Factors Affecting Web Page Similarity

Anastasios Tombros and Zeeshan Ali

Department of Computer Science, Queen Mary University of London, London, U.K
tassos@dcs.qmul.ac.uk, zeeshan@odl.qmul.ac.uk

Abstract. Tools that allow effective information organisation, access and navigation are becoming increasingly important on the Web. Similarity between web pages is a concept that is central to such tools. In this paper, we examine the effect that content and layout-related aspects of web pages have on web page similarity. We consider the textual content contained within common HTML tags, the structural layout of pages, and the query terms contained within pages. Our study shows that combinations of factors can yield more promising results than individual factors, and that different aspects of web pages affect similarities between pages in a different manner. We found a number of factors that, when taken into account, can result in effective measures of similarity between web pages. Query information in particular, proved to be important for the effective organisation of web pages.

1 Introduction

The World Wide Web provides large repositories of electronically stored information. The size, dynamic nature and diversity of content of this information necessitate the development of effective search tools. Web search engines are today one of the most frequently used tools for retrieving information from the web [18]. Apart from research into methods for effective retrieval of information on the Web, there has also been a considerable increase in research into methods for effective information organisation, access and navigation [16]. For such research problems, relationships (i.e. similarities) between web pages become important. Some contexts in which the notion of similarity finds uses include cluster-based search engines (e.g. Vivisimo, iBoogie[1]), web communities [23], the related pages function of search engines [5, 10], identification of duplicate web pages [1], collaborative filtering, and visualisation [17].

Given the importance of page similarity on the Web, it is essential to understand how similarity is determined in this context. Current similarity approaches typically use information from the hyperlink structure of web pages [5, 9], the textual content of the pages [1, 7], and from the structural layout of the pages [3, 14, 27]. A number of approaches combine different sources of information; the most typical combination is that of link and textual content [16, 19, 26].

[1] http://www.vivisimo.com, http://iboogie.tv

D.E. Losada and J.M. Fernández-Luna (Eds.): ECIR 2005, LNCS 3408, pp. 487–501, 2005.
© Springer-Verlag Berlin Heidelberg 2005

The main motivation for this work has been to systematically look into factors that determine similarities between web pages. In past studies on page similarity, the effect of different aspects of web pages (e.g. content, layout, etc.) on the effectiveness of similarity measures has not been thoroughly investigated. In this study we focus on three aspects of information that is available from web pages: the textual information contained within common HTML tags, the structural layout of pages, and the query terms present in web pages. We systematically investigate the effect of these sources by varying their relative importance in the resulting similarity measures, and by examining the effect of the variations in the effectiveness of the similarity measures. Our approach for using multiple sources of evidence is motivated by results of previous research that have suggested that a single source of information for detecting web page similarity is unlikely to be the most effective [3, 7].

In the rest of this paper, we first present some related work on similarity measures on the Web in section 2, then in section 3 we present the details of our investigation, in section 4 we present and analyse the results and in section 5 we conclude and draw some pointers for taking this work further.

2 Related Work

Similarity between documents in information retrieval (IR) is typically measured as a degree of content overlap [24]. Inter-document similarities in IR have been extensively researched, due to their application to areas such as clustering, visualization, etc. [20]. The concept of similarity is also central to the Cluster Hypothesis [13], which states that documents relevant to the same queries tend to be more similar to each other than to non-relevant ones. The hypothesis has been investigated in a number of different contexts [12, 22] where it has been linked to the effectiveness of document clustering for IR.

In web IR, the success of link-based evidence for effective retrieval [15], has led researchers to look into using the same evidence to determine inter-page similarities [5, 9] (e.g. the more links two pages have in common the more similar they are, etc.). Haveliwala and his colleagues [10] have suggested that most web pages under link-based measures have orthogonal vectors. A further possible of link-based measures is that they make it difficult to discover similarity relations for relatively new web pages, which have not been cited enough.

