# Link Information as a Similarity Measure in Web Classification

Marco Cristo[1,2], Pável Calado[1], Edleno Silva de Moura[3], Nivio Ziviani[1], and Berthier Ribeiro-Neto[1]

[1] Federal University of Minas Gerais, Computer Science Department, Belo Horizonte — MG, Brazil {marco, pavel, nivio, berthier}@dcc.ufmg.br
[2] Fucapi, Technology Foundation, Manaus — AM, Brazil
[3] Federal University of Amazonas, Computer Science Department, Manaus — AM, Brazil edleno@dcc.ufmg.br

**Abstract.** The objective of this paper is to study how the link structure of the Web can be used to derive a similarity measure between documents. We evaluate five different measures and determine how accurate they are in predicting the subject of Web pages. Experiments with a Web directory indicate that the use of links from external pages greatly increases the quality of the results. Gains as high as 45.9 points in $F_1$ were obtained, when compared to a text-based classifier. Among the similarity measures tested in this work, co-citation presented the best performance in determining if two Web pages are related. This work provides an important insight on how similarity measures can be derived from links and applied to Web IR problems.

## 1 Introduction

The World Wide Web has become a main focus of research in information retrieval (IR). Its unique characteristics, like the increasing volume of data, the volatility of its documents, or the wide array of user's interests, make it a challenging environment for traditional IR solutions. On the other hand, the Web provides ground to explore a new set of possibilities. Multimedia documents, semi-structured data, user behavior logs, and many other sources of information allow a whole new range of IR algorithms to be tested. This work focuses on one such source of information widely available in the Web: its link structure.

It is possible to infer two different meanings from links between Web pages. First, if two pages are linked, we can assume that their subjects are related. Second, if a page is pointed by many other pages, we can assume that its content is important. These two assumptions have been successfully used in Web IR for tasks like page ranking [1–3], finding site homepages [4], and document classification [5–9].

In this work we evaluate how the link structure of the Web can be used to determine a measure of similarity between documents. We experiment with five different similarity measures and determine how accurate they are in predicting the subject of Web pages. We argue that a good similarity measure will be able to accurately determine if two Web documents are topic-related. Thus, we expect that such measure will be effective in classifying documents into a set of pre-defined categories.

To validate this assumption, tests were performed using a *kNN* classifier on a Web directory containing approximately 44,000 documents. Experiments show that the use of links from pages outside the directory greatly increases the quality of the results. Among the similarity measures considered in this work, the co-citation measure presents the best performance in determining if two pages are related.

These measures have never been directly compared in a Web environment and the results shown here provide an important insight on how they perform. Since they are simple to compute and use highly available information, we expect them to be applicable to several other Web related problems, such as Web document clustering [10, 11], finding similar pages [12], or building visual retrieval systems [13].

## 2 Related Work

The issue of document similarity is of central importance to Information Retrieval. Although the most widely used measure up to today is still the *cosine similarity* in the *vector space model* [14], it is known that using different approaches will influence retrieval effectiveness. For this reason, many alternatives have been proposed. Tombros and van Rijsbergen [11], for instance, show that a similarity measure that depends on the users queries may lead to better results in document clustering. In [15], Zhang and Rasmussen show that the traditional cosine similarity measure can be improved, when combined with a distance measure. And, for detecting changes in Web pages, Flesca and Masciari [16] propose a similarity measure that compares not only text, but also HTML trees.

Using link information as a way of finding related Web documents has also been proposed. One example is the Companion algorithm [17], which we describe in Sect. 3.4, where links are used to determine a set of pages related to a given initial page. He et al. [10] use link information to assign weights the edges of a graph representing hyperlink structure. Graph partitioning algorithms are then used to split the set of pages into clusters. In a work more similar to ours, three measures of linkage similarity are compared to a human evaluation of similarity between Web pages [12]. The authors come to quite different conclusions, however, mainly due to the collection used— a set of academic sites from the U.K. This collection has a very different link structure where, for instance, many of the pages link to each other, a phenomena that we cannot expect in a Web directory (or the Web in general [18]).

