A Comparative Study of Citations and Links in Document Classification

Thierson Couto[1] Marco Cristo[1] Marcos André Gonçalves[1]
Pável Calado[2] Nivio Ziviani[1] Edleno Moura[4] Berthier Ribeiro-Neto[1 3]

Federal University of Minas Gerais
[1]Computer Science Department
Belo Horizonte, Brazil
{thierson,marco,mgoncalv, nivio,berthier,}@dcc.ufmg.br

[2]IST/INESC-ID
Lisboa, Portugal
{pavel.calado}@tagus.ist.utl.pt

[3]Google Engineering
Belo Horizonte, Brazil
{berthier}@google.com

Federal University of Amazonas
[4]Computer Science Department
Manaus, Brazil
{edleno}@dcc.ufam.edu.br

ABSTRACT

It is well known that links are an important source of information when dealing with Web collections. However, the question remains on whether the same techniques that are used on the Web can be applied to collections of documents containing citations between scientific papers. In this work we present a comparative study of digital library citations and Web links, in the context of automatic text classification. We show that there are in fact differences between citations and links in this context. For the comparison, we run a series of experiments using a digital library of computer science papers and a Web directory. In our reference collections, measures based on co-citation tend to perform better for pages in the Web directory, with gains up to 37% over text based classifiers, while measures based on bibliographic coupling perform better in a digital library. We also propose a simple and effective way of combining a traditional text based classifier with a citation-link based classifier. This combination is based on the notion of classifier reliability and presented gains of up to 14% in micro-averaged F1 in the Web collection. However, no significant gain was obtained in the digital library. Finally, a user study was performed to further investigate the causes for these results. We discovered that misclassifications by the citation-link based classifiers are in fact difficult cases, hard to classify even for humans.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.5.3 [**Pattern Recognition**]: Applications—*Text processing*

General Terms

Algorithms, Experimentation

Keywords

Web directories, digital libraries, text classification, links

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
JCDL'06, June 11–15, 2006, Chapel Hill, North Carolina, USA.
Copyright 2006 ACM 1-59593-354-9/06/0006 ...$5.00.

1. INTRODUCTION

Traditional document classification techniques usually rely on textual features, like terms and expressions extracted, from documents. However, Web pages and documents in digital libraries often present other components that can be explored by classifiers, one of the most interesting ones being the citations or links between them. In the Web we have explicit *links* between Web pages and in digital libraries we have *citations* (or references) between scientific documents. In both cases, a citation-link is an explicit statement of one author that his or her document is somehow related to another document. Link information has been used in document classification [5, 9, 11] and is specially important when documents are noisy or contain little text, a circumstance where pure traditional content based techniques are known to perform poorly [6].

In this work we investigate similarities and differences of the role of citations and links applied to the task of document classification. Our motivation is based on previous success in the use of link based information in a Web classification task [5, 9], as well as unanswered questions raised by these previous works. One of our hypotheses is that citations and links have both similarities and differences in the way they are used, which can impact algorithms and results. In particular, we hypothesize that links are noisier than citations, and therefore, citations may provide better results than links when used within similar algorithms.

Experiments performed over a Web directory and a digital library confirm some of these hypotheses, while disproving others. We found that measures derived from citations and links can be used to learn reliable and effective text classi-

fiers based on the kNN model. By reliable we mean that when the classifier assigns a class to a document with high probability, the class is the correct one most of the time. Conversely, if the classifier assigns a class to a document with low probability,the class is generally incorrect most of the time. By effective we mean that experiments performed with 10 fold cross validation have reached values of macro and micro F1 superior to state-of-the-art text based classifiers, in both collections.

There are also similarities in the failures of classifiers based on citations and based on links. Failures occur when a test document shares more links with training documents belonging to the wrong classes than with those of the correct class. We suspected that these cases are hard even for human beings to classify, since the test documents afford to have some kind of strong relation to documents of the wrong class. In order to confirm this hypothesis we conducted a user study, asking volunteers to classify a random sample of these supposedly difficult cases. The experiment shows that most cases are in fact difficult and that there is little consensus among human classifiers regarding the correct class of a same document.

We also found differences between citations and links with respect to the best measures to be used. Web pages in directories are better classified with the co-citation similarity measure, while documents in digital libraries are better classified with the bibliographic coupling and Amsler measures. We discuss further about this fact and explain why this happens in Section 5.

Combination of evidence is a well known method to boost classification performance. We revisit the topic here in the light of our new results. We find that features of the environment (i.e., the collections) in which this information occurs have important implications for algorithms that try to combine the citation-link information with more traditional text based approaches. A simple method based on the reliability of these evidence performed very well in the Web Cadê directory collection, but it has provided only a marginal gain in the digital library collection. Reasons for this behavior are discussed in Section 8.

2. RELATED WORK

Citations among documents were first used as a source of information in bibliometric science. In 1963, Kessler introduced the notion of bibliographic coupling [17], a measure that can be used to determine documents with similar topics. Later, the measure of co-citation was introduced by Small in [26]. Both measures have been used as complementary sources of information for document retrieval and classification [1, 2, 23] and as a means to evaluate the importance of scientific journals [13].

The ideas used for citations among documents were later naturally transposed to the Web environment. However, several distinctions must be made between the Web and the domain of scientific publications. For instance, unlike web pages, papers are peer reviewed, thus ensuring the referencing of other important papers on the same subject. Also, Web links can have different functions, such as site navigation. The adaptation of these bibliometric techniques to the web environment, has given rise to several algorithms for improving IR tasks in the Web.

Seminal work was done by Brin and Page [3] and Kleinberg [18], who proposed algorithms capable of using the Web's link structure to derive a measure of importance for web pages. Such measure could then be applied in document ranking, greatly improving on the results achieved by traditional text-based methods. Once discovered the richness, and effectiveness, of the available link information, many other approaches followed. These approaches were not only dedicated to ranking web pages, but also to IR tasks that included finding topic related documents [10], discovering web communities [19], or classifying Web pages [27].

The work presented here focuses precisely on the problem of document classification. Classification of web pages using links has already deserved a wide attention from the IR community. Several works explored the Web's link structure through techniques such as using link anchor text to improve the target page description [12,14,29], applying machine learning algorithms to exploit patterns present in the link structure [8, 11, 16], and using the Web's link graph as support for the classification of neighboring pages [6, 22].

