IJITKM Volume 8 • Number 2 Jan-Jun 2015 pp. 8-14 (ISSN 0973-4414)

Use of HTML Tags in Web Search

Manjit Singh1, Dr.Dheerendra Singh2, Dr.Surender Singh3
1, Ambala College of Engineering and Applied research,Haryana,India
2Shaheed Udham Singh College of Engg and Technology, Tangori, Mohali, Punjab, India
3 HCTM Technical Campus,Haryana,India
manjitbehniwal@rediffmail.com,professordsingh@gmail.com,ssjangra20@gmail.com

ABSTRACT
In this paper we first describe in brief as to how web search is done with or without HTML tags and then briefly survey the available literature on web search in which HTML tags have been used. Brief introduction to the work that we have currently undertaken on use of HTML Tags in web search is also given.
Keywords: Web search, HTML Tags.

1. INTRODUCTION

Information Retrieval is the science of storing data, searching data, and for searching information within data. In computing context, data are raw entities manageable by a computer device. These include, but are not limited to, text documents, web documents, images, graphs, videos, and audio clips. After several decades of effort, research, and development, the IR field has matured substantially and has now became widely diverse and pervasive. It is now capable of retrieving in just a few seconds textual and non-textual information out of millions of documents. The invention of the Internet in late 60's and the World Wide Web in early 90's is now meeting considerably the requirements of users to search the web and locate online information in a consistent and efficient manner. For this purpose search engine have now been developed. In essence, a web search engine such as Google, Yahoo, and Bing is an online IR system whose purpose is to grab web contents dispersed over the web and index them into searchable databases that allow Internet users to retrieve relevant web pages.

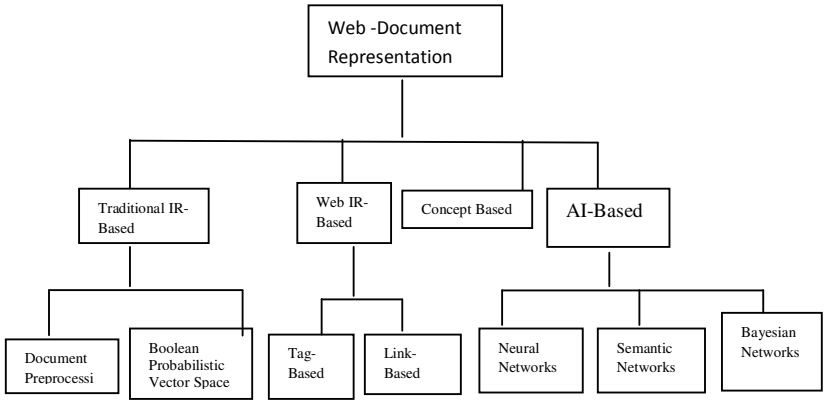
Alessandro Micarelli [] Most of the web pages available on the Internet are written using Hyper Text Markup Language (HTML)which is composed of a set of markup tags that describe the content, the presentation, and the layout of the web page. Unfortunately, little work has been done to effectively utilized HTML documents in information retrieval context by associating the inner structure of the HTML language with the semantics of the document.

2. WHAT IS WEB SEARCH

Web search means extracting required information from the internet. It is well known that, each one of us can benefit from a large quantity of information, available on-line on the internet .It is available in several, essentially standard, formats, such as HTML, XML and XHTML text pages and jpeg and tif graphic ones. More complex formats, particularly multimedia (audio and/or video) usually require a longer search and interaction with the Web, as well as more sophisticated fruition tools installed on user clients, now widely accessible. The quantity of information available on the World Wide Web is increasing exponentially, and this boom has paved the way for a new era, creating new opportunities in many different fields, such as e-business, e-commerce, e-marketing, e-finance and e-learning, just to mention some of the most interesting ones.

