International Journal of Latest Trends in Engineering and Technology
Vol.(8)Issue(1), pp.192-196
DOI: http://dx.doi.org/10.21172/1.81.025
e-ISSN:2278-621X

LINK ANALYSIS: EVIDENCES, RELEVANCE MEASURES, TYPES OF WEB SEARCH AND QUERIES USING HYPERTEXT RETRIEVAL

Hemangini S. Patel[1] and Apurva A. Desai[2]

Abstract-Information Retrieval is a vast field of research and evaluation is done in this area of research is rapid. With evaluation in the notion of relevance, it turned out to be a difficult concept to employ. Categories of web search are also evaluated and types of queries are also varies. So, this paper indicates the importance of contextual information other than text on page such as document former, anchor text for better performance via hypertext retrieval as evidences.

Keywords – Information Retrieval, Relevance, Precision, Recall, Web search, Link Analysis

I. INTRODUCTION

Web information is heterogeneous and scattered on the Web and has been growing rapidly to become widely accessible and publish. Due to these expansions looking for the range of related information is an extremely difficult as well as rigid task. Search engines serve important role in this retrieval task via indexing a Web, however they are not satisfy the user's information need because of requests of user contains only 2 or 3 terms. As few terms sent to search tools frequently directs towards noisy links in the reply. This is an outcome of that not using of the documents contextual information in the indexing stage. The documents context is represented as the current document is suggested via hyperlinks, semantic network, or adjacent text in retrieval process. To enhance the documents local index via information mined from its neighbours the context is used. Experiments done by Chibane and Doan [1] showed improvement in precision by using this context information for certain types of queries.

Users are concerns in web pages and consider context information of web pages as neighbourhood information which comes from the hyperlink information of page and directly concerned with web page. According to recent researches done in the area of information retrieval it is noticed that to enrich the performance of web search the utilization of link structure analysis provides essential information. Most of the ranking systems combine link and content information for better quality results in ranking. The two most popular elementary algorithms such as PageRank algorithm of Google [2 ,3] and HITs algorithm of Kleinberg's [4] are make use of the hyperlink structure amongst the Web page. Various expansions of link analysis

[1] *Bhagwan Mahavir College of Computer Application (BCA) Bharthana,Vesu, Surat, Gujarat, India*
[2] *Department of Computer Science, Veer Narmad South Gujarat University, Surat, Gujarat, India*

algorithms, via Web pages context defined as enhanced neighbourhood information suggested by hyperlinks and their weight is calculated based on the query terms sent by the end user.

II. Evidences Used For Improving Retrieval Performance

Lee and Croft [5] have studied various categories of information for getting better retrieval performance allied with web documents like document former, anchor text etc.

Document former: Previous findings, such as PageRank [2,3] and HITs [4], recommended that the web's hyperlink structure gives vital information and be able to efficiently published as document formers. Further important factors for retrieval performance of web page are page design and text within page along with link analysis that forming the web documents. All the links such as in-links, out-links are part of document in link analysis.

Anchor text: Eiron and McCurley [6] argued that anchor text akin to real queries in relation to term allocation and length, in this manner it links the information gap among query and document depictions. Previous findings [7, 8, 9, 10] specified with the aim of the amalgamation of anchor text be capable of entry site discovering and particularly valuable for operations like clustering and labelling process. Dou et al. [11] considered the relationship between target sites and anchor texts, and the anchor from related websites should be high weighted than the unrelated ones. Research done by Koolen [12] confirmed to further the utilization of anchor text be able to effective for the presently leading test collection of TREC ClueWeb09. Metzler et al. [13] demonstrated that exploration effectiveness closely bases on the collection of anchors, and specified models related to anchor aggregation be able to tackle the difficulty of link sparsity.

Song et al. [14] showed that they are the foremost to use web page titles for element mining in an unconfirmed manner since web page titles review the key notions of the content of web page and valuable for entity extraction from web page.

III. Relevence

Relevance is a vital notion in the area of information retrieval. Some of explanations about relevance are given by Koolen [12] what makes a document relevant are topical relevance, user relevance, Utility, Pertinence etc.

Topical relevance related to the subject awareness of relevance, which represents relevance as the relation among the subject content of the information requested by user and the existing subject awareness. This is also strongly related to the notion of aboutness. *User relevance* tracks from the user context. User relevance relates to background of user state. *Pertinence* is the relation between the subject knowledge of documents and the underlying information need. The information need involves the knowledge state of the user, which the system has no access and can only guess at. A document can only be relevant if the user can understand the content and if it contains information that changes the knowledge state of the user. *Utility* holds all notion relating not only topic-relatedness but also eminence, innovation, significance, reliability and many former things. The *situational relevance* means user's individual situation and personal view. It involves the problem at hand. It is inferred from criteria such as "usefulness in decision making, appropriateness of information in resolution of a problem, reduction of uncertainty, and the like".

In Web-centric search tasks, the assumed user model is of a user first trying to locate the entry page to a particular Web site and use the links on this entry page to navigate to pages that satisfy the information need. The entry page itself might not contain the information to satisfy the user,

but gives access to the rest of the site and allows the user to browse, representing a first step in a longer session. The relevance of entry pages is based not only on topical relevance, but also on user relevance, utility or situational relevance and pertinence. The traditional ad hoc retrieval methodology of TREC treats each search result as an individual document and assumes the user only wants pages that contain the information that satisfies information need.