A number of approaches have used evidence from the textual content of web pages to calculate similarities [1, 2, 6, 7, 10]. Different aspects of content have been used. For example, [2, 6, 10] have used hyperlinks and anchor text associated to hyperlinks as a succinct representation of the content of the target page. Referred pages are then typically indexed by some form of aggregation of the anchor texts of their incoming links [10]. In [6] it was also demonstrated that anchor text resembles query text in terms of length and term distribution and that it is also less ambiguous than query text, resulting in more coherent retrieval results. This evidence from past work demonstrates that anchor text is an important aspect of the content of a web page. Some other content-based approaches [7] have used proper

names in web pages to "boost" the effectiveness of similarity measures. A widely used approach for detecting duplicate web pages was also proposed in [1], where a set of contiguous terms, or shingles, extracted from pages are considered, and the number of matching shingles determines the degree of similarity between pages.

A different source used to determine similarity is the structural layout of pages [3, 8, 14, 27]. The rationale of structure-based approaches is that pages containing similar information would also have a similar structure [27], or a similar layout and look-and-feel [14]. An approach that has utilised tag frequency information from web pages to determine their similarity is reported in [3]. This approach is based on the assumption that tag frequencies reflect some inherent characteristics of a web page and correlate with its structure. A number of measures of structural similarity between pages were developed in [3], and their effectiveness was compared to measures of similarity using the text of documents alone. The results showed that certain improvements are introduced by the structure-based approaches; however, the authors emphasise that it is unlikely that structure information alone will be an effective enough source of evidence. They also stress the importance of combining different sources of evidence for the calculation of inter-page similarities. Other approaches that make use of the hierarchical structure of web pages to calculate similarities, include [14, 27]. In such approaches a tree representation of the HTML structure of pages is used to calculate similarities. In general, tree-based approaches have proven to be computationally expensive [14].

A number of approaches have also used combinations of link and content-based approaches to page similarity [16, 17, 26]. Such approaches typically combine content and link similarities. Most of this work however, is limited in that the effectiveness of the resulting similarity measures is not evaluated.

Evidence from past work suggests that a single source of evidence is unlikely to provide the most effective input for measuring similarities. Based on this observation, in our approach we examine three different sources of evidence that are available to us from the content of web pages. First, we look into textual content that is contained within different HTML tags. This source has been investigated for its effect on retrieval effectiveness [4], but not on web page similarity. Second, we investigate how structural layout information can be used to detect similarity, and how it can be combined with other sources. The combination of structural information with other sources of evidence has not been investigated by previous research. Third, we use information provided by the query to increase the similarities between pages that are likely to be relevant to the same queries. This query-based approach to similarity has been shown to be effective for cluster-based document retrieval [20], but it has not been investigated in the context of web pages. In the next section we present the details of our research approach.

3 Research Approach

We first present the three different sources of evidence that we use in section 3.1. Then, in section 3.2 we outline the details of the experiments, and in section 3.3 we describe the evaluation approach used.

3.1 Sources of Evidence

We use three different sources of evidence to calculate similarities between pages, and we examine the relative effectiveness of each source by adjusting its importance in linear combination formulas. The main aim of this approach is not to establish optimal values for the weights of the different sources, but rather to investigate how the different sources affect similarities between pages.

HTML Tags. We use a number of the most common HTML tags typically found in web pages. We place tags in classes depending on their semantic connotations. In Table 1 we present the eight classes we used along with the tags contained within each class. Content within each individual class is indexed separately, and treated as the representation of each class. In [4] a similar approach was employed for indexing web pages: index terms were assigned different weights depending on the classes in which they occurred. This work demonstrated that a significant improvement in retrieval effectiveness is introduced when using tag information. We aim to examine whether the same holds when calculating similarities between web pages.

Titles and *headings* are deemed to be good author-provided representations of the main contents of a page [4]. *Tables* and *lists* offer both a good representation of the layout and general look of a page, and an effective means of capturing information that may be salient within a web page. When calculating similarities based on the table class, we do not take into account numbers, as this may lead to increased noise in the calculations. For the *font* class, we hypothesise that, similar to titles and headings, authors will use special font options for content they deem salient within a page. *Images* are also a significant source of information. We assume that pages that use images in a similar textual context will have some degree of topical relatedness. We implement this assumption by using the text contained within an image window (text withing the bounds of the P tag in which the image is used) and the text in the ALT option of the IMG tag as the representation of the image class.

