Combining Link-Based and Content-Based Methods for Web Document Classification

Pável Calado Marco Cristo Edleno Moura
Nivio Ziviani Berthier Ribeiro-Neto Marcos André Gonçalves

Dep. of Computer Science
Fed. Univ. of Minas Gerais
Belo Horizonte, MG, Brazil
{pavel,marco}@dcc.ufmg.br
{berthier,nivio}@dcc.ufmg.br

Dep. of Computer Science
Fed. Univ. of Amazonas
Manaus, AM, Brazil
edleno@dcc.ufmg.br

Virginia Tech
Dep. of Computer Science
Blacksburg, VA, USA
mgoncalv@vt.edu

ABSTRACT

This paper studies how link information can be used to improve classification results for Web collections. We evaluate four different measures of subject similarity, derived from the Web link structure, and determine how accurate they are in predicting document categories. Using a Bayesian network model, we combine these measures with the results obtained by traditional content-based classifiers. Experiments on a Web directory show that best results are achieved when links from pages outside the directory are considered. Link information alone is able to obtain gains of up to 46 points in F_1, when compared to a traditional content-based classifier. The combination with content-based methods can further improve the results, but too much noise may be introduced, since the text of Web pages is a much less reliable source of information. This work provides an important insight on which measures derived from links are more appropriate to compare Web documents and how these measures can be combined with content-based algorithms to improve the effectiveness of Web classification.

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

Classification, Web, link analysis, Bayesian networks

CIKM'03, November 3–8, 2003, New Orleans, Louisiana, USA.

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 contents, or the wide array of user's interests, make it a challenging environment for traditional IR solutions. On the other hand, the Web also provides grounds 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: its link structure.

It is possible to infer at least 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 others, we can assume that its content is important. These two assumptions have been successfully used in Web IR for tasks like page ranking [2, 16], finding site homepages [11], and document classification [4, 5, 7, 14, 25]. This work concerns the particular case of document classification, an especially important task that can be used to construct on-line directories, to improve the precision of Web searching, and even to help the interactions between user and search engines [28].

Content-based classifiers are known to perform poorly in the Web [4, 10]. Web documents are usually noisy and with little text, containing images, scripts and other types of data unusable by text classifiers. Furthermore, they can be created by many different authors, with no coherence in style, language or structure. For this reason, any evidence available, other than textual content, can be useful for classification. In this work, we experiment with different similarity measures derived from link structure and determine how accurate they are in predicting the subject of Web pages. We then show how these measures can be combined with the results of traditional content-based classifiers and explore how this combination can be used to improve classification effectiveness. The combination is obtained through a Bayesian network model and is independent of the link measures and classification algorithms used.

Four different link-based similarity measures were tested: bibliographic coupling, co-citation, Amsler and Companion. These were combined with three content-based classifiers: *kNN*, support vector machine, and Naive Bayes. Experiments performed on two sets of classified documents, extracted from a Web directory, indicate that best results are achieved when links from pages outside the directory are considered. Also, similarity measures that make use of in-links yielded better results, since directory Web pages tend to have more in-links than out-links. Link information alone is able to obtain gains of up to 46 points in F_1, when compared to a traditional content-based classifier. The combination with content-based methods can further improve the results, but too much noise may be introduced, since the text of Web pages is a much less reliable source of information. This work provides an important insight on which measures derived from links are more appropriate to compare Web documents and how these measures can be combined with content-based algorithms to improve the effectiveness of Web classification.

2. RELATED WORK

Link information has been previously proposed as a way of finding related Web documents. The Companion algorithm [6], for instance, uses links to determine pages related to a given initial page. Its functionality is briefly described in Section 3. On a different approach, He et al. [12] propose a clustering algorithm that groups Web pages by operating on the graph defined by their link structure. Co-citation and text similarity measures are used to assign weights to the edges of the graph and partitioning algorithms are used to split the set of pages into clusters. In [29], three measures of linkage similarity are compared to a human evaluation of similarity between Web pages. 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 [17]).