In this paper, we evaluate the linkage similarity measures by applying them to a classification algorithm. Several other works in the literature have reported the successful use of links as a means to improve classification performance. Using the taxonomy presented in Sun et al. [19], we can summarize these efforts in three main approaches: hypertext, link analysis, and neighborhood.

In the hypertext approach, Web pages are represented by context features, such as terms extracted from linked pages, anchor text (text describing the links), paragraphs surrounding the links, and the headlines that structurally precede the sections where links occur. Yang et al. [20] show that the use of terms from linked documents works better when neighboring documents are all in the same class. Similarly, Furnkranz et al. [21], Glover et al. [22] and Sun et al. [19] achieved good results by using anchor text, and the paragraphs and headlines that surround the links.

In the link analysis approach, learning algorithms are applied to handle both the text components in Web pages and the linkage among them. Slattery and Mitchel [6] exploit the hyperlink topology using a HITS based algorithm [2] to discover test set regularities. Joachims et al. [7] studied the combination of support vector machine kernel functions representing co-citation and content information. By using a combination of link-based and content-based probabilistic methods, Cohn et al [8] improved classification performance over a content-based baseline. Fisher and Everson [9] extended this work by showing that link information is useful when the document collection has a sufficiently high link density and the links are of sufficiently high quality.

Finally, in the neighborhood approach, the document category is estimated based on category assignments of already classified neighboring pages. Chakrabarti et al [5] showed that co-citation based strategies are better than those using immediate neighbors. Oh et al [23] improved this approach by using a filtering process to select the linked documents.

Our method is based on the link analysis approach, but it differs from previous works in the fact that our focus is not build a link based classifier, but to analyze what linkage similarity measures could be best used by a link based classifier. We evaluate a set of different approaches to extract information from the links and determine which ones provide the best results.

## 3 Linkage Similarity Measures

To determine the similarity of subject between Web pages we used five different similarity measures derived from their link structure: co-citation, bibliographic coupling, Amsler, Companion with authority degrees, and Companion with hub degrees. The first three were introduced in bibliometric science, as measures of how related two scientific papers are [24–26]. In this work, we evaluate how they perform when applied to the Web environment, where we assume that links between Web pages have the same role as citations between scientific papers. The Companion algorithm was proposed by Dean and Henzinger [17], as a method to find Web pages related to each other. Here, we use it to provide a value of similarity between documents. We now describe in detail each of the proposed linkage similarity measures.

### 3.1 Co-Citation

Co-citation was first proposed by Small [26] as a similarity measure between scientific papers. Two papers are co-cited if a third paper has citations to both of them. This reflects the assumption that the author of a scientific paper will cite only papers related to his own work. Although Web links have many differences from citations, we can assume that many of them have the same meaning, i.e., a Web page author will insert links to pages related to his own page. In this case, we can apply co-citation to Web documents by treating links as citations. We say that two pages are co-cited if a third page has links to both of them.

To further refine this idea, let $d$ be a Web page and let $P_d$ be the set of pages that link to $d$, called the *parents* of $d$. The co-citation similarity between two pages $d_1$ and $d_2$ is defined as:

$$cocitation(d_1, d_2) = \frac{P_{d_1} \cap P_{d_2}}{|P_{d_1} \cup P_{d_2}|} \tag{1}$$

Equation (1) tells us that, the more parents $d_1$ and $d_2$ have in common, the more related they are. This value is normalized by the total set of parents, so that the co-citation similarity varies between 0 and 1. If both $P_{d_1}$ and $P_{d_2}$ are empty, we define the co-citation similarity as zero.

### 3.2 Bibliographic Coupling

Also with the goal of determining the similarity between papers, Kessler [24] introduced the measure of bibliographic coupling. Two documents share one unit of bibliographic coupling if both cite a same paper. The idea is based on the notion that paper authors who work on the same subject tend to cite the same papers. As for co-citation, we can apply this principle to the Web. We assume that two authors of Web pages on the same subject tend to insert links to the same pages. Thus, we say that two pages have one unit of bibliographic coupling between them if they link to the same page.