For the *anchor* class, we follow a different approach to [2, 10]. We suggest that the anchor text coupled with the hyperlink can be a good measure of similarity

Table 1. The eight classes and associated HTML tags

Class name	HTML tags
Content	All legal tags
Title	TITLE
Heading	H1-H6
Table	TABLE, TR, TD, TH
List	LI, OL, UL
Anchor	A and anchor window
Font	STRONG, B, I, U
Image	IMG and image window

of web pages. We assume that in a situation where two pages P_1 and P_2 refer to the same page P_3, the two referring pages are likely to be about a related topic. We only consider this textual information if two pages point to the same page. We should however note that we do not take into account the referred page. Further, to enhance the semantic context in which pages are referred, we use both the text included within the bounds of the anchor tag, as well as the text included within an anchor window defined by the paragraph (P tag) in which the anchor appears.

We use the *content* class as a baseline, as it encompasses the entire textual content of a web page. We are therefore interested in examining how close to this baseline the other classes, or the combinations of the classes, can get.

This approach may be susceptible to some problems: not all authors use the same semantics with HTML tags, web page content may also contain spam content, pages generated automatically may follow different stylistic conventions, the anchor and image windows may introduce noise in the similarity calculations. However, we believe that this approach is general enough to provide us with evidence about the importance of different HTML tags at detecting similarity.

Structure. To determine the structural layout of a page, we examine the frequency with which tags occur in the page. We assume that this frequency will provide an indication of the page's general layout and structure, and may be an effective source of evidence for detecting the similarity of structurally-similar pages. For example, hub pages providing links to other pages will have similar distributions of the anchor and list-related tags.

To measure the structural similarity between pages, we use the tag frequency distribution analysis (TFDA) measure proposed in [3]. The frequency of HTML tags is used to calculate similarities as follows. Let $TagF_{ti}$ and $TagF_{tj}$ be the frequencies of the same tag t in pages P_i and P_j, n the total number of tags, w_t the weight for the *t-th* tag and $\sum_{t=1}^{n} w_t = 1$. Similarity can then be computed as:

$$S(P_i, P_j) = 1 - \sum_{t=1}^{n} (TagF_{ti} - TagF_{tj})^2 * w_t \quad (1)$$

Values are normalised to fall between zero and one. The weights w_t are calculated to be proportional to the ratios of the different tags.

A drawback of this approach is that it does not take into account the order in which tags appear in pages. On the other had, TFDA can be implemented efficiently for on-line calculations and is an effective means of detecting the broad categories to which pages belong (e.g. hub pages, etc.) [3]. In this paper we extend the work in [3] by combining structural similarity with other sources of evidence.

Query. Recent research in inter-document similarity [20, 21] has suggested that similarity measures that take the query into account are more effective than conventional measures. This class of similarity measures is called query-sensitive (QSSM). QSSM are based on the assumption that documents that are jointly relevant (*co-relevant*) to a query, display an inherent similarity that is dictated

by the query itself. QSSM aim to detect this inherent similarity by viewing the query terms as the salient features that define the context under which similarity is examined for an IR task. Conventional similarity measures (e.g. cosine coefficient) are enhanced by the inclusion of a query-sensitive component which introduces a dynamic nature to similarity; similarity values for the same pair of documents are different for different queries. This dynamic component is shown in equation 2, which is based on the cosine coefficient formula:

$$Sim(D_i, D_j, Q) = \frac{\sum_{k=1}^{n} c_k \cdot q_k}{\sqrt{\sum_{k=1}^{n} c_k^2 \cdot \sum_{k=1}^{n} q_k^2}} \tag{2}$$

where $Q = \{q_1, q_2, \ldots, q_n\}$ is the query vector, D_i and D_j are the two document vectors, and $C = D_i \cap D_j = \{c_1, c_2, \ldots, c_k, \ldots, c_n\}$ is a vector which contains the common terms of documents D_i and D_j.

This equation essentially enhances the similarity of pairs of documents that have many query terms, as well as many other content terms, in common. The dynamic component is combined with a standard cosine coefficient measure between the two documents to yield the overall value for a QSSM, as equation 3 shows. Functions that were investigated in [20, 21] were a linear combination and the product of the two components.

$$Sim(D_i, D_j|Q) = f(Sim(D_i, D_j), Sim(D_i, D_j, Q) \tag{3}$$

QSSM have been shown to be significantly more effective than conventional similarity measures at detecting the similarity of co-relevant documents, and have also been shown to significantly increase the effectiveness of cluster-based IR [20]. However, the effectiveness of these measures for web pages has not been investigated. In this paper, we use the query as a source of evidence, and we use equation 2 to measure the query component of similarity measures.

