Capturing Semantics of Web Page using TAGs-Based Approach for Information Retrieval

R .Vishnu Priya
School of Computing Science and Engineering
VIT University, Chennai Campus
{vishnupriya.r@vit.ac.in }

***Abstract*—** A web search engine has been developed and used for retrieving the relevant information. One of the critical issues of the engine is to provide relevant information to the user query. This is due to the ranking mechanism used by the conventional system is not effective. This demand has spurred to develop an approach that supports to improve the precision of retrieval. In this paper, the terms related to the TAGs of web pages are considered and weighted according to the characteristics. This ranking will retrieve the users' desired web pages in top-k results. From the experimental result, we have observed that the performance of the proposed approach is encouraging compared to the recent web search engine.

Keywords- HTML TAGs; Ranking; Retrieval.

I. INTRODUCTION

Information retrieval (IR) is a field of study that helps to extract relevant and needed information from a large collection of text documents in the web. On the Web, the documents are web pages. Web search is a most important application of IR and it's challenging issues is finding the high quality web pages that are relevant to the user query. However, the existing web search engines like Google, MSN, Yahoo are cannot display users' required web pages in top-k results.

The most popular search engine says Google uses page rank algorithm[11], where the rank of a page is calculated using two factors says the weighted sum of the ranks of all pages having links to the page and the numbers of click made by the user on the page. Since the logic of ranking mechanism of Google is well-known, some organizations have increased the page rank of their pages in web by manipulating numbers of back links to the page using some tactics (start-online-business-guide). Further, the number of clicks on the page increases the page rank and can be done using a simple source code. Due to this fact, all commercial web pages, social networking web pages like Facebook, Google+, Orkut and irrelevant images are ranked higher by current search engines. This result to frequently navigate through the result pages and it consumes time for users'.

The web search engines have developed another technique for ranking based on keyword search technique. Initially, this technique has used only frequency of occurrence of term for ranking. Later, the semantics of web pages are captured from the syntax of HTML and the engines rank the web pages based on semantics rather than using only terms. Google captures the semantics of web page using features of web pages like page, URL, domain, header, body, heading and link. However, it considers only few features with assigning equal priority which will lead to less progression.

From the above discussions, it is understood that the better ranking criterion is required. This will allow search engines to display the best relevant pages to the user in response to his/her queries. In this paper, we address this problem and develop a novel TAGs based approach. In ranking mechanism, the semantics of web page is captured using the features of web pages say HTML TAGs. These semantics are captured through assigning different weights for all HTML TAGs. For this, the terms and its associated TAGs extract from the web pages and each term is weighted based on the property of the TAGs with which it's associated. A weighting scheme is proposed to give emphases to important terms, qualifying how well they semantically explain webpage and distinguish them from each other. It will retrieve the users' desired web pages in top-k results. The rest of the paper is organized as follows. Section 2 presents the related work and the proposed technique is explained in Section 3. In Section 4, the experimental results are given and we conclude the paper in the last section.

II. RELATED WORKS

In this section, the previously delegated algorithms related to text and image retrieval along with query reformulation are discussed. A recursive connectivity algorithm [2] based on the distance between pages called 'Distance Rank' are considered to compute ranks of web pages. The distance is calculated by summing up the weights of existing links in a shortest path. However, various other important issues are not addressed and it is known that the content of hyperlink varies based on designer's perspective. This aspect favors only the odd pages and even very recent and relevant pages uploaded recently without many inbound links are ignored. Li et al [9] used graph based model to represent multiple interrelated pages. The r-radius steiner graphs are extracted using the set by removing the non-steiner nodes from the corresponding r-radius graphs and ranked the result to return the top-K