More formally, let $d$ be a Web page. We define $C_d$ as the set of pages that $d$ links to, also called the *children* of $d$. Bibliographic coupling between two pages $d_1$ and $d_2$ is defined as:

$$bibcoupling(d_1, d_2) = \frac{C_{d_1} \cap C_{d_2}}{|C_{d_1} \cup C_{d_2}|} \tag{2}$$

According to (2), the more children in common page $d_1$ has with page $d_2$, the more related they are. This value is normalized by the total set of children, to fit between 0 and 1. If both $C_{d_1}$ and $C_{d_2}$ are empty, we define the bibliographic coupling similarity as zero.

### 3.3 Amsler

In an attempt to take the most advantage of the information available in citations between papers, Amsler [25] proposed a measure of similarity that combines both co-citation and bibliographic coupling. According to Amsler, two papers $A$ and $B$ are related if (1) $A$ and $B$ are cited by the same paper, (2) $A$ and $B$ cite the same paper, or (3) $A$ cites a third paper $C$ that cites $B$. As for the previous measures, we can apply the Amsler similarity measure to Web pages, replacing citations by links.

Let $d$ be a Web page, let $P_d$ be the set of parents of $d$, and let $C_d$ be the set of children of $d$. The Amsler similarity between two pages $d_1$ and $d_2$ is defined as:

$$amsler(d_1, d_2) = \frac{(P_{d_1} \cup C_{d_1}) \cap (P_{d_2} \cup C_{d_2})}{|(P_{d_1} \cup C_{d_1}) \cup (P_{d_2} \cup C_{d_2})|} \tag{3}$$

Equation (3) tell us that, the more links (either parents or children) $d_1$ and $d_2$ have in common, the more they are related. The measure is normalized by the total number of links. If neither $d_1$ nor $d_2$ have any children or parents, the similarity is defined as zero.

### 3.4 Companion

On a different approach, the Companion algorithm was proposed by Dean and Henzinger in [17]. Given a Web page $d$, the algorithm finds a set of pages related to $d$ by examining its link structure. Companion is able to return a degree of how related each page is to $d$. This degree can be used as a similarity measure between $d$ and other pages.

To find a set of pages related to a page $d$, the Companion algorithm has two main steps: (1) build a vicinity Graph of $d$ and (2) compute the degrees of similarity. In step 1, pages that are linked to $d$ are retrieved. We build the set $\mathcal{V}$, the vicinity of $d$, that contains the parents of $d$, the children of the parents of $d$, the children of $d$, and the parents of the children of $d$. This is the set of pages related to $d$.

In step 2 we compute the degree to which the pages in $\mathcal{V}$ are related to $d$. To do this, we consider the pages in $\mathcal{V}$ and the links among them as a graph, called the vicinity graph of $d$. This graph is then processed by the HITS algorithm [2]. The HITS algorithm returns the degree of *authority* and *hub* of each page in $\mathcal{V}$. Intuitively, a good authority is a page with important information on a given subject. A good hub is a page that links to many good authorities. In practice, the degrees of authority and hub are computed recursively: a page is a good hub if it links to many good authorities and a good authority if it is linked by many good hubs.

Once HITS is applied, we can choose to use the degree of authority or hub (or a combination of both) as a measure of similarity between $d$ and each page in $\mathcal{V}$. We define the similarity between $d$ and any page that is not in $\mathcal{V}$ as zero. In this work we experimented with the Companion algorithm using either the authority or the hub degree in isolation as a similarity measure.

For a more detailed description of the Companion and HITS algorithms, the reader is referred to [17] and [2], respectively.

## 4 The *kNN* Classifier

The measures described in Sect. 3 can be used to calculate the similarity between any two Web documents. To be useful, these measures should be able to correctly determine if two Web pages are on the same subject. In order to test this assumption, we applied them in a Web classification task.