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. [27], 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 describing the links, paragraphs surrounding the links, and the headlines that structurally precede them. Furnkranz et al. [8], Glover et al. [9] and Sun et al. [27] achieved good results by using anchor text, and the paragraphs and headlines that surround the links, whereas Yang et al. [32] show that the use of terms from linked documents works better when neighboring documents are all in the same class.

In the link analysis approach, learning algorithms are applied to handle both the text components of the Web pages and the linkage among them. Slattery and Craven [25], for instance, explore the hyperlink topology using a HITS based algorithm [16] to discover test set regularities. Joachims et al. [14] studied the combination of support vector machine kernel functions representing co-citation and content information. Similarly, Cohn et al. [5] show that classification performance can be improved by using a combination of link-based and content-based probabilistic methods. Fisher and Everson [7] 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 high quality.

Finally, in the neighborhood approach, the document category is estimated based on category assignments of already classified neighboring pages. The algorithm proposed by Chakrabarti et al. [4] uses the known classes of training documents to estimate the class of the neighboring test documents. Their work shows that co-citation based strategies are better than those using immediate neighbors. Oh et al. [20] improved on this work by using a filtering process to further refine the set of linked documents to be used.

Our method mixes the link analysis and neighborhood approaches, differing from previous works in two main issues. First, we analyze several distinct linkage similarity measures and determine which ones provide the best results in predicting the category of a document. Second, we propose a Bayesian network model that takes advantage of both the information given by a content-based classifier and the information given by the document link structure. This model is independent of the classifier used, thus allowing us to study different classifier/link measure combinations. It also provides a formal and flexible way to test and combine new link-based and content-based algorithms in future works.

3. LINK-BASED SIMILARITY MEASURES

To determine the similarity of subject between two Web pages we used four different similarity measures derived from link structure: co-citation, bibliographic coupling, Amsler, and Companion.

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 are somewhat different 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.

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:

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

Eq. (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.

Also with the goal of determining the similarity between papers, Kessler [15] 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 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. 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:

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

Thus, 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.

In an attempt to take the most advantage of the information available in citations between papers, Amsler [1] 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. Thus, 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:

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

Eq. (3) tell us that, the more links (either parents or children) d_1 and d_2 have in common, the more they are related.

Finally, on a different approach, Dean and Henzinger proposed the Companion algorithm [6]. Given a Web page d, the algorithm finds a set of pages related to d by examining its links. Companion is able to return a degree of how related the topic of each page in this set is to the topic of page 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. In step 1, 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. This graph is then processed by the HITS algorithm [16], which returns a degree of *authority* and *hubness* for 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. Companion uses the degree of authority as a measure of similarity between d and each page in $\mathcal{V}$. For a more detailed description of the Companion and HITS algorithms, the user is referred to [6] and [16], respectively.

4. CONTENT-BASED CLASSIFIERS

In this work, we experimented with three well-known text classifiers: Naive Bayes, *kNN* and Support Vector Machine (SVM). These methods use different classification strategies and have been extensively evaluated for text classification on reference collections, thus offering a strong baseline for comparison.

The Naive Bayes classifier [18] uses a density estimation approach, based on a term independence hypothesis, to determine the probability of a category given a document. The *kNN* 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 in [31] and uses conventional *TF-IDF* weighting schemes as a similarity measure among the documents. Finally, the SVM classifier is a relatively new method, first used in text classification by Joachims [13]. It works over a vector space, where the problem is to find a hyperplane with the maximal margin of separation between two classes. This hyperplane can be uniquely constructed by solving a constrained quadratic optimization problem, by means of quadratic programming techniques.

Following, we present a Bayesian network model that allows the combination of the three classifiers here described with the link similarity measures presented in Section 3.

5. COMBINING LINK INFORMATION WITH A CONTENT-BASED CLASSIFIER

To combine link-based and content-based information, we propose the use of a Bayesian network model [21]. Bayesian networks have been successfully applied to several information retrieval tasks [22, 30] and have been shown especially useful when combining distinct sources of evidence [23]. They provide a graphical formalism for explicitly representing relationships among variables of a probability distribution and can derive any probability regarding such variables.

The network in Figure 1 shows the proposed model, where nodes represents pieces of information in the domain. To each node is associated a binary random variable, which takes the value 1 to mean that the corresponding information will be taken into account in our computation. In this case, we say the information was *observed*.