In this paper, we apply similar techniques to classify documents, not only on the Web, but also on a collection of computer science papers. However, instead of focusing on the effectiveness of the classification method, we intend to study how the different characteristics of links and citations can impact the method's results. We analyze several distinct link- and citation-based similarity measures and determine which ones provide the best results in each environment. Following, we evaluate how effective these measures are in improving the results of a text-based classifier.

This paper continues and extends preliminary work presented in [5] and [9] by including a more detailed set of experiments and more analytical, quantitative user-based studies. All of these allowed us to reach better conclusions regarding the use of citation-link based similarity measures in document classification

3. DATASETS

In this section we describe the citation and link collections used in our experiments. Section 3.1 describes ACM8, a subcollection of the ACM digital library. Section 3.2 describes the Cade12, a collection of Web pages derived from the the Cadê directory.

3.1 Citations: the ACM Digital Library

Our study of citations was performed with the ACM8 collection, a sub-collection of the ACM Digital Library[1]. All the text contained in the title and abstract, when available, was used to index the documents. Notice that many citations in the original ACM Digital Library could not be traced to the corresponding paper for a number of reasons. Among them, the fact that many cited papers do not belong to this digital library and also due to the imprecise process used to match the citation text to the corresponding paper [20]. High precision and recall in this pre-processing phase is hard to be achieved for many reasons, including differences in the writing style for the names of authors and conferences in the citations. This problem is particularly exacerbated in the case of the ACM Digital Library, since most citations were obtained with OCR after scanning with the introduction of many errors, making the matching process even harder.

[1] `http://portal.acm.org/dl.cfm`

To simulate a more realistic situation in which most citations are available, we selected a subset of the ACM Digital Library having only documents with at least four out-citations. This is a very reasonable assumption since most papers of the ACM Digital Library (even short ones) have more than four citations. In fact, the average number of citations in the ACM Digital Library is 11.23.

The resulting ACM8 collection is a set of 6,680 documents, without stop words, labeled under the 8 first-level categories of the ACM Digital Library taxonomy: *Hardware*, *Computer Systems Organization*, *Software*, *Theory of Computation*, *Mathematics of Computing*, *Information Systems*, *Computing Methodologies*, and *Computing Milieux*). Classes *General Literature*, *Data* and *Computer Applications* of the taxonomy were not used because they have less than 20 documents in this sub-collection. Similarly to our Web collection, each paper is classified into only one category.

Figure 1 shows the category distributions for the ACM8 collection. Notice that the ACM8 collection also has a very skewed distribution. The two most popular categories represent more than 50% of all documents.

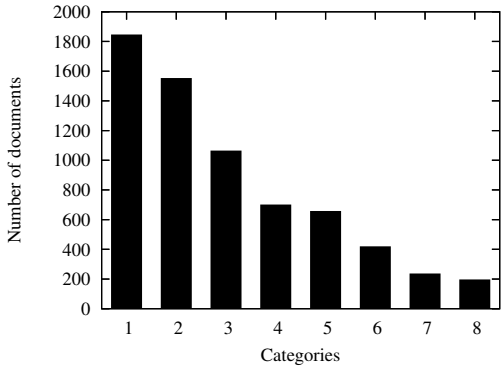

Figure 1: Category distribution for the ACM8 collection.

Statistics	ACM8
Internal citations	11,510
Citations from ACM8 documents to external documents	40,387
Number of ACM8 with no in-citations	1,941
Average of in-citations by document	4.72
Average of out-citations by document	7.77

Table 1: Statistics for the ACM8 collection.

Table 1 shows some statistics about citations in the ACM8 collection. Citations from ACM8 articles to external documents correspond to 77.8% of the citations in the collection. These are citations to documents that belong to the ACM Digital Library but were not included in the ACM8 collection and also to documents that do not take part in the ACM Digital Library[2]. Since we have no citation information about the external documents, in-citations can be derived only from internal citations, while out-citations can be derived from all citations. Thus the number of in-citations in the ACM8 collection is 11,510, while the number of out-citations is almost four times higher.

[2]Information about these documents came from the DBLP(http://dblp.uni-trier.de/) collection.

Figure 2 shows the distribution of in-citations and out-citations for the ACM8 collection. It can be seen that the majority of documents has less in-citations than out-citations.

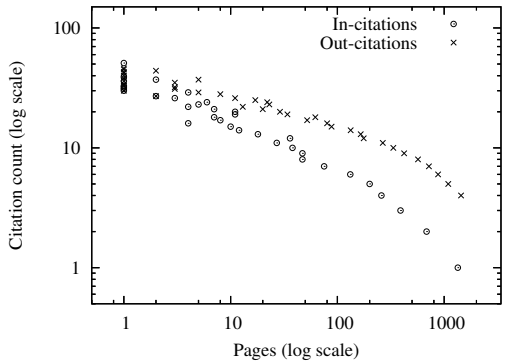

Figure 2: Citation distribution for the ACM8 collection.

3.2 Links: the Cade12 Collection

We used in our experiments a collection of pages indexed by the Brazilian Web directory Cadê[3], referred to as Cade12 collection. All the pages in the Cadê directory were manually classified by human experts. Since they were also indexed by the TodoBr search engine [25][4], we built the Cade12 collection by obtaining text and links directly from the TodoBr database. The content of each document in Cade12 collection is composed of the text contained in the body and title of the corresponding Web page, discarding HTML tags.

The resulting collection is composed of 44,099 documents, containing a vocabulary of 191,962 distinct terms. But in our experiments we used only the 10,000 terms with best infogain. The documents were labeled under 12 first-level classes of the Cadê directory: *Computers*, *Culture*, *Education*, *Health*, *News*, *Internet*, *Recreation*, *Science*, *Services*, *Shopping*, *Society*, and *Sports*. Figure 3 shows the category distribution for the Cade12 collection. Notice that the collection has a skewed distribution and the three most popular categories represent more than half of all documents.