The representations of documents come from the traditional IR research field, whose original goal was to retrieve relevant documents matching user needs, expressed through queries. With the advent of the Web and of the HTML language, used to write the vast majority of web documents, other document representations have been suggested, not only based on the single terms that made them up, as occurs for classic IR, but also on other features typical of the hypertext environment, expressed through hyperlinks and/or HTML tags. These exploits the characteristics of hypertext languages of documents: a term t of a Web document d is also given a weight also according to the HTML tags. For example, if a term t appears between the two HTML tags <TITLE> and < /TITLE>, it means it is part of the document title, and must thus be significantly weighted, as it could be correlated to the semantics of the very document, unlike a term included between the two HTML tags <BODY> and < /BODY>, a part of a plain text, where its importance could be relative.

The techniques currently in use to extract web information may be depicted as under:



3. Information Retrieval Techniques

The traditional Information Retrieval document modeling techniques for the retrieval of relevant documents from a collection that were proposed when the Internet was only just dawning are as under

Document Pre-processing

A document is stored on the Web in several formats: HTML, TEXT,PDF etc, but in all cases, the information it contains is to be enriched with particular character sequences, not intended for the user but for the computer visualization driver, such as the browser in the case of HTMLWeb pages. However, we shall deal only with HTML documents.

A traditional IR system extracts significant lines for the user in order to select the correct terms to be represented on the Web page. This term-selection process is a part of the document pre-processing task which is broken down into four phases namely.

(i) HTML Tag Removal**:** This phase consists in removing all HTML instructions (i.e., tags) from an HTML page. The simplest approach excludes all terms belonging to HTML tags. Only text part remains:

(ii) Stopwords Removal: Not all terms of a document are necessarily relevant. Some frequently used terms, within the document itself, tend to be removed: these terms are known as Stopwords (i.e., "a", "the", "in", "to"; or pronouns: "I", "he", "she", "it" etc.).

(iii) Stemming : The goal of this phase is to reduce a term to its morphologic root, in order to recognize morphologic variations of the word itself. For example, the root comput is the reduced version of "comput-er", "comput-ational", "comput-ation" and "compute". The morphologic analysis must be specific for every language, and can be extremely complex. The simplest stemming systems just identify and remove suffixes and prefixes.

(iv)Term Weighting.:Terms are weighted according to a given weighting model which may include local weighting, global weighting or both. If local weights are used, then term weights are normally expressed as term frequencies, tf. If global weights are used, the weight of a term is given by IDF values. Thet common weighting scheme is one in which local and global weights are used and referred to as tf*IDF weighting.

Models for information Retrieval

Several models are in use for information retrieval. These are

(i) Boolean Model

This model is based on Set Theory and on Boolean Algebra: it ascribes a binary value to the weight w_i of a term ti accordingly to its appearance (or non-appearance) in a document $d_k \in D$. In this way, the document d_k is represented by a vector d_k:

$d_k = \{(t_1, w_{1k}), (t_2, w_{2k}), \ldots, (t_m, w_{mk})\}$

where $w_{ik} = 1$ iff $t_i \in d_k$, $w_{ik} = 0$ otherwise and where terms t_i are all the index terms belonging to the ITD of the collection D.

(ii) Probabilistic Model

In the Probabilistic Model, a document (even a query) is modeled through a binary weight vector, as is done in the Boolean Model. The difference is to be seen in the model for the calculation of the query-document similarity function. The probabilistic model tries to answer the following Basic Question : What is the probability that a certain document is relevant to a certain query? The objective of asking the Basic Question is to rank documents according to their probability of relevance. This maximizes the system effectiveness, as retrieved documents are ranked by decreasing probability of relevance. The user

inspects the ranked list of documents and assess their relevance by him/herself. Assuming that terms are distributed differently in relevant and non-relevant documents, one could base the representation and retrieval of documents on term distribution. Both for query q and for the document d_j , index terms are represented by binary weights: $w_{ij} \in \{0, 1\}$ and $w_{iq} \in \{0, 1\}$ and the similarity function $sim(d_j, q)$, i.e., the function that calculates the query-document similarity, is the following:

$sim(d_j, q) = P(R/d_j)/P(\overline{R}/d_j)$

where R is the set of documents known as relevant documents while the set $\overline{R}$ is the complement of set R, namely the set of non-relevant documents. $P(R|d_j)$ is the probability that document d_j will be relevant for the query q, and $P(\overline{R}|d_j)$ is the probability that d_j will not be relevant for q.