The root nodes, labelled D_1 through D_N, represent our prior knowledge about the problem, i.e., a set of classified documents (the training set). Node C represents a category. A category is constituted by a set of classified documents. Thus, there are edges from nodes D_j to node C, representing the fact that observing a set of classified documents will influence the observation of a category.

Nodes T_1 through T_K and L_1 through L_K represent the documents to be classified (the test set) under two different contexts. Each node T_i represents evidence from the content-based classifier indicating that test document i belongs to category C. Since this evidence depends on the training documents, there are edges from each node D_j to every node T_i. Thus, observing a set of training documents will influence the fact that we observe test document i as belonging to category C.

Similarly, each node L_i represents evidence, given by the link-based similarity measure, that document i belongs to category C. There is an edge from a node D_j to a node L_i if the training document j is related to test document i. We say that two documents i and j are related if their linkage similarity is greater than zero, as given by one of the link similarities described in Section 3. Thus, the fact that we observe test document i as belonging to category C is influenced by the observation of the training documents that are related to i.

Finally, nodes F_1 through F_K represent the final evidence that each test document belongs to category C. This evidence depends on both the content-based and the link based evidences, as shown by the incoming edges.

Given these definitions, we can now use the network to determine the probability that a test document i belongs to category C, i.e., the probability of observing the final evidence regarding document i, given that category C was observed, $P(F_i = 1|C = 1)$. This translates to the following equation[1]:

$$P(f_i|c) = \eta \sum_{\mathbf{d}} P(f_i|\mathbf{d})\ P(c|\mathbf{d})\ P(\mathbf{d}) \qquad (4)$$

[1] To simplify our notation we represent the probabilities $P(X = 1)$ as $P(x)$ and $P(X = 0)$ as $P(\overline{x})$.

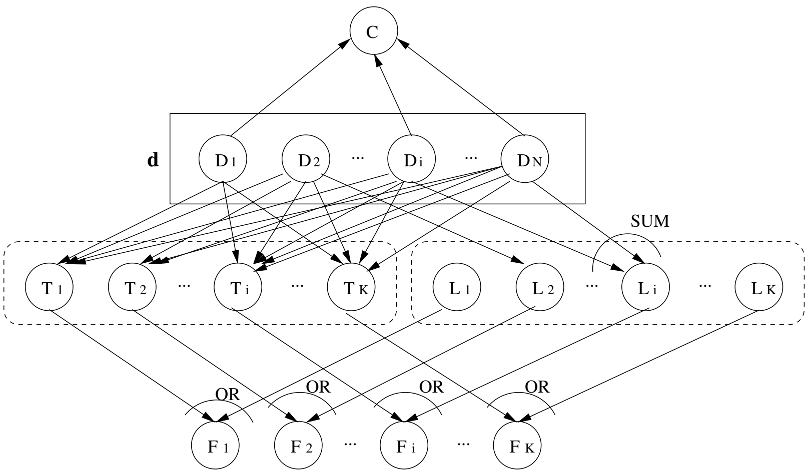

Figure 1: Bayesian network model to combine evidence from the content-based classifier with evidence from the link structure.

where $\eta = 1/P(c)$ is a normalizing constant and $\mathbf{d}$ is a possible state of all the variables D_j.

The probability $P(f_i|\mathbf{d})$ represents the combination of content-based and link-based evidences. To quantify it, we can use any function that obeys the axioms of probability [21]. For our experiments, we define $P(f_i|\mathbf{d})$ as a disjunction of the evidence from the classification algorithm and the link structure. This means that, for the final evidence to be observed, it is enough to observe one of the content-based or link-based evidences. Eq. (4) thus becomes:

$$P(f_i|c) = \eta \sum_{\mathbf{d}} \Big(1 - \big(1 - P(t_i|\mathbf{d})\big)\big(1 - P(l_i|\mathbf{d})\big)\Big) \quad P(c|\mathbf{d})\ P(\mathbf{d}) \qquad (5)$$

Eq. (5) is the general equation to compute the probability of a document belonging to a given category. We now need only to define the probabilities $P(t_i|\mathbf{d})$, $P(l_i|\mathbf{d})$, $P(c|\mathbf{d})$, and $P(\mathbf{d})$.