answers. However, it is prohibitively expensive to discover the rich structural relationships graph, adjacency matrix and steiner graph, as each page consists of many internal links from other pages. In turn, links from many other websites will be pointing to them. Thus, this structured relation search approach is suitable for a specific domain and may not be suitable for WWW. Park E-K et al [12] has proposed Retrieval Status Value (RSV) for each document using several stages. It assigns more score if the search query terms in the title, in the same sentence apart from adjusting sentence and the document being pointed by anchor tag. All the scores are summed and RSV value of a document is calculated. The ranked documents are finally undergoing stratifying and re-ranking stage. However, it is suitable only for named page finding task and do not work well for the short queries. The value for sentence-query similarity for a document varies once there is a change in query terms and it is a time consuming process. The titles defined within the bodies of web pages are more relevant compared to the content in title field as they are more noticeable to the readers. Addressing this idea, the approach [16] as an automatic extraction of the title from the bodies of web pages. DOM-tree and vision based methods are used for title extraction. However, this approach considers only title of the document, the information present in other TAGs is ignored and thus the low precision of retrieval. Akritidis et al[1] has developed a new Metasearch engine where it does not maintain an own document index and the author weighted each document based on the query terms containing in the zones like title, snippet or URL as many times as possible. However, it has not considered the importance of all the zones for scoring a document. From the above, it is observed that the research on web page retrieval focused with developing techniques to make use of information available in all HTML TAGs for increasing effectiveness of retrieval.

Considering all above issues, a novel TAGs-Pattern based approach is proposed, where concepts conveyed by the users' in the queries are captured and the web pages related to the concepts are ranked based on the weight assigned to the web pages. The weight for each web page is calculated using features (HTML Tags) of web pages. Hence, it helps to retrieve the users' desired web pages in top-k results.

III. TAGs-Pattern Based Approach

There are no strict and identical data structures or schemas, which web pages should strictly follow. As a result, there is an increasing need to better deal with the unstructured nature of web pages for capturing the semantic knowledge of this nature. Usually, web designers are designing the web pages in a well-defined HTML format could hold some preliminary web data structures says TAGs and this structure can primarily help the appearance quality of web pages. Hence, in this approach, this appearance quality (i.e. TAGs) is used for capturing the semantic nature of web pages. Further, in addition to textual information, there are several types of web data such as images, audio and video are exist on the web pages. The appearance quality (i.e. TAGs) of texts and images of the web pages could potentially use to understand the page. Therefore, in this proposed work, this quality is considered to capture the semantics of web pages.

A. TAGs Based Approach for Texts

Let us consider that a web contains a large number of web pages, says $WP = \{wp_1, wp_2, \ldots wp_p\}$, which designed using various HTML TAGs $TG = \{TG_1, TG_2, \ldots TG_q\}$ and are set of terms occurred between *TL*, where *p,q,r* are the total number of web pages, TAGs and terms respectively. Usually, different TAG has different nature/ property and based on the nature/ property the weight is assigned. There are totally 94 HTML TAGs used for designing the web pages and it is divided into 6 distinct groups based on its characteristic, where TAGs in the first level are <meta> and <title>. Since, it's used to describe the content of the page. For instance, if the user searching term lies between <meta> and <title> TAGs of the particular page then the whole web page describes the details related to the searching term. It conveys more details about the searching term and the highest weight is assigned to the terms present between these TAGs. TAGs which briefly/ elaborately/ pictorially represent the content namely <caption>, <img>, <h1>, <a>, <marquee>, <blink> and <cite> are placed in the second level. Usually, the <caption> describes the details of the searching term briefly in the form of a table, <img> is used to pictorially represent the term, <a> and <h1> are elaborately described the related details and the last two TAGs namely <marquee> as well as <blink> signifies the current important information about the term. It describes the details of searching term as a part and it is less compared to the first level. Thus, the texts represented by these TAGs gains a next level of importance.

Similarly, the TAGs in Level-3 category relate to font style namely <b>, <u>, etc., and totally there are 14 font style TAGs. This helps to visually highlight and distinguish the appearance of text on the web page for conveying the important information to the user. For instance, the user searching term is in highlighted form in a particular web page, then the sentence related to the highlighted text is used to describe the details about the searching term and it is the sentences which convey less detail comparing to second level. Hence, the texts available in these TAGs are given next level of importance. The TAGs in Level-4 are used for normal description of text and may not express essential information compared to the text present within other TAGs in the higher levels. Such TAGs are <p>, <dd>, <h3>, <h4>, <h5>, <body>, <basefont>, <td>, <li> and <ins>. Some TAGs in Level-5 are used for design purpose namely <button>, <label> etc., and totally 19 TAGs are related to GUI design. In general, no texts are found between the TAGs in Level-6 namely <html>, <head>, <noscript>, <table> etc., where totally 35 TAGs exist related to this type. This TAGs based model is developed based on the way it used to describe the details about the searching term and it helps to effectively capture the semantics of web pages.