(iii) Vector Space Model

Vector-space models, by placing terms, documents, and queries in a term-document space and computing similarities between the queries and the terms or documents, allow the results of a query to be ranked according to the similarity measure used. Similarity between documents and query is calculated as follows:

$$Sim(Q, D_i) = \frac{\sum_i w_{Q,j} w_{i,j}}{\sqrt{\sum_j w_{Q,j}^2} \sqrt{\sum_i w_{i,j}^2}}$$

However it has been observed that HTML Tags significantly improve retrieval of information in web search. The work that has been done in this direction is reviewed in the next section.

4. WEB SEARCH USING HTML TAGS

A review of the available literature shows that web search is a usually being done using features that are extracted from the web page-text only . However, in this a web users usually faces two problems. These are (i) Some of the retrieved documents are not related to the user query(this is called low precision). (ii) and many of the relevant documents are not retrieved (this is called low recall) Most of these problems seem to arise because the retrieval is based on text documents. In order to overcome these problems it is necessary to design to more effective web search algorithms which make use of text segments marked by HTML tags also. These tags have specific meanings which can be utilized to improve the performance of documents retrieval system.

A brief Review of Literature on use of HTML Tags in Web Search:

Molinari et al.[1] proposed an approach to indexing of HTML documents, based on the assumption that tags provide the text with different levels of importance with respect to the document content. A significance degree of an index term can be computed by weighting the term occurrences according to the "importance" associated with the tags in which they appear.

Cutler et al.[2]proposed a method for making use of the structures and hyperlinks of HTML documents to improve the effectiveness of retrieving HTML documents. Their method partitions the occurrences of terms into six classes (title, H1-H2, H3-H6, anchor, strong and plain text) and adjusts the traditional term weighting scheme by incorporating different important factors to term occurrences in different classes. They observed that the terms in the Strong and Anchor classes are the most useful for improving the retrieval effectiveness. They calculates optimal assigned weights as given in Table 1:

A Class Importance Vector, CIV =(civ_1, civ_2, . . . civ_6) is formed, where civi ≥ 0 represents a weight determined experimentally by using genetic algorithms. That weight is the contribution of the i_{th} tag class to the formation of the overall weight of term **t** in document d. Another vector, TFV (Term Frequency Vector) which contains the occurrence frequency of term t in document d for each tag class .

TFV = (tif_1, tif_2, . . . tif_6)

Weight w, to be associated with term t, can be calculated by the formula:

w_t = (CIV •TFV)idf

where idf is the inverse document frequency of term t in the documents of the collection D.

At the end, document vector d = (w_1, w_2, . . . , w_n) is built. It contains the weight of each term of the ITD with respect to the document, to be used to carry out the query-similarity operation.

Kwon et al [3] felt that automatic categorization is only the suitable method to deal with scaling problem of the World Wide Web. They proposed a Web page classifier based on an adaptation of k-Nearest Neighbor (k-NN) approach. To improve the performance of k-NN approach, they supplement k-NN approach with a feature selection method and a term-weighting scheme using markup tags, and reform document-document similarity measure used in vector space model. In their experiments on a Korean commercial Web directory, their proposed methods improved the performance of classification.

Andrea et al [4] proposed an indexing model of HTML documents. In this model the index term weight is computed by weighting the term occurrences differently, according to the tags in which they appear.

