To maintain the retrieval efficiency of search engines prominent evaluation techniques have been used by researchers over the years viz-a-viz standard Cranfield/TREC model of information retrieval system evaluation. The history of evaluating IR systems, roots in the practices of documentation, and especially of science librarianship. Bradford compiled subject bibliographies, primarily on request of a scientist or a group of scientists. The goal of such bibliographies was to identify all of the documents pertaining to the subject, and to not include in the bibliography any documents which did not pertain to the subject. The phrase "pertaining to the subject" as meaning (eventually) "relevant to the inquirer's query", making relevance of a document the basic criterion of evaluation, and therefore leading to the measures of recall and precision, emulating the "all and only" of the subject bibliography. (Belkin, 2010)

Generations of Web based technologies/ Web Generations:

Web 1.0

Web 1.0 refers to static web pages connected with hyperlinks. It is the initial phase of web called "read -only" web whereby not many people were able to put their content online and only few others could exploit the same.

Web 2.0

Web 2.0 or the "read-write" web emphasis on user generated content by using dynamic web pages, which are precisely user/access oriented. There is more sense of ease and accessibility. Web 2.0" was reportedly first conceptualized and made popular by Tim O'Reilly and Dale Dougherty of O'Reilly Media in 2004 to describe the trends and business models that survived the technology sector market crash of the 1990s (O'Reilly, 2005). Web 2.0 refers to the social use of the Web which allows people to collaborate, to get actively involved in

ISSN (Online): 2395- 6593

creating content, to generate knowledge and to share information online. (Grosseck G, 2009). In Contrast to passive access in Web 1.0 platform, Web 2.0 offers interactive access to wide range of user friendly technologies. The library and Information centres felt the need to utilize the number of avenues created by the Web 2.0 interoperable tools and be a part of major technological change. The Web 2.0 has facilitated libraries sharing of information at reduced/economical costs especially in reference management and information retrieval. As projected by (O'Reilly, 2005) Web 2.0 has transformed the way people interact in online environment with each other. The development of Web 2.0 (social web) has revolutionized the interaction patterns among web users.

Web 3.0

The Semantic Web is a Web of actionable information—information derived from data through a semantic theory for interpreting the symbols. The semantic theory provides an account of "meaning" in which the logical connection of terms establishes interoperability between systems. (Nigel, 2006). Web 3.0 refers to the use of emerging technologies such as Resource Description Framework (RDF), Web Ontology Language (OWL), Cloud computing, semantic web to facilitate the development, organization and sharing of user-generated web content through seamless collaboration among all users. Semantic markups allow machines to understand and link information available on the Web.

Web 4.0

Web 4.0 is still in its initial phase of development with no exact definition. It is believed to work like humans do with intelligence. The dream behind the symbiotic web (Web 4.0) is interaction between humans and machines in symbiosis. Although there is no exact idea about web 4.0 and its technologies, but it is obvious that the web is moving toward using artificial intelligence to become as an intelligent web. **(Sareh, 2012)**

Web 5.0

This phase is an idea in progress and definition is yet unknown. Web 5.0 is believed to be a sensory emotive space where we will be able to move the web from an emotionally flat environment to a space of rich interactions. (Kambil, 2008)

Implications of Web

The Web has radically emerged and hence changed the way people used to communicate/interact at different levels in pre-web era. Over the decades of development, the prosperity of any nation/unit depends upon the availability of information; it is to say that web has a role to play in transforming the world to information society. The more access to information leads to more opportunities to lead the Information-driven society. With the evolution of Information Society, the urge to organise the abundant information available is ever increasing. The Online

Information Retrieval system has offered a way to address such issues with much flexibility and professionalism.

Information Retrieval Systems:

Most existing information retrieval systems are based, either directly or indirectly, on models of the
traditional information retrieval (IR) system. These retrieval models specify how to create representation for textual documents, and how these representations and information needs should be
compared with each others in order to estimate the likelihood that a document will be judged relevant.
The estimates of the relevance of documents to a given query are the basis for the document rankings
that are now a familiar part of IR systems (Yang, 2012). The exponential growth of digital/electronic information poses a challenge of providing access to desired content or relevant resources. The core technique to meet such needs is Vector Space Model coupled with term weighing scheme that weights terms in a document according to their significance with respect to the context in which they appear. (Bassil, 2012). Most information retrieval systems are easily accessible even by a neophyte but then the relevance of documents retrieved ultimately determines the efficiency of information retrieval system. An efficiency/ efficacy of any Information Retrieval System depend upon the versatility of its index. The index terms describes document content by using content analysis techniques/methods. The aim if IR system is to retrieve high recall searches in order to return more relevant documents for each query. The better the index is the more will be its precision ratio among the overall recall against any query/search formulated by the user.