We start by defining the probability $P(t_i|\mathbf{d})$ that document i belongs to category C, given that the set of documents indicated by $\mathbf{d}$ was observed. To do this, let $\mathcal{C}$ be a set of documents labelled as belonging to category C, let $\overline{\mathcal{C}}$ be a set of documents labelled as not belonging to C, and let $class(i, \mathcal{C}, \overline{\mathcal{C}})$ be a function that returns a value of association between document i and category C, based on the labelled document sets. We define:

$$P(t_i|\mathbf{d}) = class(i, \mathcal{C}_{\mathbf{d}}, \overline{\mathcal{C}}_{\mathbf{d}}) \qquad (6)$$

where $\mathcal{C}_{\mathbf{d}}$ is the set of documents indicated as observed by $\mathbf{d}$ and $\overline{\mathcal{C}}_{\mathbf{d}}$ is the set of documents indicated as not observed by $\mathbf{d}$. The function *class* represents a content-based classifier and, for our experiments, the returned value is given by either the SVM, the *kNN*, or the Naive Bayes algorithms. We assume that this value is normalized such that $0 \leq class(i, \mathcal{C}, \overline{\mathcal{C}}) \leq 1$.

We now define the probability $P(l_i|\mathbf{d})$. Let $\mathcal{V}(i)$ be the set of training documents related to document i (notice that these documents are represented by the parent nodes of node L_i). Let $link(i, j)$ be the similarity between document i and document j, as given by one of the linkage similarity measures described in Section 3. We define:

$$P(l_i|\mathbf{d}) = \alpha \sum_{j \in \mathcal{V}(i) | d_j = 1} link(i, j) \qquad (7)$$

where α is a normalizing constant used to keep the sum between 0 and 1. Evidence represented by L_i is, therefore, defined as the sum of the values given by the linkage similarity between document i and all of its related documents that are indicated by $\mathbf{d}$ as observed. Thus, the more training documents related to i belong to a given category, the greater the probability that i belongs to the same category.

The probability $P(c|\mathbf{d})$ is now used to select only the training documents that belong to the category we want to process. We define $P(c|\mathbf{d})$ as:

$$P(c|\mathbf{d}) = \begin{cases} 1 & \text{if } \forall_i,\ d_i = 1 \text{ iff } i \in \mathcal{C} \\ 0 & \text{if otherwise} \end{cases} \qquad (8)$$

where $\mathcal{C}$ is the set of training documents that belong to category C.

Finally, since we have no initial preference as to what set of training documents is more probable of being observed, we can regard the *a priori* probability $P(\mathbf{d})$ as a constant for every $\mathbf{d}$. By applying Eqs. (6), (7), and (8) to Eq. (5) we obtain the final equation to compute the probability that document i belongs to category c:

$$P(f_i|c) = \rho\Big(1 - \big(1 - class(i, \mathcal{C}_{\mathbf{d}}, \overline{\mathcal{C}}_{\mathbf{d}})\big) \big(1 - \alpha \sum_{j \in \mathcal{V}(i) | d_j = 1} link(i, j)\big)\Big) \qquad (9)$$

where $\rho = P(\mathbf{d})/P(c)$ is a normalizing constant and $\mathbf{d}$ is the state where only the documents labelled as belonging to class C are active.

We observe that the belief network is used here as a modeling framework and not as an inference engine. While more complex designs are possible, our simple representation is powerful enough to allow modeling important relationships between documents, classes, and link structure.

6. EXPERIMENTS

6.1 Methodology

To evaluate all the linkage similarity measures and the effects of the combination, experiments were performed using a set of classified Web pages, extracted from the Cadê directory (`http://www.cade.com.br/`). Cadê is a Brazilian Web directory, pointing to Web pages that were classified by human experts. To obtain the contents of the classified pages we used the database of Brazilian Web pages crawled by the TodoBR search engine [24] (`http://www.todobr.com.br/`).