Kim et al [5] proposed a text mining approach to Web document retrieval. It uses the tag information of HTML documents to improve retrieval performance. They used Five HTML tags <TITLE>,<H>,<B>,<I>and <A> for the

experiments . The title and header tags use words in a Web page marked as a part of the title or headings to classify that page. The bold and italic tags are taken because they are used to emphasize words. Since the hyperlinks of a document generally are connected to the related documents. So, the anchor tag is also used. They define a similarity measure using tag weights as:

$$\text{sim}(d, q) = \sum_{k=1}^{T} a_{dk} w_{dk} w_{qk}$$

where w_{dk} is the weight of the k_{th} term in the document d, w_{qk} is the weight of k_{th} term in the query q, T is the number of terms, and α_{dk} is the weight of term k in the document d with respect to the tag weights.

Their experimental results show that significant HTML tags of Web-documents can be properly selected and the document retrieval systems can take advantage of this information to outperform the traditional information retrieval approaches which are based on plain texts.They also applied Genetic algorithms, to search significant HTML tags and get the optimal weights.

Byurhan Hyusein et al [6] studied the feasibility of HTML tags to represent the contents of Web documents. They studied title text, anchor text ,emphasized text of the HTML documents and all terms with frequency greater than one, which appears in the whole HTML document. In their study they have presented as to how the different parts of the Web documents contribute to the average precision in the process of search. The anchor and emphasized text were observed to be helpful in indicating the contents of the Web documents. The terms with frequency greater than one were found useful to index and for improving the precision of the retrieved results.

Daniel Ramage et al [8] demonstrate how user-generated tags from large scale social bookmarking websites such as del.icio.us can be used as a complementary data source to page text and anchor text for improving automatic clustering of web pages. They have shown that tagging data improves the performance of two automatic clustering algorithms when compared to clustering on page text alone. They observed that a simple modification to the widely used K-means algorithm enables it to better exploit the inclusion of tagging data. Use of a novel algorithm—MMLDA (Multi-Multinomial latent Dirichlet allocation) makes even better use of the complementary similarity information held in a document's words and tags on a general web clustering task.

Youssef Bassil [9] presented an original image information retrieval model to index and retrieve web images embedded within HTML documents. The model uses CBIR techniques to retrieve images based on their content and keyword-based IR techniques to retrieve images based on their enclosing textual metadata. Terms of the textual metadata were weighted using a novel term weighting scheme called VTF-IDF(Variable Term Frequency-Inverse Document Frequency) which assigns variable weights for terms depending on the HTML tags they appear in. Subsequently, this combination of using graphical content alongside with textual metadata, in addition to a weighting scheme that evaluates terms according to their semantic significance with respect to the HTML tag structure, has led to high image retrieval Precision rate that outperformed other traditional and existing image IR models.For this they define a VTF-IDF term weighting scheme as:

$$w_{ij} = (tf_{ij} * \text{variable\_weight}) * idf_j$$

where w_{ij} is the final weight for term j, tf_{ij} is the frequency of term j enclosing image i, idf_j is the Inverse Document Frequency of term j, and variable_weight is a number whose different values are outlined in Table 1.

M.A. Shelke et al.[10]observed that automatically clustering WebPages into semantically relevant classes, results into improved search and browsing on the web. Typically, webpage clustering algorithms only use features that are extracted from the web page-text. However, the advent of social-bookmarking websites like Stumble Upon and Delicious, has led to a huge amount of user-generated content such as the tag information that is associated with the WebPages. In their work, they use a subspace based feature extraction approach which leverages tag information to complement the page-contents of a webpage to extract highly discriminative features, with the goal of improved clustering performance. They consider page-text and users generated tags as two separate views of the data, and obtain a shared subspace that maximizes the correlation between these two views. K-means clustering algorithm is applied using this subspace. They compare their subspace based approach with existing clustering method i.e. word only and show that the subspace based approach leads to improved performance on the webpage clustering task.

Youssef Bassil et al[11] proposed a semantic-sensitive web vector model called SWVM and a term weighting scheme called BTF-IDF intended for web IR systems. Their proposed model exploits the HTML tag structure of web documents to deduce the semantic importance of specific terms with respect to others. Subsequently, terms that appear in certain pre-specified tags are assigned a higher weight. Additionally, synonyms for every single term in the document are generated, then assigned a weight equal to their corresponding terms, and stored as extra features in the vector model.