Evaluation of IR systems:

As the number of electronic resources grows it is crucial to profit from powerful tools to index and retrieve documents efficiently. Information retrieval is a science and practice of storing documents and retrieving information from billions of sites and the number is ever increasing. With the exponential growth of digital information, IR systems return bulk of results for any query formulated by the information seeker. To return relevant results, evaluation techniques have been adopted since long. (Voorhees, 2002) broadly classifies evaluation of IR systems into two divisions, system evaluation and user based evaluation. The user based evaluation is preferred over system evaluation as the goal of IR system is to meet the user expectations in terms of information needs, which can be judged by the user more precisely. However the user based evaluation is very expensive. The system evaluation works on ranking of documents in order of their relevance.

Vector Space Model:

Vector space model is one of the most popular statistical retrieval models since it uses statistical information to determine the relevance between the document and the query. In this model, the document is represented as a vector of keywords from

the respective document. The corresponding weights for each keyword determine its importance in the document and also in the collection. Similarly, the query is also a vector representation of keywords in the query and also has corresponding weights denoting the importance of the respective keywords in the query. Katta, 2009). The vector space model considers a distribution (variable) as a vector. Between two vectors one can compute a cosine that varies from 0 to 1. This measure was used by Salton in information retrieval field. The VSM allows us to describe any object with many characteristics in a gradual form. Thus, a vector is characterized by the frequency of occurrences that it has in the characteristics. Priego, 2003).

Precision and Recall

Both precision and recall are based on measure of relevance. Precision can be seen as a measure of exactness or quality, whereas recall is a measure of completeness or quantity. In simple terms, high precision means that an algorithm returned substantially more relevant results than irrelevant ones, while high recall means that an algorithm returned most of the relevant results. Precision and Recall are standard metrics expressing the quality of information retrieval methods. These measures are commonly used to check the retrieval efficiency of information retrieval system. They are based on the comparison of an expected result and the effective result of the evaluated system. These results are considered as a set of items, e.g., the documents to be retrieved. In information retrieval, precision measures the ratio of relevant results returned against users query (true positives) over the total number of returned results (true positives and false positives). In logical terms this determines the efficiency/correctness of information retrieval system. Precision is normally a measure of accuracy i.e., the returned results need to be correct. Recall measures the ratio of relevant results returned against users query (true positives) over the total number of expected results to be returned (true positives and true negatives). (Euzenat, 2007). The most widely used measure is the relevance-based measure of **recall** and **precision**. With respect to a given query, the entire space of documents can be partitioned into four sets: Relevant to the user and retrieved by the system; relevant but not retrieved; irrelevant and retrieved; irrelevant and not retrieved. The recall and precision are defined based on these four sets.

Precision= number of retrieved relevant documents/ total number of retrieved documents

Recall= number of retrieved relevant documents/ total number of relevant documents

Ontology:

For better understanding the meaning of data, ontology is being used, which is the one of the major components of semantic web and knowledge representation. Ontology is the most intelligent way of describing a domain, which can be shared, visualized and understood easily. It gives us the freedom to search for any topic after specifying proper domain. (Mukhopadhyay, 2011)The ambiguity of words poses a problem of returning relevant results to a

users query. A mechanism/concept that addresses such issues in web environment is web ontology. (Stein, 2011) concentrates on how the effectiveness of standard information retrieval systems can be enhanced with semantic technologies like ontologies. Ontologies are knowledge models that can represent knowledge of any universe of discourse by describing how concepts of a domain are related. The basic idea is that these ontologies can be used to tackle the problem of ambiguous words and hence improve the retrieval effectiveness.