To evaluate the linkage similarity measures, we used a strategy based on a nearest neighbor classifier. This classifier assigns a category label to a test document, based on the categories attributed to the $k$ most similar documents in the training set. The most widely used such algorithm was introduced by Yang [27] and is referred to, in this work, as *kNN*. The *kNN* algorithm was chosen since it is simple, efficient, and makes a direct use of similarity information.

In the *kNN* algorithm, to a given test document $d$ is assigned a relevance score $s_{c_i,d}$ associating $d$ to each candidate category $c_i$. This score is defined as:

$$s_{c_i,d} = \sum_{d' \in \mathcal{N}_k(d)} similarity(d, d') f(c_i, d') \tag{4}$$

where $\mathcal{N}_k(d)$ are the $k$ nearest neighbors (the most similar documents) of $d$ in the training set and $f(c_i, d')$ is a function that returns 1 if document $d'$ belongs to category $c_i$ and 0 otherwise. Traditionally, documents are represented by vectors of term weights and the similarity between two documents is measured by the cosine of the angle between them. Term weights are computed using one of the conventional TF-IDF schemes [28], in which the weight of term $t$ in document $d$ is defined as:

$$w_{d,t} = (1 + \log_2 \ f_{t,d}) \times \log_2 \frac{N}{f_t} \tag{5}$$

where $f_{t,d}$ is the number of occurrences of $t$ in document $d$, $N$ is the number of training documents, and $f_t$ is the number of training documents containing $t$. Based on the computed scores, we determine the top ranking category and assign it to the test document. This text-based version of the $kNN$ classifier was used in our experiments as the baseline for comparison.

To test the linkage similarity measures, (1), (2), (3) and the values returned by the Companion algorithm were used in place of the cosine similarity in (4). This allowed us to test all measures under the same set of conditions, and evaluate how accurate they are in predicting the subject of Web pages.

## 5 Experiments

### 5.1 The Test Collection

We performed experiments using a set of classified Web pages extracted from the Cadê Web directory [29]. This directory points to Brazilian Web pages that were classified by human experts. To obtain the content of the classified pages we used a database composed of Brazilian Web pages, crawled by the TodoBR [30] search engine.

We constructed two sub-collections using the data available on Cadê: Cade12 and Cade188. Cade12 is a set of 44,099 pages labelled using the first level categories of Cadê (Computers, Culture, Education, Health, Internet, News, Recreation, Science, Services, Shopping, Society, and Sports). Cade188 is a subset of Cade12, without the pages originally classified in the first level category. Thus, Cade188 corresponds to a set of 42,004 pages relabelled using the second level categories of Cadê (Biology, Chemistry, Dance, Music, Schools, Universities, etc.). Each Web page is classified into only one category. Figures 1, 2, and 3 show the category distributions for these collections. Notice that the two collections have skewed distributions. In Cade12, the three most popular categories represent more than 50% of all documents. The most popular category, *Services*, has 9,081 documents while the least popular, *Shopping*, has 715 documents. In Cade188, 50% of the documents are in just 10% of the categories. The most popular category, *Society:People*, has 3,675 documents while the least popular, *Internet:Tutorials*, has 24 documents. Cade12 and Cade188 have vocabularies of 192,580 and 168,869 unique words, respectively, after removing stop words.

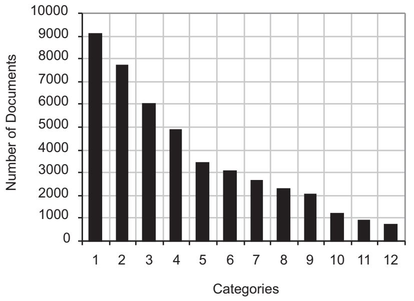


**Fig. 1.** Category distribution for Cade12.

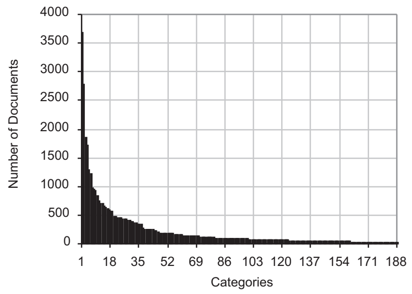


**Fig. 2.** Category distribution for Cade188.

Information about the links related to the Cadê pages was also extracted from the TodoBR collection. TodoBR provides 40,871,504 links between Web pages (an average of 6.9 links per page). We extracted from this set all the links related to the pages of our two experimental sub-collections.