Ammar Sami Al-Dallal[12] proposed a searching model that is based on GA to retrieve HTML documents. Their model is composed of two main units. The first unit indexes the HTML documents. The second unit is a somewhat modified GA mechanism. They achieved a high Recall and Precision with HTML Documents by applying Genetic Algorithm.The weight assigned to HTML tags used in the inverted index are given in the given Table 1.

Pooja Mudgil et al [13] have implemented an indexing mechanism to store the keyword present in the document with their contextual senses.It also focuses the importance of keywords in different HTML Tags. The mechanism removes the stop words, stems the keyword and after that creates the index. The data structure Trie fasten the search for matched results from the Inverted Index. It also helps the user to process the user query with fast and more relevant results.

Sukrati Pathak et al [14] proposed a new approach for information retrieval of structured documents and through this approach, they are able to consider the possible dependencies among the fields that form the document structure. This model employs classification of HTML tags according to priorities. They defined four classes title (C1), header (C2), emphasized (C3) and Delimiters (C4) to cover the most of functionalities of tags that are mainly relevant to text retrieval.They also defined E-IOWA (Extended-Induced Ordered Averaging Order) which is used as a new method for calculating order inducing variables in HTML web domain. In this study they found that if query terms are present in more than one class in documents, then E-IOWA technique proposed by them is able to aggregate the rank of all those documents in more efficient way and present a final ranked list.

Shihab et al[15] presented a new semantic similarity measure for capturing the alikeness among the web pages. In their proposed approach, each page is represented by a vector of weighted keywords. As a result the similarity measure for two pages involves the similarity computation between the respective weighted keyword sets. Subsequently, the pages are grouped by the fuzzy C-means clustering algorithm. Numerous experiments performed by them confirm that their similarity measure helps efficiently grouping the semantically related pages. Their semantic similarity measure outperforms others by giving lower intra-cluster distance and higher inter-cluster distance. They assigned weights to different HTML Tags as given in Table 1:

Table1:Tags with their assigned weights positions

HTML Tags	**Michal Cutler et al(1997)** (Assigned Weight to Tags)	**Youssef Bassil(Feb 2012)** (Assigned Weight to Tags)	**Ammar Sami Al-Dallal(Aug 2012)** (Assigned Weight to Tags)	**Shihab et al(Sep 2014)** (Assigned Weight to Tags)
<title>	4		6	10
<h1>	6	+10	5	9
<h2>	6	+10	5	8
<h3>	1	……	5	7
<h4>	1	……	…….	6
<h5>	1	……	…….	5
<h6>	1	……	…….	4
<a>	8	……	4	3
<i>	8	……	3	2
<b>	8	……	3	2
<p>	…….	1	…….	1
<body>	…….	……	1	……

STRONG	8	……	…….	……
EM	8	……	…….	……
U	8	……	…….	……
DL	8	……	…….	……
OL	8	……	…….	……
UL	8	……	…….	……
<alt>	…….	+10	…….	
Image's filename	…….	+20	…….	
Image's class-label	…….	+20	…….	

5. CONCLUSION & FUTURE WORK

Keeping in view on the available literature on use of HTML Tags in web search, we propose to cluster HTML documents using K-means and Evolutionary algorithms. Clustering of database will be done off line before query processing so that to answer a query ,the system does not need to search the entire database and instead searches just a limited number of candidate web documents.

6. ACKNOWLEDGEMENTS

The author is grateful to authorities of Ambala College of Engineering & Applied Reserch(ACE),Ambala for providing necessary facilities.I would also like to express a deep sense of gratitude and thanks to my Supervisors Dr. Dheerendra Singh ,Professor and Head,Department of Computer Science & Engineering and Dr. Surender Singh, Associate Professor, Department of Information Technology, Without their wise counsel and able guidance, it would have been impossible to complete the paper in this manner. I am also thankful to Dr. Chander Mohan, Professor, Department of CSE for his help and valuable suggestions in writing of this paper.