Conclusion:

Online information retrieval systems have gained importance owing to advanced technologies and increasing metrics of access to these technologies over the years. The users/consumers of information have increased manifold due to ease of access and web interface availability. However access to quality information/ relevant documents can be determined by applying quality evaluation measures. Evaluation techniques particularly vector space model and precision and recall ratio have further assisted in computing retrieval efficiency of information retrieval systems. Since the inception of web, there has been continuous rapid flow of information from all ends, hence creating a tricky situation for indexers and one such issue is ambiguity of words that define documents, metadata. A measure/technique called web ontology has semantically reduced the ambiguity of words, hence increasing retrieval efficiency. The future information retrieval system is believed to be more user friendly in terms of access and better/ refined evaluation measures.

References:

Bassil, Y., & Semaan, P. (February 01, 2012). Semantic-sensitive web information retrieval model for HTML documents. European Journal of Scientific Research, 69, 4, 550-559.

Bassil, Youssef. (2012). Hybrid Information Retrieval Model For Web Images. International Journal of Computer Science & Emerging Technologies, 3, 1, 23-31

Belkin, N. "On the evaluation of interactive information retrieval systems", Rutgers University Community Repository, 2010. DOI: http://dx.doi.org/doi:10.7282/T3SF2TJK

Brophy, J., & Bawden, D. (January 01, 2005). Is Google enough? Comparison of an internet search engine with academic library resources. Aslib Proceedings, 57, 6, 498-512.

Euzenat, Jerome (2007) Semantic precision and recall for ontology alignment evaluation. Proc. 20th

International Joint Conference on Artifcial Intelligence (IJCAI), Jan 2007, Hyderabad, India. AAAI Press, pp.348-353, 2007.

Grosseck, G. (June 30, 2009). To use or not to use web 2.0 in higher education?. Procedia - Social and Behavioral Sciences, 1, 1, 478-482.

Kambil, A. (October 31, 2008). What is your Web 5.0 strategy?. *Journal of Business Strategy*, *29*, 6, 56-58.

Katta, Deepthi (2009) "A study of relevance feedback in vector space model" (2009). UNLV Theses, Dissertations, Professional Papers, and Capstones. Paper 1123.

Mukhopadhyay, Debajyoti, Chakrabarti, Chandrima, & Chakravorty, Sounak. (2011). A New Semantic Web Approach for Constructing, Searching and Modifying Ontology Dynamically.

O'Reilly,T(2005). What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. Communications and Strategies, 65, 17-38.

Priego, J. L. O. (October 01, 2003). A Vector Space Model as a methodological approach to the Triple Helix dimensionality: A comparative study of Biology and Biomedicine Centres of two European National Research Councils from a Webometric view. *Scientometrics : an International Journal for All Quantitative Aspects of the Science of Science, Communication in Science and Science Policy,58,* 2, 429-443.

Pushpa C N, Thriveni J, Venugopal K R, & L M Patnaik. (February 01, 2013). WEB SEARCH ENGINE BASED SEMANTIC SIMILARITY MEASURE BETWEEN WORDS USING PATTERN RETRIEVAL ALGORITHM. *Computer Science & Information Technology, 3,* 1, 1-11.

Sareh Aghaei, Mohammad Ali Nematbakhsh, & Hadi Khosravi Farsani. (February 01, 2012). EVOLUTION OF THE WORLD WIDE WEB: FROM WEB 1.0 TO WEB 4.0. *International Journal of Web & Semantic Technology, 3,* 1, 1-10.

Shadbolt, N., Berners-Lee, T., & Hall, W. (May 01, 2006). The Semantic Web Revisited. *Ieee Intelligent Systems, 21,* 3, 96-101.

Tenopir, Carol, "Online Systems for Information Access and Retrieval" (2008). School of Information Sciences -- Faculty Publications and Other Works. http://trace.tennessee.edu/utk_infoscie pubs/40

Stein L, Tomassen. (2011). *Conceptual Ontology Enrichment for Web Information Retrieval* (thesis). Norwegian University of Science and Technology, Norway

Voorhees, E. M. (January 01, 2002). The Philosophy of Information Retrieval Evaluation. *Lecture Notes in Computer Science,* 2406, 355-370.

Yang, C.-Y., & Wu, S.-J. (April 01, 2012). Semantic web information retrieval based on the wordnet. International Journal of Digital Content Technology and Its Applications, 6,6, 294-302.