The link information of the Cade12 collection was extracted from the set of 40,871,504 links of the TodoBR database. As observed by the authors in [4], the richer the link information considered, the better the accuracy obtained by link based classifiers. In fact, this was an important reason for choosing Cadê. With Cadê we are not restricted to a limited source of links since Cadê is a subset of TodoBR, which is a large collection containing most of the link information available in Brazilian Web pages.

Notice that pages belonging to the Cadê site itself are used to compose the directory hierarchy. For instance, the Cadê Science page is a directory page, which links to science related pages indexed by Cadê. We do not use these pages for calculating the link information measures in our experiments, because they provide information on the categories of the remaining pages and could cause a bias in the results. For the same reason we do not use pages found in

[3]`http://www.cade.com.br/`
[4]`http://www.todobr.com.br/`

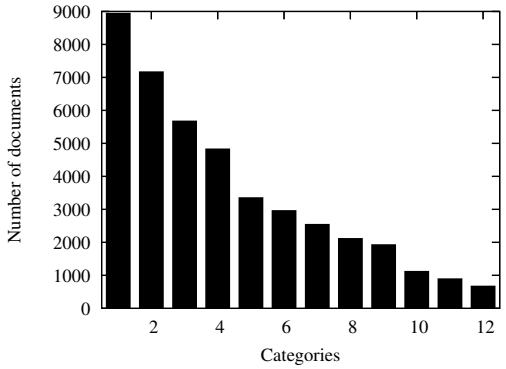

Figure 3: Category distribution for the Cade12 collection.

the TodoBR collection *similar* to Cadê pages. We consider a page in TodoBR similar to a page in Cadê if they share 70% or more of their out-links. Pages from directories other than Cadê were also discarded since these share many out-links with Cadê. These pages had gone unnoticed in previous works, which caused a slight increase in the precision of the results for the link based classifiers.

Table 2 shows statistics about link information of the Cade12 collection without considering directory pages.

Statistics	Cadê without directory pages
Internal Links	3,830
Links from external pages to Cadê pages	554,592
Links from Cadê pages to external pages	5,894
Cadê pages with no in-links	4,392
Cadê pages with no out-links	40,723
Mean of in-links by document	12.57
Mean of out-links by document	0.13

Table 2: Link statistics for the Cade12 collection.

Figure 4 presents the distribution of in-links and out-links in the Cade12 collection. Notice that most pages have no out-links at all, but the majority does have in-links.

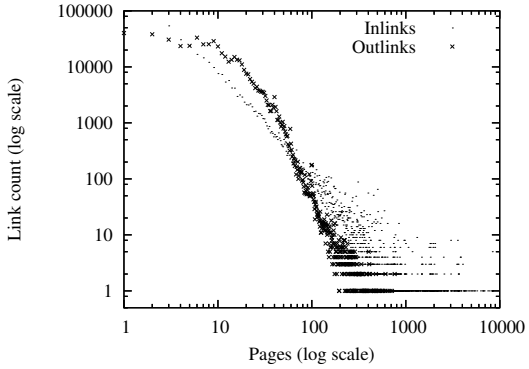

Figure 4: Link distribution for the Cadê collection.

4. BIBLIOMETRIC SIMILARITY MEASURES

In this section we present the bibliometric similarity measures we used with both citation based and link based classifiers: Co-citation [26], Bibliographic coupling [17], and Amsler [1].

Let d be a web page and let P_d be the set of pages that cites (or link to) d, called the *parents* of d. The co-citation similarity between two pages d_1 and d_2 is defined as:

$$\text{co-citation}(d_1, d_2) = \frac{P_{d_1} \cap P_{d_2}}{|P_{d_1} \cup P_{d_2}|} \quad (1)$$

Eq. (1) shows 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.

We define C_d as the set of pages that d cites (or links to), also called the *children* of d. Bibliographic coupling between two pages d_1 and d_2 is defined as:

$$\text{bib-coupling}(d_1, d_2) = \frac{C_{d_1} \cap C_{d_2}}{|C_{d_1} \cup C_{d_2}|} \quad (2)$$

According to Eq. (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.

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:

$$\text{Amsler}(d_1, d_2) = \frac{(P_{d_1} \cup C_{d_1}) \cap (P_{d_2} \cup C_{d_2})}{|(P_{d_1} \cup C_{d_1}) \cup (P_{d_2} \cup C_{d_2})|} \quad (3)$$

Eq. (3) tell us that, the more citations or 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.

Since bibliometric measures are functions that map pairs of documents into real numbers, we can obtain a doc-doc matrix for each of the bibliometric measure defined. Notice that because of the definitions of the similarity measures, all diagonal values of these matrices are equal to 1. Bibliometric similarity matrices are used as input to the citation-link based classifiers, that are presented in next section.

5. EXPERIMENTS

For each collection studied we developed several classifiers based on the three bibliometric similarities defined above, which we call citation-link based classifiers. These were compared to traditional text-only classifiers. In both cases, we learned classifiers based on the k Nearest Neighbor (kNN) and the Support Vector Machine (SVM) methods.

A kNN classifier assigns a class label to a test document based on the classes attributed to the k most similar documents in the training set, according to some similarity measure. In the kNN algorithm [28], each test document d is assigned a score s_{d,c_i}, which is defined as:

$$s_{d,c_i} = \sum_{d_t \in N_K(d)} = similarity(d, d_t) \times f(c_i, d_t), \quad (4)$$

where $N_k(d)$ are the k neighbors (the most similar documents) of d in the training set and $f(c_i, d_t)$ is a function that returns 1 if the training document d_t belongs to class c_i and 0 otherwise. The classifier assigns to test document d the class c_i with the highest score.

In our experiments we learned kNN classifiers for each bibliometric similarity measure by substituting the similarity $similarity(d, d_t)$ function in Eq. (4) for the value of the pair (d, d_t) in the corresponding bibliomentric similarity matrix.

Text-based kNN classifiers were learned using the cosine measure as the similarity function. With the cosine similarity measure each document is considered a vector of term weights and the measure corresponds to the cosine of the angle between the two vectors. We use TF-IDF [24] as the weight of a term t in a document d, defined as:

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

where $f_{t,d}$ is the number of times term t occurs in document d, N is the number of training documents, and f_t is the number of documents that contain term t.

We experimented with different values for k. Since values greater than 30 did not cause any significant change in the results we fixed k equal to 30 in all kNN classifiers we used.

