Utilizing the Subjective Intent of Authoring Formats to Perform Focused Web Crawling

Hok Peng Leung and Wynne Hsu

School of Computing
National University of Singapore
Lower Kent Ridge Road, Singapore 119260

{leunghok, whsu} @comp.nus.edu.sg

Introduction

A successful web information retrieval system requires the ability to determine quickly and accurately whether a document or a link should be further explored. Current state-of-the-art web search engines typically use the meta-information in the HTML header to determine the relevancy of the documents. However, many documents on the web do not have such HTML header information. On the other hand, most web documents are formatted carefully to convey some messages to the readers. The hidden information, embedded in these formatting tags, serves as a good source for determining the relevancy of a document with respect to the query context. In this paper, we propose a fast and accurate approach to determining the relevancy of a document by taking into account the information embedded within these formatting tags. Using such information, we are able to quickly narrow down the scope of our search to the most promising sites. In addition, a new query formulation strategy is proposed to further improve the accuracy of the new approach. Based on this new approach, a crawling strategy has been proposed. A number of experiments have been conducted to test the effectiveness of the proposed approach and the crawling strategy. Experiment results indicate that we are able to achieve a significant improvement over the standard information retrieval algorithm based on tf*idf. Furthermore, our algorithm, unlike the tf*idf scheme, does not require the whole document space to be known in advance. This feature makes our algorithm suitable to be used on the web where it is impossible to known in advance the entire document space.

Related Work

A number of researchers have tried to improve the search engine performance by utilizing different information available from the web documents. Marchiori [1] had the idea of considering not only the textual information within a page but also those pages that are pointed to by that page. The assumption here is that hyperlinks that are inside a page must have some correlation with the content of the page itself. This may not be true in general. On the other hand, Junghoo [2] proposed the use of back links crawlers for finding the most useful information. However, to compute this measure, one needs to know in advance the entire web space which is certainly not practical. Seung-Jin [3] implemented a tool, called web view, that can show the decomposition of an html page in graphical form.

As far as we know, no one has used the HTML format tags information to aid in the document retrieval on the web. However, we know of a number of researchers who make use of the HTML tags information to perform additional feature/knowledge extraction task. For example, Tao-Guan [4] presented KPS – a Web Information Mining Algorithm where he used HTML tags for his style similarity comparison.

Our HTML-based Text-Emphasis cum Query Reformulation Approach

The motivation for HTML-Based Text Emphasis is to take into consideration the subjective intent of the author during retrieval by allocating different weights to the different HTML tags used to format the document. This is a promising approach for web search because most web documents are carefully formatted by their authors and thus utilizing such information embedded in these tags can provide valuable feedback with regard to the relevancy of the document to a query. The algorithm is based on the idea of recursively assigning higher weights to those terms that are enclosed within some text-emphasis tags. In addition, query phrase formulation is used to process the query as composition of phrases rather than as individual terms. For example, for a query Q="Natural Language Processing", a document that contains the string "Natural Language…Processing" is more relevant than those documents that contain single words such as "Natural", or "Language", or "Processing." In this strategy, we first generate all the possible phrases that can be formed from such query string. They include:

1. (Natural Language Processing)
2. (Natural) & (Language Processing)
3. (Natural Language) & (Processing)
4. (Natural) & (Language) & (Processing)

For each phrase, we attempt to match it to the document iteratively, starting from the longest phrase (case 1), to the shortest phrase (case 4). The results are then weighted and averaged to give the overall similarity measure. Figure 1 shows the algorithm details for our approach.

ALGORITHM Similarity(String query, String datafile)

1. Let query = <html> concat query concat </html>
2. Strip the datafile of its special characters.
3. Process_HTML_Text_Emphasis(datafile)
4. Let **$V1 = W_0V_0 + W_1V_1 + \ldots + W_nV_n$**, where V_i represent the words that do not appear in the query and W_i is term frequency of V_i
   Let **V2** = {p| p is the longest matching phrase appearing in the datafile when compared with the query}
5. Let k = $\sum_{\forall w \in V1} w_i$
6. Query_Phrases_formulation(query)
7. for each query_phrase formulation do
   - 7.1 Let **$Q_i = P_1 + P_2 + \ldots + P_n$**, where Q_i is the i^{th} query formulation and P_j (1<=j<=n) are the phrases in the Q_i.
   - 7.2 Count the number of occurrence of P_j in ***V2*** and store them in a vector $V = W_{p1}P_1 + W_{p2}P_2 + \ldots + W_{pn}P_n$, where W_{pj} represent the term frequency of phrase P_j

     Let match= $\sum_{j=1}^{j=n} w_{pj}$

   - 7.3 Those words in **V2** that ***are not matched*** are counted and stored in a Vector $R = W_{r1}r_1 + W_{r2}r_2 + \ldots + W_{rm}r_m$, where W_{rt} is the term frequency of word r_t

     Let residue= $\sum_{t=1}^{t=m} w_{rt}$

   - 7.4 Let Normalization = SQRT(k + residue + match)
   - 7.5 Let local_sim= DOT_PRODUCT(Q_i,**V**)
   - 7.6 Normalize local_sim = local_sim Divide by Normalization
   - 7.7 Weight the local sim accordingly and add it to global sim.
8. Normalize global_sim according to the weighting scheme used

Figure 1. The Proposed Similarity Algorithm.