Table 1 summarizes the link data obtained. It was divided into two types: the *internal links*, which are links between pages classified by Cadê, and the *external links*, which are links where the target or the source page is in TodoBR, but not in the set of pages classified by Cadê. This distinction is important to verify whether the external information provided by TodoBR can be used to improve the results.

We call *hierarchy pages* those that belong to the Cadê site itself and are used to compose the directory hierarchy. For instance, the Cadê Science page, which

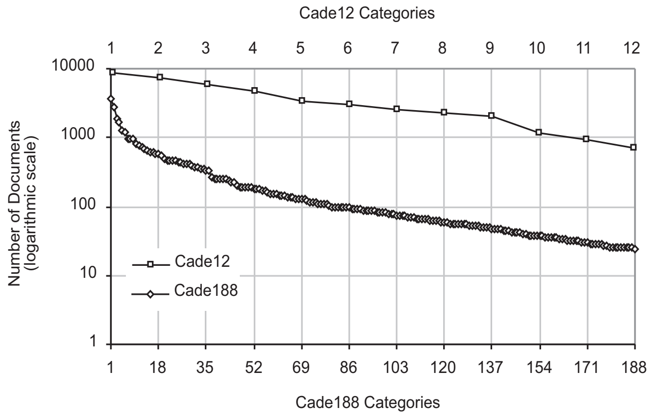


**Fig. 3.** Compared distributions for Cade12 and Cade188.

links to science related sites. As can be seen, hierarchy pages represent a great part of the internal links in the Cadê collection. Since hierarchy pages provide information on the categories of the remaining pages (for instance, the Science hierarchy page links only to science related pages), they were not used when calculating the link information measures for our experiments.

Table 1 also shows that external pages provide a rich source of link data. About 96% of the Cadê pages are linked by external pages while less than 4% link to external pages. This was an important reason for using Cadê in our experiments. With Cadê we can obtain information about external links extracted from TodoBR and verify how useful this information can be during the classification process. This is only possible because Cadê is a subset of TodoBR, which is a large collection containing most of the link information available in Brazilian Web pages. This is not the case with most other classification collections where, in order to obtain more link information, it would be necessary to collect a huge amount of Web pages, or to have access to another search engine database, as we did with TodoBR.

| Statistics | Whole Cadê | Cadê without Hierarchy Pages |
|---|---|---|
| Internal Links | 45,548 | 3,830 |
| Links from external pages to Cadê pages | 570,404 | 570,337 |
| Links from Cadê pages to external pages | 7,584 | 5,894 |
| Cadê pages with no in-links | 2,556 | 1,625 |
| Cadê pages with no out-links | 40,917 | 40,723 |

**Table 1.** Link statistics for the Cadê collection.

### 5.2 Methodology and Evaluation

To perform the experiments, we used 10-fold cross validation [31]. Each dataset was randomly split in ten parts, such that in each run, a different part was used as a test set while the remaining were used as a training set. This split on training and test sets was the same in all experiments. The final results of each experiment represent the average of the ten runs.

To make sure that the results for conventional *kNN* are not biased by an inappropriate choice of parameters, different feature selections were conducted, using term frequency and information gain [32]. To each number of features, different values for $k$ were tested. The best performance parameters were $k = 30$, feature selection by information gain, and using 15,000 features.