The SVM classifier [15] works over a vector space where the problem is to find a hyper-plane with the maximal margin of separation between two classes. In our experiments we learned SVM classifiers with the *Radial Basis Function*(RBF) Kernel, using the SVMLIB software [7].

The input to each of our bibliometric SVM classifiers was derived directly from the corresponding bibliometric similarity matrix. Each document d_j related to a document d in the matrix is considered a feature of d in the input of SVM classifier, and the value of this feature is the value of the bibliometric similarity of the pair (d, d_j) in the similarity matrix.

In all classification experiments we used 10-fold cross validation and we evaluate each run using macro and micro F_1 measures. The final results of each experiment represent the average of the ten runs for both measures.

5.1 The ACM8 Collection

Table 3 presents the micro-averaged and macro-averaged F_1 values for the citation-link and text based classifiers over the ACM8 collection.

Method	Similarity	$micF_1$	$macF_1$	Gains (%) over text classifier	
				$micF_1$	$macF_1$
kNN	co-citation	61.60	52.56	-20	-25.5
	bib-coupling	83.20	78.29	8.1	10.9
	Amsler	84.43	79.41	9.7	12.5
	Cosine	76.95	70.57	–	–
SVM	co-citation	59.33	49.98	-24.3	-32.2
	bib-coupling	80.72	74.59	2,9	1.0
	Amsler	83.08	77.08	5.93	4.5
	TF-IDF	78.43	73.77	–	–

Table 3: Macro-averaged and micro-average F_1 results for the ACM8 collection for kNN and SVM classifiers.

We take the results of text based classifiers of each classification method as the baseline for the method. The two last columns of the table show the percentage of gain (or loss) of each classifier over the content classifier for the method.

In citation classifiers, co-citation presented the worst overall results over all the classifiers. Since co-citation is a measure of the number of in-citations two documents have in common and most of the documents of the ACM8 collection have few in-citations, co-citation is not sufficiently precise for the classifier to decide the class of a test document. In fact, 85% of the documents that the kNN with co-citation failed to classify have less than 4 in-citations, and 73% of the mistakes were for documents with one or zero in-citations. On the other hand, of the 61.75% of documents that kNN with co-citation correctly classified, only 28% of them have zero[5] or one in-citation. So we can presume that the problem of kNN with co-citation is not due to the co-citation measure itself but to the lack of in-citations in the collection.

Citation classifiers using the Amsler similarity were the best performers for both methods. However, results are only slightly better than for bib-coupling. Since the Amsler similarity is a kind of combination between co-citation and bib-coupling, we can conclude that bib-coupling contributed most to the results. Also, because bib-coupling is a measure of similarity between two documents based on the out-citations they have in common, we can also conclude that out-citation evidence was the one responsible for the best results. We note also that the text in documents of the ACM8 collection, despite being short, seems not to be noisy, since content base classifiers also presented a good performance.

Finally, Table 3 also shows that citation information is better used to learn classifiers based on the kNN method, while textual information is better used with SVM.

5.2 The Cade12 Collection

The same set of experiments applied on the ACM8 collection using citation based classifiers was also applied to the Cade12 collection using link based classifiers. The results are shown in Table 4.

Method	Similarity	$micF_1$	$macF_1$	Gains (%) over text classifier	
				$micF_1$	$macF_1$
kNN	co-citation	68.51	75.60	36.9	51.1
	bib-coupling	22.09	5.39	-55.8	-87.9
	Amsler	68.56	75.53	37.0	50.9
	Text cosine	50.03	44.50	–	–
SVM	co-citation	68.91	76.9	27.2	55.7
	bib-coupling	24.08	6.40	-55.6	-87.0
	Amsler	68.09	74.8	25.6	55.5
	Text	54.18	49.38	–	–

Table 4: Macro-averaged and micro-average F_1 results for the Cade12 collection for kNN and SVM classifiers. kNN Text Cosine and SVM Text are used as baselines for comparison.

Contrary to the ACM8 collection, for link based classifiers the bib-coupling measure presented the worst results in Cade12, whereas co-citation performed better. The reason is that the there are much more in-links than out-links in Cade12 collection, as shown in Figure 4. In fact, 99%

[5]The implementation of the kNN used assigns the most frequent class in a collection to a test document when the test document has no neighbors.

of the documents that kNN misclassified using bib-coupling measure have no out-links at all. However, overall accuracy increases by 50% when we consider only pages with three or more out-links. Thus, we conclude that the weak overall performance of this classifier is due to the lack of out-links in Cade12. The performances of the classifiers based on Amsler and co-citation measure are very similar. This is not surprising since the Amsler measure reduces to the co-citation measure in the case where very little out-link information is available. In sum, the co-citation measure is the best in the Cade12 collection due to the rich in-link information present.

In the Cade12 collection, text based classifiers presented a very weak performance, especially if compared to its performance in the ACM8 collection. This is due to the fact that Web pages are usually noisy and contain little text, lacking coherence in style, language, and structure. These problems are less common in the papers of a digital library. As a result, the quality of the text evidence in the ACM8 collection is better than in the Cade12 collection.

As the above experiments show, classification in a digital library works better for bib-coupling, while classification of pages indexed by a Web directory works better for co-citation. This happens because the characteristics of the documents and the way such collections are organized affect the number of in-links (or in-citations) and out-links (or out-citations) in each kind of collection. In a digital library, almost all articles cite many other articles inside or outside the digital library boundary. However, only citations to external documents are available, while citations from external documents are not. Also the number of articles inside the collection that are cited by many other articles is comparatively few, since the number of authoritative articles in a collection is scarce. Data from the ACM8 collection, shown in Table 1 and Figure 2, corroborate the above conclusion. Thus out-citations tend to be more frequent than in-citations, which explains why bibliometric measures based on out-citations (bib-coupling and Amsler) tend to work better than co-citation.

