**Information Retrieval manuscript No.**
(will be inserted by the editor)

# Classifying Documents with Link-Based Bibliometric Measures

**T. Couto · N. Ziviani · P. Calado**

**· M. Cristo · M. Gonçalves**

**· E. S. de Moura · W. Brandão**

**Abstract** Automatic document classification can be used to organize documents in a digital library, construct on-line directories, improve the precision of web searching, or help the interactions between user and search engines. In this paper we explore how linkage information inherent to different document collections can be used to enhance the effectiveness of classification algorithms. We have experimented with three link-based bibliometric measures, co-citation, bibliographic coupling and Amsler, on three different document collections: a digital library of computer science papers, a web directory and an on-line encyclopedia. Results show that both hyperlink and citation

T. Couto, N. Ziviani M. Golçalves and W. Brandão
Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte, Brazil;
Tel.: +55 31 3499 5860
Fax: +55 31 3499 5858
E-mail: thierson,nivio,mgoncalv,wladmir@dcc.ufmg.br

P. Calado
IST/INESC-ID, Lisboa, Portugal
E-mail: pavel.calado@tagus.ist.utl.pt

M. Cristo
FUCAPI - Analysis, Research and Tech. Innovation Center, Manaus, Brazil
E-mail: marco.cristo@fucapi.br

E. S. de Moura
Department of Computer Science, Federal University of Amazonas, Manaus, Brazil
E-mail: edleno@dcc.ufam.edu.br

information can be used to learn reliable and effective classifiers based on a $k$NN classifier. In one of the test collections used, we obtained improvements of up to 69.8% of macro-averaged $F_1$ over the traditional text-based $k$NN classifier, considered as the baseline measure in our experiments. We also present alternative ways of combining bibliometric based classifiers with text based classifiers. Finally, we conducted studies to analyze the situation in which the bibliometric-based classifiers failed and show that in such cases it is hard to reach consensus regarding the correct classes, even for human judges.

**Keywords** Text classification, links, web directories, digital libraries

## 1 Introduction

Document classification is an especially important task that can be used to organize documents in a digital library, construct on-line directories, improve the precision of web searching, and even help the interactions between user and search engines (Terveen et al. 1999).

Traditional document classification techniques use words and expressions extracted from documents, as the only relevant features useful for determining to which categories a document belongs. They rely, therefore, only on textual information. However, many important document collections, such as web pages and digital libraries contain other components that can be explored by classifiers, such as the text structure, its formatting or, as is the focus of this work, some type of link structure connecting the documents. For web pages, for instance, these links can be derived directly from the *hyperlinks*

between pages. For digital libraries, they can be derived from the *citations*[1] between documents.

We know that citations and links are used with multiple motivations (Smith 2004). For instance, citations can be used to provide background information, give credit to the authors of an idea, or discuss or criticize existing work. They can be also used as rhetorical devices to persuade the reader of claims, praise colleagues, or receive grants and awards (Moed 2005). Web links, besides having the same functionality, can be also used for advertising, in-site navigation, providing access to databases, among others. In both cases, we assume that a link is a statement of an author that his or her document is somehow related to another document in the collection. In fact, several works in IR have successfully used links and citations as evidence in tasks such as finding site homepages (Hawking and Craswell 2001) and document classification (Chakrabarti et al. 1998; Slattery and Mitchell 2000; Joachims et al. 2001; Cohn and Hofmann 2001; Fisher and Everson 2003), which is the focus of this work.

Link information can be specially important in document classification when documents are noisy or contain little text, a circumstance where traditional content based techniques are known to perform poorly (Chakrabarti et al. 1998; Gövert et al. 1999). This is often the case of web documents, which 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. Thus, any evidence other than textual content may be useful for classification.

[1] In this article we consider a *citation* as a direct link from a citing article to a cited one.

Following this assumption, we present a comparative study of the use of bibliometric similarity measures for classifying documents. Three different link based bibliometric measures were used: co-citation (Small 1973; Marshakova 1973), bibliographic coupling (Kessler 1963) and Amsler (Amsler 1972). These measures were first used in (Egghe and Rousseau 1990) in the context of citations among scientific documents. In this work, we experiment with them in three different contexts: a digital library of scientific papers, a web directory, and an on-line encyclopedia. In each case, the measures are derived either from hyperlinks between web pages or citations between articles.