Two sub-collections were constructed using the data available on Cadê: Cade12 and Cade188. Cade12 is a set of 44,099 pages labelled using the 12 first level categories of Cadê (Computers, Culture, Education, Health, News, Internet, Recreation, Science, Services, Shopping, Society, and Sports). Cade188 corresponds to a set of 42,004 pages labelled using the 188 second level categories of Cadê (Biology, Chemistry, Dance, Music, Schools, Universities, etc.). Each Web page is classified into only one category. As shown in Figure 2, both collections have very skewed distributions. For instance, in Cade12, the three most popular categories represent more than 50% of all documents whereas, in Cade188, 50% of the documents are in just 10% of the categories. Cade12 and Cade188 have vocabularies of 192,580 and 168,869 unique words, respectively, after removing stop words.

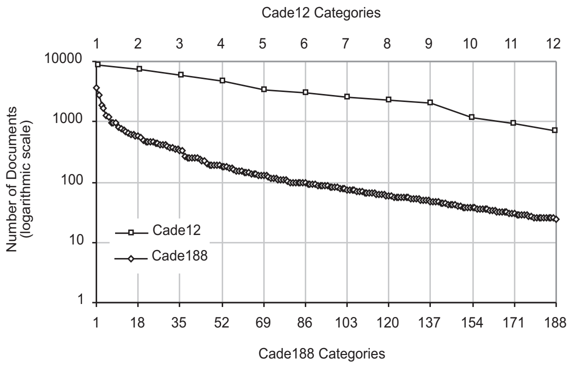

Figure 2: Compared distributions for the Cade12 and Cade188 collections.

Information about the links related to the Cadê pages was also collected 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. Links connecting different pages of the same site were discarded. Table 1 summarizes the data obtained. Links were divided into two types: *internal*, which are links between pages classified by Cadê, and *external*, 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.

It is interesting to note that external pages are 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 and verify how useful they are for the classification process. This is only possible because Cadê is a subset of TodoBR, which is a large collection that mirrors the link distribution of the Web [3]. This is not the case in most other reference 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.

Statistics	N. of links
Internal Links	3,830
Links *from* external pages	570,337
Links *to* external pages	5,894
Cadê pages with no in-links	1,625
Cadê pages with no out-links	40,723

Table 1: Link statistics for the Cadê collection.

To experiment with the text-based classifiers, we used 10-fold cross validation. The final results of each experiment represent the average of ten runs. To make sure that the results are not biased by an inappropriate choice of parameters, several experiments were performed and, in all cases, we report the best results obtained. Thus, for the *kNN* classifier, the value of k was set to 30 and 15,000 features were considered. For the SVM classifier, a linear kernel was used and 10,000 features were considered. For the Naive Bayes classifier, 15,000 features were considered. For all algorithms, the features were selected using the information gain method [19].

The performance of the presented methods was evaluated using the conventional precision, recall and F_1 measures. 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 defined as $\frac{2pr}{p+r}$. Macro-averaging and micro-averaging [32] were applied to get single performance values over all classification tasks.

6.2 Results

We started by analyzing how each source of information, content and links, performs when used in isolation. This can be accomplished by setting the probabilities $P(t_i|\mathbf{d})$ or $P(l_i|\mathbf{d})$ to zero, according to the evidence to be tested. Table 2 shows the micro-averaged F_1 values for the content-based classifiers and linkage similarity measures used in isolation. The linkage similarity measures were computed using only internal links (marked (i)) and both internal and external links (marked (i+e)). The highest values for each classifier and similarity measure are shown in bold face.

The content-based classifiers, as expected, show poor results, indicating that the text of the Web directory pages does not provide enough information to reliably classify the documents. The F_1 values were much lower for the Cade188 collection, since the classifiers tend to perform worst in collections where the class distribution is more skewed. The best results were achieved by SVM on the Cade12 collection and by *kNN* on the Cade188 directory, with F_1 values of 40.86 and 24.45, respectively.