3.2 Experimental Environment

Our experimental approach consists of using a set of documents, queries and relevance assessments to evaluate the effectiveness of various similarity measures. We used the WT2g test collection, from the TREC-8 Web track [11], for our study. This collection includes about 250,000 web pages from a web crawl carried out in 1997. We also used TREC-8 topics 401-450.

It should be noted that the WT2g collection has shown to be inadequate for evaluating the retrieval effectiveness of link-based IR approaches. In [11] it was reported that this collection, which forms part of the larger VLC2 collection, does not contain a large enough number of inter-server links within its pages for link-based methods to be sufficiently evaluated. Although this limitation is significant for link-based approaches, it does not affect the validity of this study, as we do not use any form of link information.

With regards to the TREC queries used, we considered only the *title* part the queries, as we deem this to be more representative of the way searchers formulate queries. The average length of the *title* section was 2.7 terms. For each query, we retrieve the top 100 documents and use them for our study. In [12, 20, 21] it has been demonstrated that using relationships from among documents ranked high by an IR system in response to a query, is more effective than using relationships from entire document collections.

The IR system we used in this study is the Lucene system[2]. We applied standard stemming as provided by Lucene to preprocess the web pages, and we used an extended stop-word list by including non-meaningful terms that are commonly included in HTML pages. Standard *tf-idf* weights were used for document and query terms. Retrieval was performed using the cosine coefficient for matching between documents and queries. The cosine coefficient was also used as the basic formula for measuring page similarity.

3.3 Evaluation Measure

Our evaluation approach is based on the view that for IR tasks, effective similarity measures should structure the information space in such a manner that, for each query, co-relevant documents (web pages) should be closer to each other than to non-relevant documents. This evaluation approach is also reflected by the cluster hypothesis [13]. Based on this approach, the effectiveness of a similarity measure is gauged by its effectiveness at placing co-relevant documents close to each other; this also facilitates the direct comparison of the effectiveness between different similarity measures. The practical significance of the results of such an evaluation is that for applications such as the related pages function of search engines, for a given relevant web page, we can determine how many of its immediate neighbours are also relevant. Cluster-based IR systems would also benefit from similarity measures that result in an increased adherence to the hypothesis [13, 20, 25].

A test which is suited to this evaluation framework is the nearest neighbour (NN) test first proposed in [25]. This test consists of finding the N nearest neighbours (i.e. most similar pages) of each relevant page for a query, and of counting the number of relevant pages in this neighbourhood. The higher the number of relevant pages, the higher the adherence to the cluster hypothesis. A single value that corresponds to the number of relevant pages contained in the NN set (we used a value of 5 for the test, the same that Voorhees used for her experiments) can be obtained when averaging over all of the relevant pages for all the queries in the WT2g collection.

We use the Wilcoxon signed-ranks test to look at the statistical significance of results. This test does not make any assumptions about the distribution of the values that it is comparing. The test assumes that there is information in the magnitude of the differences between paired observations, as well as in the sign of the differences. We consider results to be statistically significant at $p<0.05$.

[2] http://jakarta.apache.org/lucene

4 Results

In this section we present the results of our study. We begin by presenting results from the use of individual sources in 4.1, then in 4.2 we present results from the combination of sources, and in 4.3 we discuss the main findings.

4.1 Individual Sources

HTML Tags and Query Terms. In Table 2 we present the results of the 5NN test when each of the eight tag classes is considered on its own: similarity is calculated based only on matching terms between pages in the respective tag class in the table. The second column of the table contains the average 5NN values for each class, and the third column contains the average number of terms in each class, after stop-word removal.

The results demonstrate that the content class provides the most effective source of evidence. All differences between content and the other classes are significant at levels <0.02. The title class is the second most effective among the tags. What is surprising is that with an average of only 4.4 terms per page for this class, similarity calculations are relatively effective. This demonstrates that titles of web pages are good descriptors of the topical content of pages. The same, to a slightly lesser degree, applies to the heading class. A further observation from this table is that images and anchors are not particularly good sources of evidence. One potential reason for this is the amount of noise that may be introduced by the anchor and image windows used.