The performance of the presented methods was evaluated using the conventional precision, recall and $F_1$ measures [33]. Precision $p$ is defined as the proportion of correctly classified examples in the set of all examples assigned to the target class. Recall $r$ is defined as the proportion of correctly classified examples out of all the examples having the target class. $F_1$ is a combination of precision and recall in a way that gives them equal weight. $F_1$ is defined as:

$$F_1 = \frac{2pr}{p + r} \tag{6}$$

To compute the final $F_1$ values, we used macro-averaging and micro-averaging. For macro-averaging, recall, precision, and $F_1$ scores were first computed for individual categories and then averaged over all categories. For micro-averaging, the decisions for all categories were counted in a joint pool. Since the datasets used in the experiments are single label per document, micro-averaged recall, precision and $F_1$ are the same. Thus, the micro-averaged scores will be referred to as just micro-averaged $F_1$.

### 5.3 Experimental Results

Table 2 shows the $F_1$ figures for five different similarity measures obtained for the Cade12 and Cade188 collections: Amsler, bibliographic coupling, co-citation, Companion using authority degrees, and Companion using hub degrees. Only internal links were considered. As a baseline for comparison we show the results for the *kNN* classifier using the TF-IDF weighting scheme, as explained in Sect. 4.

We observe that all the results were below the baseline values. By considering only internal links, much of the link structure information of the collection is lost. In fact, as shown in Table 1, about 98% of the link information in the collection comes from external pages. This lack of information does not allow us to draw any definite conclusions.

When we make use of external links, however, results are much improved. Table 3 shows the $F_1$ figures for the Cade12 and Cade188 collections using both internal and external links. The figures for Amsler, co-citation, and the Companion algorithm using authority degrees are well above the baseline, showing

| *kNN* Similarity Measures | Cade12 | | Cade188 | |
|---|---|---|---|---|
| | $macF_1$ | $micF_1$ | $macF_1$ | $micF_1$ |
| Amsler | **16.02** | **22.44** | 4.83 | **8.87** |
| Bibliographic Coupling | 15.12 | 21.79 | 3.95 | 8.31 |
| Co-citation | 15.31 | 21.81 | 4.67 | 8.55 |
| Companion authority | 15.88 | 22.12 | **4.89** | 8.50 |
| Companion hub | 15.31 | 22.10 | 4.64 | 8.50 |
| TF-IDF (baseline) | 35.61 | 37.26 | 22.08 | 23.33 |

**Table 2.** Macro-averaged and micro-averaged $F_1$ measures obtained with the *kNN* classifier in Cade12 and Cade188 collections, using different similarity measures. Only internal links were used.

gains as high as 36.9 and 45.9 points in micro-averaged $F_1$, for the Cade12 and Cade188 collections, respectively. On the other hand, bibliographic coupling and the Companion algorithm using hub degree are still below the baseline.

| *kNN* Similarity Measures | Cade12 | | Cade188 | |
|---|---|---|---|---|
| | $macF_1$ | $micF_1$ | $macF_1$ | $micF_1$ |
| Amsler | 79.08 | 74.02 | 78.74 | 67.32 |
| Bibliographic Coupling | 15.12 | 22.08 | 4.10 | 8.55 |
| Co-citation | **79.25** | **74.12** | **78.80** | **69.24** |
| Companion authority | 74.88 | 70.44 | 73.68 | 63.51 |
| Companion hub | 22.46 | 25.45 | 9.91 | 11.40 |
| TF-IDF (baseline) | 35.61 | 37.26 | 22.08 | 23.33 |

**Table 3.** Macro-averaged and micro-averaged $F_1$ measures obtained with the *kNN* classifier in Cade12 and Cade188 collections, using different similarity measures. External and internal links were used.

These results can be explained. Since most of the links are *from* external pages *to* pages in the collection, i.e., they are from parents of the pages in the collection, we can expect measures that make use of parents to perform the best. Thus, co-citation, which uses the intersection of the sets of parents benefits greatly from such information. The same happens for the Companion algorithm using authority degrees and for the Amsler similarity. These last two, however, suffer from the fact that they also rely on children pages, which are not so widely available.