For the linkage similarity measures, when only internal links are available, information is clearly insufficient, thus

Source of evidence	Cade12	Cade188
kNN	39.45	**24.45**
SVM	**40.86**	24.31
Naive Bayes	39.38	22.82
Bibliographic coupling (i)	13.61	0.89
Amsler (i)	14.00	1.23
Co-citation (i)	13.84	1.11
Companion (i)	**14.30**	**1.44**
Bibliographic coupling (i+e)	13.70	0.94
Amsler (i+e)	80.49	66.91
Co-citation (i+e)	**80.70**	67.58
Companion (i+e)	75.55	**69.58**

Table 2: Micro-averaged F_1 measures obtained using the evidence provided by the content-based classifiers and linkage similarity measures in isolation. For the linkage measures, the mark (i) stands for using only internal links and the mark (i+e) stands for using internal and external links.

yielding very low F_1 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.

When links to and from external pages are used, however, link information alone was enough to achieve classification results well above those achieved by the content-based classifiers. For the Cade12 collection, the best results were obtained using the co-citation measure, showing 80.7 points in micro-averaged F_1. For the Cade188 collection, the Companion algorithm had the best performance, with 69.6 points in micro-averaged F_1.

Bibliographic coupling yielded lower F_1 values than the remaining measures. This is not surprising since it relies only on out-link information and, as shown in Section 6.1, more than 90% of the pages have no out-links. Since most of the links are *from* external pages *to* pages in the collection, we can expect measures that make use of in-links to perform the best.

We now verify the effects of combining both measures, using our proposed Bayesian Network model. Table 3 shows the micro-averaged F_1 figures obtained by combining the four different similarity measures with the results of the content-based classifiers. The highest values for each similarity/classifier combination are shown in bold face.

Again, we observe that the results using only internal links were generally poor. Although the combination showed an improvement over the use of links alone, F_1 values were below those achieved by the content-based classifiers. Only for the Naive Bayes classifier some of the figures were very slightly above the content-only baseline. Due to the lack of internal links the similarity measures introduce much noise into the classification process. Co-citation performed the best because very few documents were co-cited, thus having a zero similarity value, which left the final classifier decision to the content-based algorithms.

Results show some improvement when we make use of external links. However, and although there was a large improvement over the use of the content-based classifiers in isolation, not always the combination was able to supersede the results of the link similarity measures. For the Cade12 collection, Amsler and co-citation yield a gain of less than 1 point in micro-averaged F_1, when combined with the *kNN* classifier. When combined with the SVM and Naive Bayes classifiers both yielded a loss of about 2.8 and 21.7 points, respectively. The Companion algorithm used in isolation always performed better than when combined with a classifier. The bibliographic coupling measure was an exception, mainly because of its poor performance when used alone.

In the Cade188 collection, improvements over the link-only results are more evident. For the Amsler and co-citation similarities, gains go from 1.95 to 3.66 points in micro-averaged F_1, when combined with the *kNN* and SVM classifiers. Only combination with the Naive Bayes classifier shows poorer results. The Companion algorithm used in isolation still yields a higher F_1 value than when combined with *kNN* and Naive Bayes. However, when combined with the SVM classifier, there is a gain of 3.95 points.

A closer observation of the results shows that, in general, the link-based measures are able to correctly classify a great number of documents, showing high recall values. On the other hand, due the existence of many links that are not related to topic (e.g., navigation links, advertising, etc.), many documents are also incorrectly classified, thus hurting precision. By combining with a content-based classifier, many of these incorrectly classified documents are filtered out. Of course, many correctly classified documents are also filtered-out. If too much importance is given to the content-based classifier, too many of the correctly classified documents will be lost, and the final combination effectiveness will decrease.

The importance given to each source of evidence considered in the combination depends on the probabilities $P(t_i|\mathbf{d})$ and $P(l_i|\mathbf{d})$ (see Eqs. (6) and (7)). These probabilities are assumed to reflect the degree of "certainty" to which a classifier, or similarity measure, believes that a document i belongs to a given category. According to the proposed model, if a classifier gives a high value for $P(t_i|\mathbf{d})$, whereas a linkage similarity measure gives a low value for $P(l_i|\mathbf{d})$, more importance should be given to the classifier, since it is more "certain" that the document belongs to the given category. This assumption has a strong effect on the final results.