The last row of Table 2 presents the 5NN value for the QSSM given by equation 2. For this study we used a slightly expanded form of the TREC queries, by using the *description* section of the queries. This was based on findings in [21] which suggest that QSSM tend to be affected by query length. The new average length of the queries used is 7.2 terms (compared to 2.7 for the *title* section alone). The 5NN values given by the QSSM are the second most effective. The values obtained with the QSSM are significantly more effective than those obtained with all the tag classes, apart from the title class. Content is still signif-

Table 2. Average 5NN values and statistics for individual classes

Class	5NN	Avg. terms per page
Content	2.45	539.83
Title	2.13	4.39
Heading	2.07	8.92
Font	2.06	26.93
Anchor	1.98	63.10
List	1.97	33.02
Table	1.97	64.81
Image	1.88	13.64
Query	2.24	n/a

Table 3. Average 5NN values using tag frequency

Title	Heading	Font	Anchor	List	Table	Image
1.84	1.73	1.87	1.81	1.84	1.86	1.92

icantly more effective than QSSM. We view these results as providing evidence that the presence of query terms in pages is an effective source of evidence for page similarity.

Structural Layout. In Table 3 we present the average 5NN values resulting from the structural layout-based similarity between pages. Values in this table are calculated by matching the tag frequency of the individual classes using equation 1. No data are calculated for the content class, as this class includes all possible HTML tags.

The results based on tag frequency do not show a high degree of correlation to those in Table 2. The image class is the most effective source in this case, whereas headings and titles are less effective. Few of the differences between classes are statistically significant: the image, font, title and table classes provide significantly better results than the heading class, and the image class is also significantly more effective than the anchor class. It seems that information from tag distribution alone is not a good source for measuring page similarity. This result is in agreement to [3]. In the following section we combine tag frequency information with various other sources.

4.2 Combinations of Sources

When calculating the similarity $S(P_i, P_j)$ using combined sources, we use a linear combination of the sources: $S(P_i, P_j) = \alpha * S_{s1}(P_i, P_j) + \beta * S_{s2}(P_i, P_j)$, where α and β are adjustable parameters ($\alpha + \beta = 1$), and $S_{s_1}(P_i, P_j)$, $S_{s_2}(P_i, P_j)$ are the similarities of pages P_i and P_j according to sources s_1 and s_2. By varying α and β, we are interested in examining the relative effect that different sources have on the effectiveness of page similarity.

Combining Tag Classes. We combined pairs of tag classes and measured the average 5NN values of the resulting similarity measures. The rationale of these combinations is to examine whether specific pairs of tags provide better sources of evidence than individual tags. Our results showed that all classes benefit from combination with the title class. This result correlates with the high effectiveness of the title class reported in section 4.1. The combination with the title class has a "smoothing" effect, and the best 5NN values of the combined pairs of classes are now not significantly different to each other. The range of the best values is from 2.03 for images and titles to 2.2 for titles and tables. These improved values are generally significantly better than those obtained from individual classes. It should also be noted that, in general, the parameters α and β in

these combinations were in a the region of 0.6 and 0.4, with the highest weight attributed to the title class.

Combinations of other tag classes apart from the title class, still improve the 5NN values obtained from the individual classes, but not to the same extent as the title class did. Some promising results are obtained by the combination of lists and fonts (2.13), tables and headings (2.12) and tables and fonts (2.09). We further combined the most effective class pairs with the content class. The rationale of this comparison is to examine whether certain tag classes are worth "promoting" when measuring page similarities. For example, if two pages have substantial overall content overlap which is focused on specific tags (e.g. lists or tables), then by appropriately rewarding the tag-specific similarity we can examine the effect on the 5NN values.

The results from this study are positive. The combinations of all possible pairs of classes with the content class yield 5NN values between 2.43 and 2.49. These values represent best average values, i.e. they have been obtained at optimal settings for the α and β parameters. In general, these parameters weighted the content class more than the other classes, in ratios of 3:2. Unlike the combination of pairs of classes where the title class yielded the most effective combinations, in these results there is no clear tendency for a single class to be optimally combined with other classes and with the content class. The highest best average values are obtained with combinations of table and images with content (2.49) and with table and lists with content (2.48). These results are not significantly better than the 5NN values attained by using the content class alone.