All measures show better absolute results for the Cade12 collection. This is due to the fact that the Cade12 link per class distribution is much more balanced. The number of links among documents of the same class amounts to 22.7% of the total number of internal links for Cade12, whereas it is only 10.1% of the total number of internal links for Cade188. However, the gain relative to the baseline was higher for the Cade188 collection. This happens because the *kNN* classifier tends to perform worst in collections where the class distribution is very skewed, which is the case of Cade188, as shown in Sect. 5.1.

## 6 Conclusions

In this paper, we compared five different similarity measures based on link structure. Experiments show that, in order to have sufficient information for expressive results, pages external to the test collection must be used. Also, we observe that most external pages are parents of the pages in the collection, i.e., they have a link to the pages in the collection. For this reason, the co-citation similarity measure obtained the best results. Other measures, such as the Amsler similarity and the Companion algorithm using authorities, also show good results but are, however, affected by the fact they use out-link information, which is much scarcer.

We expect the most popular pages in the Web to be those with a high number of in-links [1,2]. These pages will also be the most interesting for Web directories, where it is preferable (and easier) to populate the hierarchy with a reasonable set of highly referenced sites, instead of a huge set of obscure pages. Thus, similarity measures that make use of in-link information are expected to be the most appropriate. This conclusion is reinforced by that fact that, although most Web pages have very few links (or no links at all), those that are highly linked have much more in-links than out-links.

The difference between results using internal or external link information also confirms that the effectiveness of the proposed measures depends highly on the link structure, as also stated in [9]. Thus, on subsets of the Web with a very different link structure, similarity measures other than co-citation may show a better performance.

Although in this work we are only evaluating link-based similarities, the contents of Web pages is a valuable source of information and should not be disregarded. The combination of content and link-based information has been shown to yield good results [3, 7], and we intend to pursue it in future work. Experiments were already initiated where the linkage similarity measures here tested are combined with content-based information. Preliminary results show that this can lead to further improvements.

Since the measures here presented were shown effective in determining the subject of Web pages, we can expect them to perform well in other IR tasks where document similarity is an important concept, such as, finding similar pages, Web page clustering, information filtering, among others. Experiments with these tasks are left for future work.

## 7 Acknowledgements

This work was supported in part by the I3DL project—grant 680154/01-9, the GERINDO project—grant MCT/CNPq/CT-INFO 552.087/02-5, the SIAM project—grant MCT/FINEP/CNPq/PRONEX 76.97.1016.00, by CNPq grant

520.916/94-8 (Nivio Ziviani), and by MCT/FCT scholarship grant SFRH/BD/-4662/2001 (Pável Calado).

A special thanks goes to Marcos André Gonçalves, for his good ideas and help in tuning up the classification algorithms.