The opposite happens with a collection of pages indexed by a Web directory. In this case, the distributions of in-links and out-links are very distinct with many pages having zero or one out-links (about 90% of the Cade12 collection has zero out-links), while only few pages have no in-links at all (approximately 6% of pages in the Cade12 collection have no in-links). This leads us to conclude that pages indexed by a directory tend to be authoritative pages when compared to other pages of the Web. A reasonable explanation to this characteristic is that, when a page is indexed by a directory, it becomes more visible by Web users and consequently new pages tend to create links to it. Thus the number of in-links tends to increase while the number of out-links tends to be constant. This explains why the differences on the performance of co-citation and bib-coupling is much more acute in the Cade12 collection than in the ACM8 collection. As a result, co-citation works better than bib-coupling in Web directories.

There are cases in both collections, however, where a test document may have many in-citations (or in-links) or out-citations (or out-links) and the classification will fail. These are considered the difficult cases and are studied in Section8.

Also notice that, according to Table 4, the SVM method was slightly better than the kNN method when we consider only citation-link measures in Cade12. However, neither SVM nor kNN presented clear performance advantage over each other when considering both collections. Since kNN is simpler and faster than SVM, it will be the method used in our experiments reported in the following sections.

6. CONFIDENCE ON CITATIONS AND LINKS

We have seen that our citation-link based classifiers are very effective, but can we trust their predictions, or, in other words, how reliable are they? Many applications rely on the scores provided by classifiers to determine how suitable is a document to a class. This is also important if we want to somehow combine the outputs of these classifiers with, for example, text based classifiers to improve performance. For such applications, an ideal classifier should provide belief estimates exactly proportional to their actual performance. In other words, given a set of documents $\mathcal{D}_p$, for which the ideal classifier assigns class labels with probability p, it should classify correctly $p \times |\mathcal{D}_p|$ documents of the set $\mathcal{D}_p$. In spite of not being an ideal classifier, the link-based kNN studied in this work presents the property of providing belief estimates proportional to its accuracy in the collections ACM8 and Cade12.

Such property can be observed in Figures 5 and 6. These figures show the accuracy values obtained for different belief degrees estimated by the kNN classifiers learned with bib-coupling and co-citation in the ACM8 collection and in the Cade12 collection. They also show the trend line, derived by means of linear regression, for the accuracy values obtained by the classifiers as well as the ideal values in which the belief degree would correspond exactly to the accuracy obtained.

We notice in all figures, except the bib-coupling graphic for Cade12, in Figure 6, that the trend line for the distribution of belief degrees is very similar to the line corresponding to the ideal values. This implies that, in general, the values provided as belief degrees correspond approximately to the accuracy obtained by the classifier. Thus, we can take these values as good estimates of how many documents will be assigned to the correct classes.

The only exception occurs for bib-coupling in the Cade12 collection, where the trend line clearly differs from the ideal line. This reflects the fact that this measure yields less reliable estimations than any other measure due to the lack of out-links in the Cade12 collection.

Cases where classifiers provide unreliable estimations are not so uncommon. In fact, providing good probability estimations is not an easy task. As observed by [21], this is due to the fact that there is no perfect classifier and, as we will show in Section 8, document classification can involve much controversy. Even human experts can disagree about how suitable is a classification.

7. EVIDENCE COMBINATION

Combination of evidence is a well known method to boost classification performance. It is specially useful if the estimations provided by the classifiers to be combined are based on independent evidence. Nevertheless, a direct combination of belief degrees may produce improper values if the estimations provided by the classifiers are unrealistic or are represented by numbers in very different scales [4].

However, if one of the classifiers to be combined presents

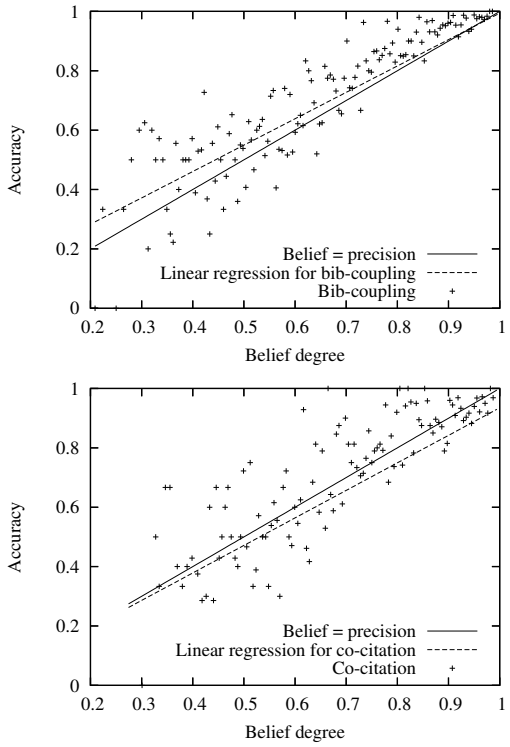

Figure 5: Accuracy per belief degree for bib-coupling and co-citation in the ACM8 collection.

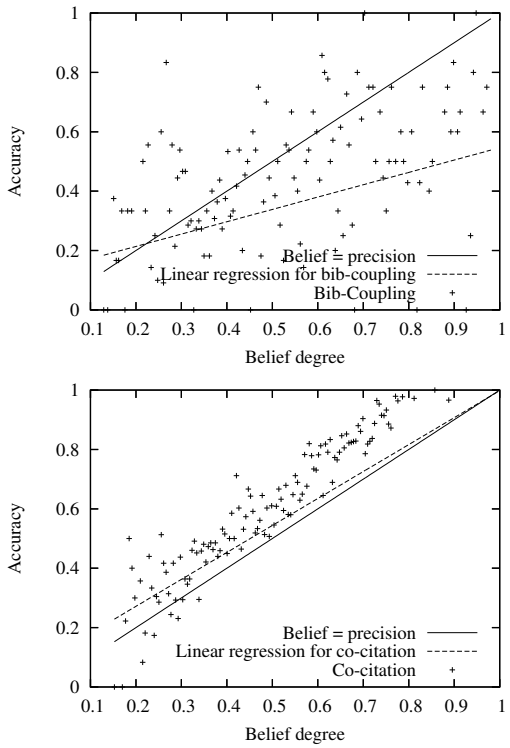

Figure 6: Accuracy per belief degree for bib-coupling and co-citation in the Cade12 collection.

high accuracy and provides reliable estimations it is possible to use it as a guide in the combination process. In other words, in the cases where the more reliable classifier assigns a document to a category with low confidence (low belief degree) we can expect it to be wrong (low accuracy). Thus, in such cases, it would be better to use the classification decision provided by the second classifier. This idea is formally presented in Listing 1.