To exemplify, consider the combination of *kNN* with any of the link measures used. In our experiments, the *kNN* classifier yielded the smallest values for $P(t_i|\mathbf{d})$. Thus, link information has a much stronger influence on the final combination results. When using external links, link information is more reliable. Since it is given more weight, the effectiveness of the combination is improved. In fact, the best overall results were obtained by combining link measures with the *kNN* classifier. When using only internal links, content information is much more reliable and, thus, combining with the *kNN* classifier yielded the lowest F_1 values.

For the SVM and Naive Bayes classifiers, the values for $P(t_i|\mathbf{d})$ are about 100 times higher than for *kNN*, giving too much importance to the text-based evidence, and thus harming the combination results. Among the similarity measures, the Companion algorithm provided the lowest values for $P(l_i|\mathbf{d})$ (about 10 times lower than co-citation, Amsler, and bibliographic coupling), and consequently, shows the lowest F_1 values when combined with the text-based classifiers.

7. CONCLUSIONS

The experiments performed with the Cadê directory show that links that are internal to the collection do not provide sufficient information for document classification. In order

Linkage Similarity Measures	Cade12			Cade188		
	kNN	SVM	NB	*kNN*	SVM	NB
Bibliographic coupling (i)	36.47	40.19	39.27	22.60	23.58	22.70
Amsler (i)	36.40	40.12	39.40	22.56	23.44	22.78
Co-citation (i)	**36.91**	**40.69**	**39.45**	**23.10**	**24.13**	22.89
Companion (i)	36.56	40.29	39.35	22.67	23.75	**22.90**
Bibliographic coupling (i+e)	36.31	40.02	39.30	22.32	23.08	22.70
Amsler (i+e)	81.26	77.65	58.80	70.57	68.91	47.01
Co-citation (i+e)	**81.55**	**77.89**	**59.03**	**71.07**	69.53	**47.31**
Companion (i+e)	73.00	63.66	42.76	68.54	**73.63**	29.82

Table 3: Micro-averaged F_1 measures obtained with three classifiers in Cade12 and Cade188 collections, using different similarity measures. The mark (i) stands for using only internal links. The mark (i+e) stands for using internal and external links.

to achieve expressive results, links to and from pages outside the directory should be used. Also, most external pages are parents of the pages in the collection, i.e., they have a link to the pages in the collection. Thus, similarity measures that make use of in-link information are expected to be the most appropriate.

Of the similarity measures tested, co-citation, Amsler, and Companion show good results. Bibliographic coupling showed a much inferior performance, due to the fact the it uses only out-link information. It may, therefore, not be the most appropriate measure for Web solutions. Using external links, all measures, except bibliographic coupling, showed better results than any of the content-based classifiers. This confirms the importance of link information for Web document classification. Since co-citation is the most straightforward measure to compute, it is a good candidate for Web IR solutions that intend to use link information to compare pages.

The combination of the link-based measures with the content-based classifiers yielded mixed results. Link information can correctly classify a large number of documents, but also introduces noise. Content-based information can filter out some of the noise, but also removes documents that were correctly classified by the linkage similarities. Results suggest that the effectiveness of the combination approach may depend on the importance given to each of the sources of evidence to be combined. More weight should be given to those that provide more reliable information.

Our combination model makes no *a priori* assumption about the importance of the sources of evidence. The probabilities to be combined depend only on the characteristics of the algorithms providing the information and on the parameters used to configure them. Thus, when evaluating if a document i belongs to a category C, the model will give more weight to the source of evidence that yields the highest probability of i belonging to C. However, the model could be modified to allow the insertion of user-defined weights, in order to fine tune the classification and provide adaptation to other collections.

Finding the ideal weights for each of the evidences to be combined is an important problem. In future works, we intend to study the effects of such weights on the results of the combination. An evaluation of when an how much one evidence should be favored over the other will be performed in different reference collections. Methods to automatically determine such weights will also be investigated. Other future works include testing alternative ways to combine link-based and content-based evidences, and evaluating the performance of the similarity measures and the combination model in other reference collections.

8. 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). Marco Cristo is supported by Fucapi, Technology Foundation, Manaus, AM, Brazil. Marcos André Gonçalves is supported by a fellowship from AOL.

9. REFERENCES

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

[2] 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.