The algorithm begins by formatting the query into HTML structure so that the same parser could be used to extract the terms from the query string as well as from the document (datafile). The next step is to strip the datafile of all special characters (e.g. è) that are encoded separately (&Egrave). After the special characters have been stripped, Process_HTML_Text_Emphasis is called to extract

individual terms/words from the document. The extracted words are then stemmed and stopped. Once this is done, Query_Phrases_formulation is invoked to generate all the possible query phrases.

Crawler's algorithm

Based on the new similarity measure, our crawler decides which is the most relevant page to begin its search. Once a page has been selected, it then determines among the many links that appear within this selected page, the most relevant link to drill down.

Experiment Results

We have performed a number of experiments to measure the precision, recall and effectiveness (time) of our proposed similarity measure algorithm. These experiments are carried out on a database of 1,400 HTML pages. The pages are pre-classified into 14 categories. Each of these categories contains a minimum of 20 pages. The size of these pages ranges from a few kilobytes to a few megabytes.

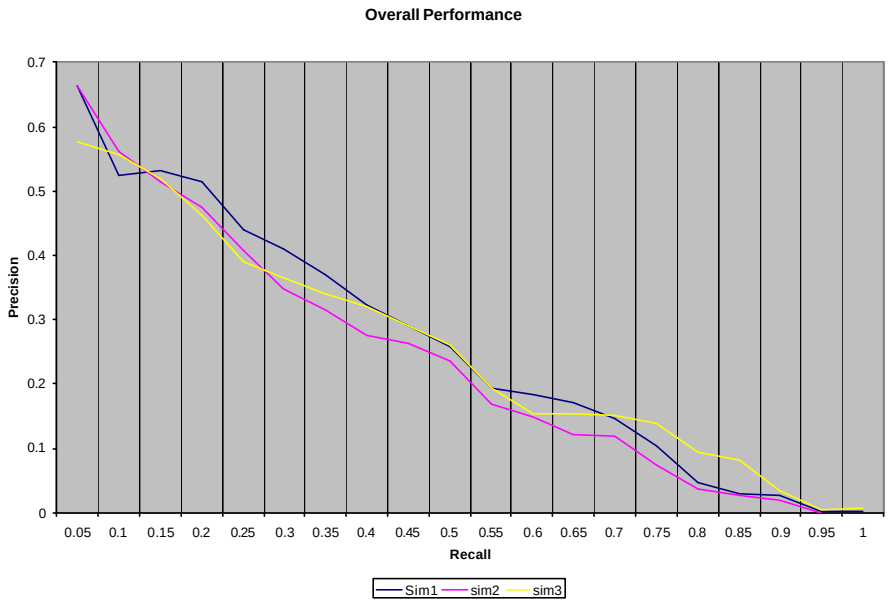

Figure 2. Overall Performance Chart.

Figure 2 shows that the overall performance of the HTML-Based Text Emphasis with Query Formulation is no worst than the standard tf*idf. In fact, **Sim1** (HTML-Based Text Emphasis with Query Formulation) outperforms both **Sim2** and **Sim3** when the recall level is 0.7 or less

Conclusions

Finding the right information on the web effectively and efficiently is a real and pressing need. In this paper, we propose a partial answer to the above need. An intelligent crawler has been implemented that has a number of unique features. The crawler allows drilling to begin from a root page or a page returned by a search engine. It uses a new similarity measure with HTML-based Text-Emphasis and query phrase formulation to intelligently select the best page and best link to drill down. This approach allows us to analyze pages that are hidden below the current page whereas current search engines fails to do so unless the pages are indexed. Experiments show that the new HTML-Based Emphasis and Query Phrase formulation algorithm is able to give better performance than the standard tf*idf.

Reference

1. Marchiori M. (1997), The Quest for Correct Information on the Web: Hyper Search Engines, Proceedings of the Sixth International World Wide Web Conference, Santa Clara, U.S.A.
2. Junghoo C, Hector G, Lawrence P (1997), Efficient crawling through URL ordering, The seventh International World Wide Web Conference (WWW7), Brisbane Australia.
3. Seung-Jin L, Ng Y(1999), Webview : A Tool for retrieving Internal Structures and Extracting Information from HTML Documents, Computer Science Department, Brigham Young University, USA
4. Tao Guan, Kam Fai Wong, KPS a web information Mining Algorithm, The eighth International World Wide Web Conference (WWW8).