Listing 1: Combining pieces of evidence.

```
1  Let A be the most reliable classifier to be combined;
2  Let B be the least reliable classifier to be combined;
3  Let A_tr be a set of points {c_Ai, y_Ai}, where c_Ai,
       0 ≤ c_Ai ≤ 1, represents the confidence of A in the
       classification given for document i in the
       training collection and y_Ai is 1 if the
       classification provided by A for i is correct and
       is 0 otherwise;
4  Let B_tr be a set of points {c_Ai, y_Bi}, where y_Bi is 1
       if the classification provided by B for i is
       correct and is 0 otherwise;
5  Let f_A(x) = b + ax be the function that best fits the
       points in A_tr;
6  Let f_B(x) = d + cx be the function that best fits the
       points in B_tr;
7  if (a == c) {
8     if (b > d)
9        p = 0;
10    else
11       p = 1;
12 } else
13    p = (b−d)/(c−a);
14 for each document i in the test collection {
15    if (c_Ai > p)
16       classification of document i is given by A;
17    else
18       classification of document i is given by B;
19 }
```

The algorithm in Listing 1 first tries to find the degree of belief p from which the most reliable classifier A tends to be always better than the least reliable classifier B (lines 1-13). For this, it obtains accuracy trend lines for A (lines 3 and 5) and B (lines 4 and 6), according to the belief degrees of A. It then finds the point p where the lines cross each other (lines 7-13) and uses this point to determine which classifiers provide the best decisions (lines 14-19). In sum, decisions from classifier A are preferable if it estimates beliefs greater than p.

To illustrate this, Figure 7 shows the accuracy trend lines obtained by text based and citation based classifiers in collection ACM8 according to the belief degree estimated by the citation based classifiers. Figure 8 shows the figures obtained by text based and link based classifiers for the Cade12 collection. The lines shown in the figures were derived, by means of linear regression, from the accuracy values provided by the classifiers. For both collections, the citation-link based classifiers were used as guides since they present better accuracy (see Tables 3 and 4).

Notice that in the Cade12 collection the classifier performances are very distinct. Their respective accuracy lines cross each other at point p corresponding to a belief degree of about 50%. When the link based belief degree falls be-

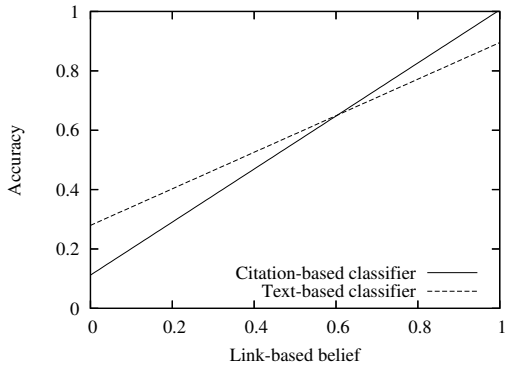

Figure 7: Accuracy per Amsler-based belief degree in the ACM8 collection.

low p, the decisions made by the text-based classifier are better. However, when the link based classifier estimates belief degrees above p, its decisions are considered more accurate. Such difference in accuracies is not so important in the ACM8 collection. In general, for any citation-link based belief degree, the classifier performances are very similar. Again, these differences observed in the ACM8 collection and the Cade12 collection are due to the already mentioned differences in the quality of the text based evidence existent in each collection.

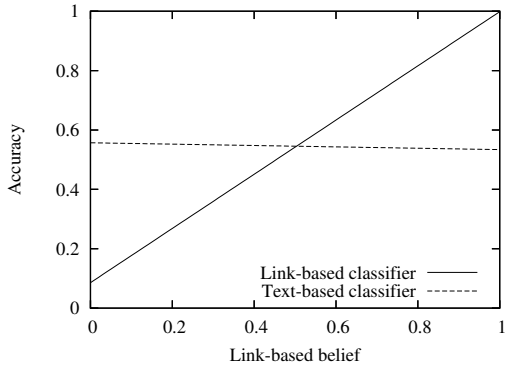

Figure 8: Accuracy per co-citation-based belief degree in the Cade12 collection.

Table 5 shows the results of the combination strategy described in Listing 1 in collections ACM8 and Cade12. Since the citation-link based classifier is used to guide the combination process, we refer to this strategy as *citation-link combination*. For comparison, Table 5 shows the results obtained by the citation-link based and text based classifiers in isolation. In particular, the citation-link based classifiers are used as our baselines because they yield the best results when considered in isolation. Finally, the table also shows the combination method described in [4], which we call *Bayesian combination*. This method uses a Bayesian network model to derive the probability $P(f_j|c)$ that a test document j belongs to class c. This probability is used to directly combine the belief degrees provided by the text based and the citation-link based classifiers and is defined as:

$$P(f_j|c) = \eta\Big[1 - \big(1 - W_t\ P(t_j|c)\big)\big(1 - W_l\ P(l_j|c)\big)\Big] \quad (6)$$

where η is a normalizing constant used to ensure that $P(f_j|c)$ fits between 0 and 1, $P(t_j|c)$ is the probability that document j belongs to class c according to the text based classifier, and $P(l_j|c)$ is the probability that document j belongs to class c according to the citation-link based classifier. Constants W_t and W_l are the weights given to the text based and to the citation-link based confidence estimations, respectively. They can be used to regulate the importance of each source of evidence on the final result. In our experiments we use weights W_t and W_l such that $\frac{W_l}{W_t} = 0.90$ for the ACM8 collection and $\frac{W_l}{W_t} = 0.20$ for the Cade12 collection. By employing these weight ratios *Bayesian combination* achieved its best performance in the experiments.

Collection	Methods	$micF_1$	$macF_1$	Gains (%) over link classifier	
				$micF_1$	$macF_1$
ACM8	Amsler	83.20	78.29	–	–
	KNN Cosine	76.95	70.57	-7.5	-9.9
	Bayesian	84.75	79.58	1.9	1.6
	Citation-link	84.07	78.90	1.0	0.8
Cade12	Co-citation	68.51	75.60	–	–
	SVM	54.18	48.41	-20.9	-36.0
	Bayesian	76.51	79.29	11.7	4.9
	Citation-link	78.04	80.39	13.9	6.3

Table 5: Macro-averaged and micro-average F_1 results for combining approaches in the ACM8 and Cade12 collections.