Combining Tag Frequencies. We split the seven tag classes for which we have tag frequency data into two groups: group A contains the anchor, list, table and image classes, while group B contains the title, heading and font classes. The first group corresponds to layout-oriented tags of a page, while the second to more content-oriented tags. We also used different settings of the parameters to assign weights to classes within each group (we used 4 parameters in this case, with the sum of the parameters equal to 1). These weights were representative of the effectiveness of the individual classes when using the TFDA-based similarity measure. We then calculated page similarities based on each of the two groups.

The best average 5NN values obtained in this study were 2.13 for group A and 2.06 for group B. For both groups, these values are significantly better than values from tag frequencies from individual classes. The difference between the two groups is not statistically significant; however, it is consistent. The consistency of the results provides evidence that the layout-oriented tags provide a better source for measuring page similarity using TFDA.

We also calculated 5NN values based on combinations of frequencies of pairs of classes. These results were not significantly better than those obtained from individual classes. However, all possible combinations of pairs consistently gave results that were higher than any of the individual parts of the pair. Further, combinations that included the image, list, table and to a lesser extent, the anchor classes, were consistently the most effective.

Table 4. Best average 5NN values for combinations of tag frequencies and tag classes

	Font	Heading	List	Title	Content
F_{image}	2.10	2.07	1.97	2.15	2.43
F_{list}	2.02	1.99	1.92	1.97	2.45
F_{anchor}	1.95	1.98	1.99	1.98	2.42
F_{table}	2.01	2.01	1.94	2.11	2.44

We also combined tag frequency information with tag classes. We calculated all combinations of individual tag classes and individual tag frequencies (including content). We present the most effective combinations in Table 4. Values in a cell of the table (5NN values) are derived by the combination of the respective row (tag frequency) and column (tag class), and correspond to the best average value attained.

The classes whose tag frequencies combined best with tag classes were the image, list, anchor and table classes. In general, combinations of tag frequencies with tag classes (except content) yield 5NN values which are significantly higher than values obtained by using individual tag frequencies. For example, using individual tag frequency from the image class, the average 5NN value was 1.92 (Table 3); combining image tag frequency with the font class the value significantly increases to 2.10. All these values are however significantly lower than those obtained by the content class alone (2.45, Table 2).

When combining tag frequencies with content, there is a significant increase in the 5NN values (compared to other tag frequency values), as this is demonstrated in column 6 of Table 4. These best values reported in the table are obtained at parameter settings that weight the content class four times more than the tag frequency information. These results are either slightly less than, or equal to, the 5NN value for the content class alone (2.45). These results are significantly higher than all other data using tag frequency information.

Combining Query with Other Sources. We combined the query-based measure given in equation 2 with the content class. This is equivalent to the linear combination function of the static and query-based similarities reported in [20, 21]. By varying the parameters α and β we vary the importance attributed to the static (i.e. cosine) and the dynamic (i.e. equation 2) components of the similarity, respectively. These results are reported in column 2 of Table 5. In columns 3, 4 and 5 we report the results from the combination of content and query sources (C+Q) with the font (F), heading (F) and title (T) tag classes respectively.

The results in Table 5 are generally higher than those obtained by using the content class alone (2.45 from Table 2). If we examine the results in column 2, we see that as the effect of the static component of the similarity increases, so does the 5NN value. The highest value in this column is achieved when the static component is weighted four times as much as the query component. These results do not agree with ones previously reported in [20, 21], where best results were

Table 5. Average 5NN values for combinations of query-based measures

$\alpha : \beta$	C+Q	(C+Q)+F	(C+Q)+H	(C+Q)+T
0.2 : 0.8	2.33	2.40	2.41	2.42
0.4 : 0.6	2.42	2.43	2.41	2.44
0.5 : 0.5	2.48	2.47	2.46	2.49
0.6 : 0.4	2.50	2.53	2.53	2.52
0.8 : 0.2	2.54	2.53	2.55	2.55

observed for higher weighting of the query component of the similarity. However, these results were obtained from different TREC test collections, and it is likely that the different properties of the collections have caused this difference. It should also be noted that in the case where α=0 and β=1, the 5NN value is equal to that reported in Table 2 for query alone (2.24), and if α=1 and β=0, it is equal to the value reported in Table 2 for content alone (2.45).

We also combined the results of the joint query and content similarity with different sources. In columns 3, 4 and 5 we report the most effective combinations with the font, heading and title classes respectively. The trend of these results is similar to those in column 2; by increasing the effect of the joint content and query similarity, we also increase the 5NN values. In a large number of cases, the values of the combinations are higher than those of content alone (2.45), or even higher of the respective values in column 2 of the same table.