## References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. In: Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia (1998) 107–117
2. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM) **46** (1999) 604–632
3. Calado, P., Ribeiro-Neto, B., Ziviani, N., Moura, E., Silva, I.: Local versus global link information in the Web. ACM Transactions On Information Systems **21** (2003) 42–63
4. Hawking, D., Craswell, N.: Overview of TREC-2001 Web track. In: The Tenth Text REtrieval Conference (TREC-2001), Gaithersburg, Maryland, USA (2001) 61–67
5. Chakrabarti, S., Dom, B.E., Indyk, P.: Enhanced hypertext categorization using hyperlinks. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, Seattle, Washington, USA (1998) 307–318
6. Slattery, S., Craven, M.: Discovering test set regularities in relational domains. In: Proceedings of ICML-00, 17th International Conference on Machine Learning, Stanford, California, USA (2000) 895–902
7. Joachims, T., Cristianini, N., Shawe-Taylor, J.: Composite kernels for hypertext categorization. In: Proceedings of ICML-01, 18th International Conference on Machine Learning, Williamstown, Massachusetts, US (2001) 250–257
8. Cohn, D., Hofmann, T.: The missing link - a probabilistic model of document content and hypertext connectivity. In Leen, T.K., Dietterich, T.G., Tresp, V., eds.: Advances in Neural Information Processing Systems 13, MIT Press (2001) 430–436
9. Fisher, M., Everson, R.: When are links useful? Experiments in text classification. In: Proceedings of the 25th annual European conference on Information Retrieval Research, ECIR 2003, Pisa, Italy (2003) 41–56
10. He, X., Zha, H., Ding, C.H.Q., Simon, H.D.: Web document clustering using hyperlink structures. Computational Statistics & Data Analysis **41** (2002) 19–45
11. Tombros, A., van Rijsbergen, C.J.: Query-sensitive similarity measures for the calculation of interdocument relationships. In: Proceedings of the 10th International Conference on Information and Knowledge Management CIKM, Altlanta, Georgia, USA (2001) 17–24
12. Thelwall, M., Wilkinson, D.: Finding similar academic Web sites with links, bibliometric couplings and colinks. Information Processing & Management (2003) (in press).
13. Olsen, K.A., Korfhage, R.R., Sochats, K.M., Spring, M.B., Williams, J.G.: Visualization of a document collection: the VIBE system. Information Processing & Management **29** (1993) 69–81
14. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill (1983)

15. Zhang, J., Rasmussen, E.M.: Developing a new similarity measure from two different perspectives. Information Processing & Management **37** (2001) 279–294
16. Flesca, S., Masciari, E.: Efficient and effective Web change detection. Data & Knoweledge Engeneering **46** (2003) 203–224
17. Dean, J., Henzinger, M.R.: Finding related pages in the World Wide Web. Computer Networks **31** (1999) 1467–1479 Also in Proceedings of the 8th International World Wide Web Conference.
18. Kumar, S.R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., Upfal, E.: The Web as a graph. In: Proceedings of the 19th Symposium on Principles of Database Systems, Dallas, Texas, USA (2000) 1–10
19. Sun, A., Lim, E.P., Ng, W.K.: Web classification using support vector machine. In: Proceedings of the Fourth International Workshop on Web Information and Data Management, McLean, Virginia, USA, ACM Press (2002) 96–99
20. Yang, Y., Slattery, S., Ghani, R.: A study of approaches to hypertext categorization. Journal of Intelligent Information Systems **18** (2002) 219–241
21. Furnkranz, J.: Exploiting structural information for text classification on the WWW. In: Proceedings of the 3rd Symposium on Intelligent Data Analysis (IDA-99), Amsterdam, Netherlands (1999) 487–498
22. Glover, E.J., Tsioutsiouliklis, K., Lawrence, S., Pennock, D.M., Flake, G.W.: Using Web structure for classifying and describing Web pages. In: Proceedings of WWW-02, International Conference on the World Wide Web, Honolulu, Hawaii, USA (2002)
23. Oh, H.J., Myaeng, S.H., Lee, M.H.: A practical hypertext catergorization method using links and incrementally available class information. In: Proceedings Of The 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece (2000) 264–271
24. Kessler, M.M.: Bibliographic coupling between scientific papers. American Documentation **14** (1963) 10–25
25. Amsler, R.: Application of citation-based automatic classification. Technical report, The University of Texas at Austin, Linguistics Research Center, Austin, Texas, USA (1972)
26. Small, H.G.: Co-citation in the scientific literature: A new measure of relationship between two documents. Journal of the American Society for Information Science **24** (1973) 265–269
27. Yang, Y.: Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland (1994) 13–22
28. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management **24** (1988) 513–523
29. The *Cadê?* Web directory. `http://www.cade.com.br/`
30. The TodoBR search engine. `http://www.todobr.com.br/`
31. Stone, M.: Cross-validation choices and assessment of statistical predictions. Journal of the Royal Statistical Society **B36** (1974) 111–147
32. Mitchell, T.: Machine Learning. McGraw-Hill (1997)
33. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, California, USA (1999) 42–49