As we can see in Table 5, the best results for the ACM8 collection were obtained by the bayesian combination. This is not surprising since for this collection both text based and citation-link based classifiers provided reliable estimations. Thus a probabilistic combination of such estimations works well. However, the gains obtained from combination strategies in the ACM8 collection were quite small. This is due to the fact that there is little to gain since the accuracy of the baseline is already high. Further, experiments show that only 5.7% of the documents in the ACM8 collection are misclassified using citation evidence and correctly classified using text evidence, which makes it even harder to obtain better results.

This is not the case for the Cade12 collection, where text based and link based classifiers present very distinct performances and the overall accuracy for the link based classifier, when taken in isolation, is not so high. As a consequence, citation-link combination yielded a gain of about 14% in $micF_1$ over the co-citation metric used in isolation. This method was also slightly better than the bayesian combination, which can be explained by the poor probabilistic estimations provided by the SVM classifier. We also notice that the gains were much higher in terms of $micF_1$. In fact, for the co-citation measure a link based classifier is capable of distributing documents for all classes, independently of their sizes, whereas the SVM classifier tends to perform better in popular classes. This results in higher $micF_1$ for the SVM classifier and higher $macF_1$ for the link based classifier. When combining the outcome of these classifiers, we normally replace lower-confidence decisions, provided by the link based classifier, by the decisions provided by the SVM classifier. As a result, the combination tends to increase the number of correct decisions for popular classes increasing, by extent, its $micF_1$.

8. USER STUDY

Motivated by the failure in improving classification results in the ACM8 collection through our combination methods, as well as by the still low performance of the best link based classifier in the Cade12 collection, we decided to perform a user study using the cases that our classifiers did not succeed in providing a correct class to a test document.

When a citation-link based classifier assigns a wrong class to a test document it does so because the document shares more links with training documents of the wrong class than it does with training documents of the correct class. Since links and citations are an explicit indication from an author that his work is somehow related to another one, we suspected that these failed cases would be hard even for humans to classify. To test this assumption, we performed an experiment to study human classification of those unsuccessful cases.

For each of our test collections we randomly chose a sample of one hundred documents that were wrongly classified by the best citation-link classifier on each collection[6]. For each sample we generated three replicas and randomly distributed the resulting three hundred documents in twenty pools, in a way that each document would be evaluated by three distinct human classifiers. We assigned a pool to each volunteer.

Each person had the option to choose between two classes or to choose both classes. One of classes was the correct one, while the other was the class indicated by the automatic classifier. However, we allowed people to have access to much more information than the automatic classifier. Besides citation information, in the case of the ACM8 sample, people could analyze the title, authors, keywords, abstract (when available), the conference name, and the citations' text. Evaluators of the Cade12 collection sample had access to the full page content and hyperlinks inside the page.

Table 6 summarizes the results for the three hundred evaluations performed on each sample of documents. Notice that in both collections more than 50% of the classifications were wrong, or assigned two classes to the document. This confirms our expectation about the difficulty of classifying the sampled documents.

Human classification	Cade12	ACM8
correct	49.00%	46.67%
wrong	24.67%	32.67%
marked both classes	26.33%	20.66%

Table 6: Results of three hundred human classification in the user study.

We also collected some statistics on the documents classified, shown in Table 7. The experiment shows that there is little consensus between users with respect to the three options presented. If we sum the values of each column of Table 7 we have that only 40% in the documents of the sample of the Cade12 collection received the same opinion by all the three volunteers. There is even less consensus in the evaluation of the ACM8 collection with only 23% of the documents receiving the same vote from all the three evaluators. More importantly, only 23% of the documents of the Cade12 collection sample, and only 13% of documents of the ACM8 collection, were assigned the correct class by all evaluators.

[6]Amsler kNN for the ACM8 collection and co-citation kNN for the Cade12 collection.

% of documents classified:	Cade12	ACM8
correctly by all volunteers	23.00	13.00
wrongly by by all volunteers	11.00	9.00
as both by all volunteers	6.00	1.00

Table 7: Percentage of documents that reached consensus by the three human classifiers, in both collections.

We also investigated users' opinion for the cases in which the citation-link classifier failed but had assigned the correct class as the second most probable. We considered only the cases when the difference between the probabilities assigned to the first and second classes was less than or equal to 0.2. We call these cases classifier "doubts". In particular we wanted to verify how difficult these cases are for human classifiers. Table 8 shows that 60% of the classifier doubts received, from at least one human evaluator, a wrong class or two classes. Also, only a few doubt cases were correctly classified by all volunteers.

We studied the cases where human classifiers failed or marked two classes. Most of the cases involved the largest and most broad classes in the two collections (*Services* for Cade12 and *Comp. Syst. Org.* and *Software* for ACM8). These classes were also involved in most of the errors of automatic classifiers.

The above results and observations tend to indicate that the failures of the citation-link classifier are really difficult cases. Even human classification, using much more information, did not achieve much success. Further, consensus on the correct class is very rare among human evaluators and the "doubt" cases are even harder ones to classify correctly.

%	Cade12	ACM8
of documents that are doubt cases	15.00	20.00
of doubts wrongly classified or that received 2 classes	60.00	65.00
of doubts correctly classified by all volunteers	20.00	15.00

Table 8: Human classification of documents that were doubt cases.

9. CONCLUSIONS

In this work, we studied the relation between citations and links by applying both in a document classification task. Extensive experimentation and analytical studies were conducted to better understand the impact, similarities and differences between these two types of evidence. Such study provided a deeper understanding on how this information can be explored, as well as on its limitations.

Experiments show that, both in a digital library of computer science papers as on a Web directory, citation and link structures are always a reliable source of information, often providing classification results above those achieved when only textual content is used. The best measure to be used, however, can vary according to the collection's characteristics. In this work, we discovered that, for Web directories, measures based on co-citation are the best performers, whereas for digital libraries containing scientific papers, bibliographic coupling measures are more appropriate.