[3] P. Calado, B. Ribeiro-Neto, N. Ziviani, E. Moura, and I. Silva. Local versus global link information in the Web. *ACM Transactions On Information Systems*, 21(1):42–63, January 2003.

[4] 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, Seattle, Washington, June 1998.

[5] 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.

[6] 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.

[7] M. Fisher and R. Everson. When are links useful? Experiments in text classification. In F. Sebastianini, editor, *Proceedings of the 25th annual European*

conference on Information Retrieval Research, ECIR 2003, pages 41–56. Springer-Verlag, Berlin, Heidelberg, DE, 2003.

[8] J. Furnkranz. Exploiting structural information for text classification on the WWW. In *Intelligent Data Analysis*, pages 487–498, 1999.

[9] 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 WWW-02, International Conference on the World Wide Web*, 2002.

[10] N. Gövert, M. Lalmas, and N. Fuhr. A probabilistic description-oriented approach for categorizing web documents. In *Proceedings of the 8th International Conference on Information and Knowledge Management CIKM 99*, pages 475–482, Kansas City, Missouri, USA, November 1999.

[11] D. Hawking and N. Craswell. Overview of TREC-2001 Web track. In *The Tenth Text REtrieval Conference (TREC-2001)*, pages 61–67, Gaithersburg, Maryland, USA, November 2001.

[12] X. He, H. Zha, C. H. Q. Ding, and H. D. Simon. Web document clustering using hyperlink structures. *Computational Statistics & Data Analysis*, 41(1):19–45, November 2002.

[13] 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.

[14] T. Joachims, N. Cristianini, and J. Shawe-Taylor. Composite kernels for hypertext categorisation. In C. Brodley and A. Danyluk, editors, *Proceedings of ICML-01, 18th International Conference on Machine Learning*, pages 250–257, Williams College, US, 2001. Morgan Kaufmann Publishers, San Francisco, US.

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

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

[17] S. R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. The Web as a graph. In *Proceedings of the 19th Symposium on Principles of Database Systems*, pages 1–10, Dallas, Texas, USA, May 2000.

[18] A. McCallum and K. Nigam. A comparison of event models for Naive Bayes text classification. In *Proceedings of AAAI/ICML-98, Workshop on Learning for Text Categorization*, pages 41–48. AAAI Press, 1998.

[19] T. Mitchell. *Machine Learning*. McGraw-Hill, March 1997.

[20] 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 international ACM SIGIR conference on Research and development in information retrieval*, pages 264–271. ACM Press, 2000.

[21] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of plausible inference*. Morgan Kaufmann Publishers, 2nd edition, 1988.

[22] B. Ribeiro-Neto and R. Muntz. A belief network model for IR. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 253–260, Zurich, Switzerland, August 1996.

[23] B. Ribeiro-Neto, I. Silva, and R. Muntz. *Soft Computing in Information Retrieval: Techniques and Applications*, chapter 11—Bayesian Network Models for IR, pages 259–291. Springer Verlag, 1st edition, 2000.

[24] 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, Cancun, Mexico, September 1999.

[25] S. Slattery and M. Craven. Discovering test set regularities in relational domains. In P. Langley, editor, *Proceedings of ICML-00, 17th International Conference on Machine Learning*, pages 895–902, Stanford, US, 2000. Morgan Kaufmann Publishers, San Francisco, US.

[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. ACM Press, 2002.

[28] L. Terveen, W. Hill, and B. Amento. Constructing, organizing, and visualizing collections of topically related Web resources. *ACM Transactions on Computer-Human Interaction*, 6(1):67–94, March 1999.

[29] M. Thelwall and D. Wilkinson. Finding similar academic Web sites with links, bibliometric couplings and colinks. *Information Processing & Management*, 2003. (in press).

[30] H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, July 1991.

[31] Y. Yang. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In W. B. Croft and e. C. J. van Rijsbergen, editors, *Proceedings of the 17rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 13–22. Springer-Verlag, 1994.

[32] Y. Yang, S. Slattery, and R. Ghani. A study of approaches to hypertext categorization. *Journal of Intelligent Information Systems*, 18(2-3):219–241, 2002.