Now, the weight "100" is distributed among these levels, where the weight "0" is manually assigned to level-6. Since, no terms are found between those TAGs. Further, GUI related TAGs do not provide details about searching the term. For instance, the user searching term exists on the label of the particular web page, that web page doesn't contribute any details related to the searching term, therefore, the lowest weight "1" is assigned to Level-5. Now, the rest "99" is distributed among the four levels using the Equ 1.

$$DWGT = \sum_{L=1}^{n}\left(\frac{(N-(T*L))}{(2*L)}\right) \quad (1)$$

where n=4, N is a normalization factor in the value as 99, T corresponds to the total number of levels, which is 4 and L stands for the respective level number.

Let us assume that the user enters the query term says $t_{qy} \subset \{wp_f\}$ such that $t_{qy} \in \{TL_3, TL_5, TL_1, TL_2\}$ respectively, where f=1to 4 and TL_i represents the TAG from $Level_i$. Using the TAGs based model, the TAGs based approach for text ranked the web pages in the order wp_3, wp_4, wp_1 and wp_2 as retrieved results. While, wp_3 describes the details related to t_{qy} as a whole web page, wp_4 as part of the web page (i.e. paragraph, table, image and so on), wp_1 as sentences and wp_2 doesn't contribute the details. Based on the way, the details are described related t_{qy}, the web pages are ranked and this shows that this model effectively ranked the user relevant web pages.

IV. Experimental Results

Google is a search engine with huge varieties of indexed web pages and evaluating with it will be more practical than searching in small-scale. The proposed approach has generated uncontrolled dataset using Google. The queries related to various domains such as colleges, Universities, Institutes, research center, flower, famous leaders, newspapers, sports, cine field, tourism etc., are provided as a query into Google and for each query result of the top-500 links are collected. This collection contains at least one qualified image on each web page and those links are given as input to the crawler for fetching those particular pages. The web pages associated with each query are saved in a separate repository, where all web pages in the same repository are semantically related and relevant to the same topic. The aim is to evaluate and investigate retrieval approaches using this heterogeneous collection of web pages that are browsed by users with various information categories. The total number of web pages in the dataset is 1, 59, 818 that cover various topics of interest and each page include the mixed and overlapping of text/ images. In this evaluation, the ground truth about Google is not known and the only way to present performance objective is through Perceived Precision (PP). Usually, online users are interested in top-10 to top-100 web pages and it is planned to evaluate PP for these ranked web pages.

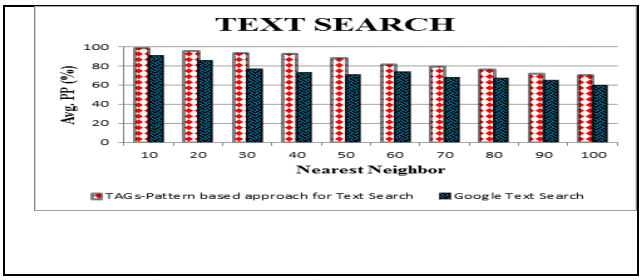

Fig. 1. Performance Metrics for Perceived Precision a. TAGs-Pattern vs TAGs for text b. TAGs-Pattern v's TAGs for image c. TAGs-Pattern for text vs Google text search d. TAGs-Pattern for image vs Google Image search

It is observed from Fig 1 that the TAGs-Pattern based approach for text at the top-10 and top-100 is 98% and 70%, while in case of Google is 90% and 59% respectively. This performance degradation of Google is due to the fact that it displays all commercial and social network pages within the top-100. For instance, while providing the keywords in Google say "National Institute of technology Trichy", the social networking pages like Facebook, Google+, twitter, Orkut of students pursuing degrees in this Institute are displayed on top-100. However, the TAGs-Pattern based approach for text displays only relevant information and not the commercial information. This shows that the weighting mechanism perform better than recent search engine.

Conclusion

Usually, retrieval system returns a large number of web pages for a user query. This is due to the ranking mechanism used by the conventional system is not efficient. In TAGs-based approach, ranking is done effectively for capturing semantics of web pages. In ranking, the semantics of the web page is captured using the features of web pages say HTML TAGs and the weight is assigned to each TAGs based on its characteristics. Web pages are analyzed and extracted texts are weighted based on the TAG with which it is associated. In experimental result, the TAGs based approaches are compared with one of the popular search engines, Google, and found that the proposed work is performs well compared to Google.