There are few significant differences in these results, mainly for different settings of the parameters in the same similarity measure (same column of the table). There are also two significant differences, for $\alpha : \beta = 0.8 : 0.2$, between the content only value of 2.45 and the content and query value of 2.54 (column 2) and the content-query and title value of 2.55 (column 5).

4.3 Discussion

In the results reported in the previous sections, the effect of the content of web pages at determining similarity is significant. The baseline set by measuring similarity using content alone was exceeded by only a few cases. Most of these cases involved the use of the query as an additional source of evidence. The significance of query terms for determining the similarity between pages is an important finding of this study. Unlike previous studies of the effectiveness of QSSM, in this study the strong effect of overall content overlap weakens the effect of query terms on page similarity. Despite this, clear effectiveness gains are introduced by the incorporation of query information.

By looking at the results using evidence from tag classes and the query, there is a general trend for the title, heading and font classes to provide effective sources for measures of similarity. These three classes seem to capture a significant amount of information related to a page's semantic content. The incorporation of content from these three classes in general improves the effectiveness of the similarity measures.

With regards to the structural-layout based similarity, results obtained with the TFDA measure were less effective than those obtained with other sources. In general, tag frequencies from single classes demonstrated low effectiveness. Improvements were introduced by combinations of classes, and by combinations of structural and content-based sources. The classes that consistently displayed a positive effect on similarity were those of anchors, images, lists and tables. The distribution of these classes within pages proved to be an effective source of evidence for calculating similarities. These tag classes may also be better at distinguishing different categories that pages belong to. For example, by looking at the distribution of anchor and list items within a page, we can infer whether this page is a hub page.

The TFDA-based similarity measures may also be introducing a certain level of noise in similarity calculations. For example, two pages may have similar tag frequency distributions but their actual topical content may be different. It would, however, be worthwhile to investigate how well structure-based similarity correlates with searchers' perception of page similarity. One can expect that structurally-motivated similarity may be better suited to searchers' intuitive interpretation of similarity than to TREC relevance assessments. This issue would need to be further investigated. Different methods for calculating structural similarity can also be investigated. Of particular interest would be measures that take into account the order of occurrence of tags in pages (e.g. [3]).

A further result from our study is that combinations of sources of evidence generally yield similarity measures that are more effective than the constituent sources alone. This was particularly evident in the case of the TFDA-based measures. When linearly combining different sources, results obtained within a region of the optimal parameter settings were also not significantly different to each other. It should be noted that the effectiveness of combination of evidence in IR is well-established [26].

The results of this study also suggest that certain query types are better suited to certain types of evidence for measuring similarity. This has been observed from a per-query analysis of the results, and has been a by-product of the current investigation. We aim to further analyse this behaviour in our data.

5 Conclusions and Further Work

In this paper we examined the effect that different aspects of web pages have on determining inter-page similarity on the Web. We looked into the textual content contained within common HTML tags, the structural layout of pages as defined by the distribution of HTML tags, and the presence of query terms in pages. The results of this study suggest that certain aspects of web pages are effective sources of information for calculating inter-page similarity. The presence of common query terms in pages was a particularly effective source of evidence. Further, the textual content of certain HTML tag classes (title, headings, font) and the tag frequency of the table, list, anchor and image classes also proved

to be effective factors for similarity calculations. Combinations of factors were more effective than individual factors.

This study can be extended by looking into user-oriented factors that determine page similarity, by using a larger dataset, and by looking more thoroughly into the dependence of certain query types on certain sources of evidence for similarity calculations. Other types of similarity measures (e.g. link-based measures) can also be examined to establish whether similar factors affect their effectiveness. The results from this study can have implications for systems that rely on the effective calculation of web page similarity, such as systems that retrieve or recommend web pages. We view this study as an important step towards understanding how similarities between web pages are determined.