7. REFERENCES

[1] [1996],Molinari, Andrea and Gabriella Pasi, "A fuzzy representation of HTML documents for information retrieval systems". In: Proceedings of the Fifth IEEE International Conference on Fuzzy Systems, vol. 1, 1996, pp. 107-112.

[2] [1997],Cutler, M., Shih, Y., and Meng, W. (1997),"Using the structure of HTML documents to improve retrieval". The USENIX Symposium on Internet Technologies and Systems, pp. 241–251. Monterey, California.

[3] [2000],Kwon, O.-W., and Lee, J.-H.," Web page classification based on k-nearest neighbor approach". IRAL '00: Proceedings of the fifth international workshop on Information retrieval with Asian languages, Hong Kong, China, pp. 9-15, 2000.

[4] [2003],Andrea Molinari Gabriella Pasi,,An indexing model of HTML documents,,In proceedings of the 2003 ACM symposium on Applied computing.

[5] [2003],Kim, S., and Zhang, B-T. (2003),"Genetic mining of html structures for effective web document retrieval". Applied Intelligence, vol. 18, no.3, pp.243-256.

[6] [2003],Byurhan Hyusein et al, "SIGNIFICANCE OF HTML TAGS FOR DOCUMENT INDEXING AND RETRIEVAL" International Conference WWW/Internet 2003.

[7][2007],Alessandro Micarelli, Filippo Sciarrone, and Mauro Marinilli,"Web Document Modeling", P. Brusilovsky, A. Kobsa, and W. Nejdl (Eds.): The Adaptive Web, LNCS 4321, pp. 155–192, 2007.

[8][2008],DanielRamage,Paul eymann,Christopher D. Manning,Hector Garcia-Molina,"Clustering the Tagged Web", Copyright 2008 ACM 978-1-60558-390-7 ...$5.00.

[9] [Feb 2012],Youssef Bassil, "Hybrid Information Retrieval Model For Web Images", International Journal of Computer Science & Emerging Technologies (IJCSET), E-ISSN: 2044-6004, Vol. 3 No. 1, February, 2012.

[10] [Mar-2012],M.A. Shelke, K.C. Sadavarte, R.K. Dhurjad ,N.P. Pandit," Improved Web Page Clustering using Words and Tags", 1st International Conference on Recent Trends in Engineering & Technology, Mar-2012, Special Issue of International Journal of electronics, Communication & Soft Computing Science & Engineering, ISSN: 2277-9477.

[11] [2012],Youssef Bassil & Paul Semaan , "Semantic-Sensitive Web Information Retrieval Model for HTML Document". European Journal of Scientific Research, ISSN 1450-216X, vol. 69(4), 2012.

[12] [Aug-2012],Ammar Sami Al-Dallal, "ENHANCING RECALL AND PRECISION OF WEB SEARCH USING GENETIC ALGORITHM", A thesis submitted for the degree of Doctor of Philosophy, School of Information Systems Computing and, Mathematics, Brunel University, August 2012.

[13] [2013],Pooja Mudgil,A. K. Sharma, Pooja Gupta," An Improved Indexing Mechanism to Index Web Documents", 2013 5th International Conference on Computational Intelligence and Communication Networks.

[14] [May-2014],Sukrati Pathak and Sakshi Mitra, " A New Web Document Retrieval Method Using Extended-IOWA (Extended-Induced Ordered Weighted Averaging) Operator on HTML Tags", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. IV (May-Jun. 2014), PP 65-74.

[15] [Sep-2014],Shihab Rahman, Dolon Chapa, and Shaily Kabir,"A New Weighted Keyword Based Similarity Measure for Clustering Webpages", International Journal of Computer and Information Technology (ISSN: 2279 – 0764) Volume 03 – Issue 05, September 2014.