IV. EVALUATION IN INFORMATION RETRIEVAL

The existence and utilization of structure of hyperlinks, is useful for information retrieval is often reveal through the evaluation using various available test collections or user developed test collections. This methodology uses a collection of documents, the information needs or requests of users and relevance judgements representing which documents in the group are pertinent to which information request. If we want to know whether or not link information is useful for IR, a baseline retrieval system that uses no link information the set of information requests, in the form of queries, is processed by them, and the returned results are re-ranked via linked based systems for the relevance judgements, after which scores are produced indicating how well they performed at finding the right documents. This allows us to compare the performance of the two systems without and with link analysis. If we are interested in knowing whether links can help finding more relevant documents, we can measure the recall (the fraction of all relevant documents that are retrieved) of via search tools and re-ranking after result retrieval. If the user is interested in the impact of link information on precision (the fraction of founded documents that are pertinent), the measure the number of pertinent documents in, for instance, the first 10 results.

There are many aspects of performance can be computed, but it is important to understand what should be determining by user. What is mainly significant for the user? Does the user desire as maximum relevant documents as probable, or to rapidly discover at least one relevant document that contains the required information? This depends on the particular context in which the user is using the retrieval system [12].

A. Web Search Tasks–

Users are searching the web for various tasks and purposes. Some challenges are like pages relevant to topic but doesn't contain the word via hyperlink analysis, quality of results and TREC topics. Broder et al. [15] Describes the types of queries such as navigational i.e. user is interested in specific web page or web site, informational i.e. user is interested on one or more pages on topic and transactional i.e. buying something, downloading or communicating etc.

One of the main task of navigational and transactional queries sent to Web search engines—are best supported using not only on-page text, but also link analysis, anchor text, click-through data and semantic analysis, among others.Types of web search formed by TREC are as Topic distillation, online service discovering, Home page discovering, Named-page discovering etc.

B. Test Collections–

Test collections may be static, user defined or topic related collections. Relevance can be approximated by topical similarity for these test collections.

C. Effectiveness measures–

The measure of success is often expressed in terms of precision and recall for the web documents. Precision is the part of documents discovered that are related. Recall is the part of related documents that are discovered.

Precision at rank n (P@n) in IR, precision is defined as the fraction of results that are classified correctly. Precision is a set-based measure. In IR, where results are typically in the form of a ranked list, precision is measured over a set of documents up to a certain rank. In the case of a results list of n retrieved documents, the precision over all documents up to that rank n is the part of documents that are evaluated as related [12]. Users are interested in top 10 or 20 results so precision is generally calculated at P@10 or P@20 for better results. Mean Average Precision (MAP) the average precision conveys the average precision at all of the ranks of the related documents. Mean Reciprocal Rank (MRR) the reciprocal rank conveys how faraway a user has downwards the results list to discover the foremost related document.

V. CONCLUSION

Contextual information is very helpful in finding relevance of web pages. So the link analysis is important for discovering relative information using hypertext and some external evidences and effectively improves the performance. Various types of relevance are discussed and depend according to users. Types of queries are also varies according to user. Two major performance measures are discussed precision and recall for related documents. Web search tasks are also evaluating like named page finding, entry page finding, ad-hoc search task etc. Test collections are also important for retrieval process.

REFERENCES

[1] I. Chibane, and B.L. Doan, "August. Impact of Contextual Information for Hypertext Documents Retrieval," In *Held in conjunction with the 6 th International and Interdisciplinary Conference on Modeling and Using Context*. 60, 2007.

[2] S. Brin, and L. Page, "The anatomy of a large-scale hyper textual Web search engine," *Computer Networks and ISDN Systems*, 30(1–7): 107–117, 1998.

[3] L. Page, S. Brin, R. Motwani, and T. Winograd., "The page rank citation ranking: Bringing order to the web," 1998.

[4] J. M. Kleinberg., "Authoritative sources in a hyperlinked environment," Journal of the ACM. 46(5) 604–632, 1999.

[5] C. J. Lee, and W. B. Croft, " Incorporating social anchors for ad hoc retrieval," In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval* 181-188, 2013.

[6] N. Eiron, and K. S. McCurley, "Analysis of anchor text for web search," In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, 459-460. ACM, 2003.

[7] N. Craswell, D. Hawking, and S. Robertson, " Effective site finding using link anchor information" In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 250-257. ACM, 2001.

[8] T. Westerveld, W. Kraaij, and D. Hiemstra, "Retrieving web pages using content, links, urls and anchors," In Tenth Text REtrieval Conference, TREC '02, 663-672, 2002.

[9] F. A. Omara, M. Amoon, N. A. El-Fishawy, and S. El-kazaz, " Analysing anchor links to enhance the web snippet clustering technique", In *Informatics and Systems (INFOS), 2012 8th International Conference on*, IEEE. SE-7 – SE – 11, 2012.

[10] Y. Zhang, and K. Lei, "Using anchor text refined by page importance to improve web retrieval", In *Computer Science & Education (ICCSE), 2012 7th International Conference on*, IEEE, 1200-1203, 2012.

[11] Z. Dou, Song, R., J. Y. Nie, and J. R. Wen,, " Using anchor texts with their hyperlink structure for web search," In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, ACM, 227-234, 2009.

[12] M. H. A. Koolen, "The meaning of structure: the value of link evidence for information retrieval," IR Publications, 2011.

[13] D. Metzler, J. Novak,, H. Cui, and S.Reddy, "Building enriched document representations using aggregated anchor text," In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, ACM, 219-226, 2009.

[14] W. Song, S. Zhao, C. Zhang, H. Wu, H. Wang, L. Liu, and H. Wang, "Exploiting Collective Hidden Structures in Webpage Titles for Open Domain Entity Extraction," In *Proceedings of the 24th International Conference on World Wide Web*, ACM, 1014-1024, 2015.

[15] A. Broder, "A taxonomy of web search," In *ACM Sigir forum*,Vol. 36, No. 2, ACM, 3-10, 2002.