References

[1] L. Akritidis, D. Katsaros and P. Bozanis, " Improved retrieval effectiveness by efficient combination of term proximity and zone scoring: A simulation-based evaluation", In: *Journal of Simulation Modelling Practice and Theory*, Vol. 22, pp. 74–91, 2012.

[2] A. M. Z. Bidoki and N.Yazdani, "DistanceRank: An intelligent ranking algorithm for web pages", *Int. J. Information Processing and Management*, Vol. 44, No. 2, pp.877–892, 2008.

[3] Y-L. Chen, Z-W. Hong and C-H.Chuang, "A knowledge-based system for extracting text-lines from mixed and overlapping text/graphics compound document images", *International Journals of Expert Systems with Applications*, Vol. 39, pp. 494–507, 2012.

[4] S. Chien and N. Immorlica, "Semantic similarity between search engine queries using temporal correlation", In: WWW '05: Proceedings of the *14th International Conferenceon World Wide Web*, ACM Press, New York, NY, USA, pp. 2–11, 2005.

[5] W.Hu, O.Wu, Z.Chen, Z. Fu and S. Maybank, "Recognition of Pornographic Web P ages by Classifying Texts and Images", *IEEE Transactions on Pattern Analysis And Machine Intelligence*, 29(6), 1019–1034, 2007.

[6] M.S. Khan, M. Muyeba, F. Coenen, D. Reid and H. Tawfik, "Finding Associations in Composite Data Sets: The CFARM Algorithm", *Int. J. of Data Warehousing and Mining*, Vol. 7, No. 3,pp. 1-29, 2011.

[7] Y.S. Koh, R. Pearl and Gillian Dobbie, "Automatic Item Weight Generation for Pattern Mining and its Application", *Int. J. of Data Warehousing and Mining*, Vol. 7, No. 3, pp.30-49, 2011.
[8] Luping Li, Stephen Petschulat, Guanting Tang, Jian Pei and Wo-Shun Luk, "Efficient and Effective Aggregate Keyword Search on Relational Databases", *Int. J. of Data Warehousing and Mining*, Vol. 8, No. 4,pp. 41-81, 2012.
[9] G .Li et al, "An effective3-in-1keyword search method over heterogeneous datasources", *Int. J. Information Systems*, Vol. 36, No.2, pp. 248–266, 2011.
[10] Z. Liu and Y.Chen, "Identifying return information for XML keyword search", In Proc. On *ACM SIGMOD international conference on management of data*, pp.329 – 340, 2007.
[11] L. Page, S. Brin, R. Motwani and T. Winograd, "The PageRank citation ranking: Bringing order to the web (Tech. Rep.)" *Stanford, CA: Stanford Digital Library Technologies Project*,1998.
[12] Park, Eui-Kyu., Dong-Yul . Ra, and Myung-Gil. Jang, "Techniques for improving web retrieval effectiveness", *Int. J. Information Processing and Management*, Vol.41, No.5, pp.1207–1223, 2005.
[13] A.Vadivel, S. Sural and A.K. Majumdar, "Image Retrieval from Web using Multiple Features", *Online Information Review, Emerald*, Vol.33, No.6, pp.1169–1188, 2009.
[14] R.Vishnu Priya and A.Vadivel, "Capturing Semantics of Web Page using Weighted Tag- tree for Information Retrieval ", In: *International Journal of Asian Business and Information Management, IGI Global*, ISSN: 1947-9638, Vol.3, No.4, pp.7-24, Oct-Dec 2012.
[15] Y. Xue et al., "Web page title extraction and its application", *Information Processing and Management*, Vol. 43, No.5, pp.1332–1347, 2007.
[16] H.-C Yang and C.-H. Lee, "Image Semantics Discovery from Web Pages for Semantic-based Image Retrieval using Self-organizing maps", *Expert Systems with Applications*, Vol.34, pp.266– 279, 2008.
[17] Zhiyong Zhang and Olfa Nasraoui (2008), "Mining search engine query logs for social filtering-based query Recommendation", *Applied Soft Computing*, Vol.8, pp.1326–1334.

http://www.start-online-business-guide.com/how-to-increase-page-ranks.html