Textual content and citation-link structures are complementary sources of information. In this work, we also take advantage of this fact to propose a procedure to combine text based and citation-link based classifiers. The classification of a document is accomplished by selecting the more appropriate classifier, based on an estimation of its reliability. This type of combination achieved gains in micro averaged F1 of up to 14% in a Web directory, although gains were much less significant in the computer science digital library used.

Finally, a user study was conducted to determine the causes of misclassifications caused by citation-link information, in both collections. This study allowed us to conclude that documents that were wrongly classified were, in fact, hard to classify even for humans and, thus, there is probably no solution for this problem using automatic classification.

For future work, we intend to investigate the application of the studied approaches to develop a system for automatic categorization of new scientific articles. We also intend to investigate the usage of citation-link based classification to automatically expand Web directories. Finally, we intend to investigate ways of adapting the bibliometric measures here presented taking into consideration the quality of links available.

10. ACKNOWLEDGEMENTS

This work was supported in part by the GERINDO project grant MCT/CNPq/CT-INFO 552087/02-5, by the 5S-QV project grant MCT/CNPq/CT-INFO 551013/2005-2, by the CNPq grant 520916/94-8 (Nivio Ziviani), by the CNPq grant 300188/95-1 (Berthier Ribeiro-Neto), by the CNPq grant 303576/04-9 (Edleno Silva de Moura), by the FCT project POSC/EIA/581494/2004 (Pável Calado). Marco Cristo is supported by Fucapi, Manaus, AM, Brazil. Thierson Couto is supported by UCG, Goiânia, GO, Brazil.

11. REFERENCES

[1] R. Amsler. Application of citation-based automatic classification. Technical report, The University of Texas at Austin, Linguistics Research Center, December 1972.

[2] J. Bichtler and E. A. Eaton III. The combined use of bibliographic coupling and cocitation for document retrieval. *Journal of the American Society for Information Science*, 31(4):278–282, July 1980.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th International World Wide Web Conference*, pages 107–117, April 1998.

[4] P. Calado, M. Cristo, M. A. Gonçalves, E. S. de Moura, B. Ribeiro-Neto, and N. Ziviani. Link-based similarity measures for the classification of web documents. *Journal of the American Society for Information Science and Technology*, 57(2):208–221, 2006.

[5] P. Calado, M. Cristo, E. Moura, B. R.-N. Nivio Ziviani, and M. A. Gonçalves. Combining link-based and content-based methods for web document classification. In *Proceedings of the 12th International Conference on Information and Knowledge Management CIKM 2003*, pages 394–401, New Orleans, LA, USA, November 2003.

[6] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 307–318, June 1998.

[7] C. Chang and C. J. Lin. Libsvm: a library for support vector machines. 2001.

[8] D. Cohn and T. Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 430–436. MIT Press, 2001.

[9] M. Cristo, P. Calado, E. Moura, and B. R.-N. Nivio Ziviani. Link information as a similarity measure in web classification. In *10th Symposium On String Processing and Information Retrieval SPIRE 2003*, volume 2857 of *Lecture Notes in Computer Science*, pages 43–55, Oct. 2003.

[10] J. Dean and M. R. Henzinger. Finding related pages in the World Wide Web. *Computer Networks*, 31(11–16):1467–1479, May 1999. Also in Proceedings of the 8th International World Wide Web Conference.

[11] M. Fisher and R. Everson. When are links useful? Experiments in text classification. In *Advances in Information Retrieval, 25th European Conference on IR Research, ECIR 2003, Proceedings*, pages 41–56, April 2003.

[12] J. Furnkranz. Exploiting structural information for text classification on the WWW. In *Proceedings of the 3rd Symposium on Intelligent Data Analysis (IDA99)*, pages 487–498, August 1999.

[13] E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479, 1972.

[14] E. J. Glover, K. Tsioutsiouliklis, S. Lawrence, D. M. Pennock, and G. W. Flake. Using Web structure for classifying and describing Web pages. In *Proceedings of the 11th International World Wide Web Conference*, May 2002.

[15] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 137–142, Chemnitz, Germany, April 1998.

[16] T. Joachims, N. Cristianini, and J. Shawe-Taylor. Composite kernels for hypertext categorisation. In *Proceedings of the 18th International Conference on Machine Learning, ICML-01*, pages 250–257, June 2001.

[17] M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14(1):10–25, January 1963.

[18] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, Sep. 1999.

[19] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the Web for emerging cyber-communities. *Computer Networks*, 31(11–16):1481–1493, May 1999. Also in Proceedings of the 8th International World Wide Web Conference.

[20] S. Lawrence, C. L. Giles, and K. D. Bollacker. Autonomous citation matching. In O. Etzioni, J. P. Müller, and J. M. Bradshaw, editors, *Proceedings of the Third Annual Conference on Autonomous Agents (AGENTS-99)*, pages 392–393. ACM Press, May 1–5 1999.

[21] R.-L. Liu and W.-J. Lin. Adaptive sampling for thresholding in document filtering and classification. *Inf. Process. Manage.*, 41(4):745–758, 2005.

[22] H.-J. Oh, S. H. Myaeng, and M.-H. Lee. A practical hypertext catergorization method using links and incrementally available class information. In *Proceedings of the 23rd annual Intl. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 264–271, July 2000.

[23] G. Salton. Associative document retrieval techniques using bibliographic information. *Journal of the ACM*, 10(4):440–457, October 1963.

[24] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, 1988.

[25] A. Silva, E. Veloso, P. Golgher, B. Ribeiro-Neto, A. Laender, and N. Ziviani. CobWeb - a crawler for the brazilian web. In *Proceedings of the String Processing and Information Retrieval Symposium (SPIRE'99)*, pages 184–191, Sept. 1999.

[26] H. G. Small. Co-citation in the scientific literature: A new measure of relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269, July 1973.

[27] A. Sun, E.-P. Lim, and W.-K. Ng. Web classification using support vector machine. In *Proceedings of the Fourth International Workshop on Web Information and Data Management*, pages 96–99, November 2002.

[28] Y. Yang. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In *Proceedings of the 17rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 13–22, July 1994.

[29] Y. Yang, S. Slattery, and R. Ghani. A study of approaches to hypertext categorization. *Journal of Intelligent Information Systems*, 18(2):219–241, March 2002.