References

1. A.Z. Broder, S.C. Glassman, M.S. Manasse, and G. Zweig. Syntactic clustering of the web. In *Proceedings of the 6th WWW Conference*, pages 1157–1166, 1997.
2. S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proceedings of the 7th WWW Conference*, pages 65–74, 1998.
3. I.F. Cruz, S. Borisov, M.A. Marks, and T.R. Webb. Measuring structural similarity among web documents: preliminary results. In *Proceedings of the 7th International Conference on Electronic Publishing*, pages 513–524, 1998.
4. M. Cutler, H. Deng, S.S. Maniccam, and W. Meng. A new study on using html structures to improve retrieval. In *Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence*, pages 406–409, 1999.
5. J. Dean and M. Henzinger. Finding related pages in the world wide web. In *Proceedings of the 8th WWW Conference*, pages 1467–1479, 1999.
6. N. Eiron and K.S. McCurley. Analysis of anchor text for web search. In *Proceedings of the 26th ACM SIGIR Conference*, pages 459–460, 2003.
7. N. Friburger and D. Maurel. Textual similarity based on proper names. In *Proceedings of the ACM SIGIR Workshop on Mathematical Formal Methods in Information Retrieval*, pages 155–167, 2002.
8. P. Ganesan, H. Garcia-Molina, and J. Widom. Exploiting hierarchical domain structure to compute similarity. *ACM Transactions on Information Systems*, 21(1):64–93, 2003.
9. M. Halkidi, B. Nguyen, I. Varlamis, and M. Vazirigiannis. Thesus: Organising web document collections based on link semantics. *VLDB Journal*, 12(4):320–332, 2003.
10. T.H. Haveliwala, A. Gionis, D. Klein, and P. Indyk. Evaluating strategies for similarity search on the web. In *Proceedings of the 11th WWW Conference*, pages 157–163, 2002.
11. D. Hawking, E. Voorhees, N. Craswell, and P. Bailey. Overview of the trec-8 web track. In *Proceedings of TREC-8*, pages 131–150, 2000.
12. M.A. Hearst and J.O. Pedersen. Re-examining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of the 19th ACM SIGIR Conference*, pages 76–84, 1996.
13. N. Jardine and C.J. van Rijsbergen. The use of hierarchical clustering in information retrieval. *Information Storage and Retrieval*, 7:217–240, 1971.

14. S. Joshi, N. Agrawal, R. Krishnapuram, and S. Negi. A bag of paths model for measuring structural similarity in web documents. In *Proceedings of the 9th ACM SIGKDD Conference*, pages 577–582, 2003.
15. J.M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
16. D.S. Modha and W.S. Spangler. Clustering hypertext with applications to web searching. In *Proceedings of the 11th ACM Conferencei on Hypertext and Hypermedia*, pages 143–152, 2000.
17. S. Mukherjea. Organizing topic-specific web information. In *Proceedings of the 11th ACM Conference on Hypertext and Hypermedia*, pages 133–141, 2000.
18. S. Ozmutlu, A. Spink, and H.C. Ozmutlu. A day in the life of web searching: an exploratory study. *Information Processing & Management*, 40(2):319–345, 2004.
19. P. Pirolli, J. Pitkow, and R. Rao. Silk from a sow's ear: extracting usable structures from the web. In *Proceedings of ACM SIGCHI Conference*, pages 118–125, 1996.
20. A. Tombros. *The effectiveness of hierarchic query-based clustering of documents for information retrieval*. PhD thesis, Department of Computing Science, University of Glasgow, 2002.
21. A. Tombros and C.J. van Rijsebrgen. Query-sensitive similarity measures for information retrieval. *Knowledge and Information Systems*, 6(5):617–642, 2004.
22. A. Tombros, R. Villa, and C.J. van Rijsbergen. The effectiveness of query-specific hierarchic clustering in information retrieval. *Information Processing & Management*, 38(4):559–582, 2002.
23. M. Toyoda and M. Kitsuregawa. Creating a web community chart for navigating related communities. In *Proceedings of the 12th ACM Conference on Hypertext and Hypermedia*, pages 103–112, 2001.
24. C.J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 2nd edition, 1979.
25. E. Voorhees. *The Effectiveness and efficiency of agglomerative hierarchic clustering in document retrieval*. PhD thesis, Department of Computer Science, Cornell University, 1985.
26. R. Weiss, B. Velez, and M. Sheldon. Hypursuit: A hierarchical network search engine that exploits content-link hypertext clustering. In *Proceedings of the 7th ACM Conference on Hypertext and Hypermedia*, pages 180–193, 1996.
27. W. Wong and A.W. Fu. Finding structure and characteristics of web documents for classification. In *Proceedings of ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 96–105, 2000.