Results for our comparative study show that both hyperlink and citation information, when properly distributed over the documents and classes, can be used to learn reliable and effective classifiers based on the $k$NN classification method. 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 ten-fold cross validation have reached values of macro-average and micro-average $F_1$ superior to state-of-the-art text based classifiers in two of the collections studied and, in the subcollection of an encyclopedia, the micro-average $F_1$ value is only marginally distinct from the one obtained with a text-based classifier learned using the SVM model.

Experiments with the three collections have shown that the link based classifiers using bibliometric measures perform well for classifying both web and digital library documents. We obtained improvements of up to 69.8% of macro F1 over a text based $k$NN classifier, used as the baseline measure in our experiments. We present empirical evidence that (i) the number of in-links and out-links is important to learn bibliomet-

ric based classifiers and (ii) the co-occurrence of in-links and out-links is important to determine the existence of bibliometric relations between documents. We also study two alternative ways of combining our link-based classifiers with traditional text-based classifiers, performing an analysis of the gains obtained by each alternative combination.

Finally, we also investigate the possible reasons for the cases of failure of the bibliometric-based classifiers. A user study was performed, where volunteers were asked to classify a random sample of documents. This study shows that most cases are in fact hard to classify even for humans, and that there is little consensus among the volunteers regarding the correct class of a same document. Also, many of the failure cases regard test documents for which the second most probable class assigned by the classifier was the correct class. Our user study confirmed that the majority of these documents are considered multi-classification cases by the volunteers.

Preliminary results of this work were presented in (Cristo et al. 2003; Calado et al. 2003; Couto et al. 2006). In particular, the main differences between this and previous work are as follows. First, we analyzed the bibliometric measures in three different collections, with a detailed analysis of the impact of the distribution of links. As a consequence, we show that each bibliometric measure behaves differently for each type of collection, according to the number and distribution of in-links and out-links. Second, we run a series of user experiments to show the possible reasons for the failures of bibliometric-based classifiers. Third, we present alternative methods to combine classifiers and compare their effectiveness to the performance of an ideal perfect combiner, showing that the gains obtained with combination are important even when they are small, since in such cases even a perfect combiner could not perform much better.

The paper is organized as follows. Section 2 discusses related work regarding the use of links in information retrieval, particularly in the task of classifying documents. Section 3 presents the bibliometric measures used in this work. Section 4 describes the classification methods used, as well as how these methods can make use of bibliometric measures. Section 5 describes the test collections used in the experiments. Section 6 presents the results of comparing classifications using bibliometric measures and only textual features. It also presents the comparison of two distinct methods for combining bibliometric information with textual information, in order to enhance document classification, and a user study conducted to investigate how difficult is the classification of those documents that the automatic classifier failed to classify correctly. Finally, Section 7 presents our conclusions and suggests future work.

## 2 Related Work

Citation links among documents were first used as a source of information in bibliometric science. In 1963, Kessler introduced the notion of bibliographic coupling (Kessler 1963), a measure that can be used to determine documents with similar topics. Later, the measure of co-citation was introduced, independently and simultaneously in (Small 1973) and (Marshakova 1973). Both measures have been used as complementary sources of information for document retrieval and classification (Salton 1963; Amsler 1972; Bichtler and Eaton III 1980) and as a means to evaluate the importance of scientific journals (Garfield 1972).

The ideas used for citation links among documents were later transposed to the Web environment (Larson 1996; Almind and Ingwersen 1997). 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, bibliographic citations are static and not reciprocal, that is, a document is never cited by an older document. On the other hand, web hyperlinks may be used for navigation purpose only, which is not the case of article citations. Despite these differences, bibliometric measures, either using hyperlinks or citations, are able to capture strong relations between two documents, because they are based on the number of linked documents in common, and not on the direct linkage between them. Our experiments show that whenever pairs of documents have common linked documents, bibliometric measures are very useful for the classification task.

In (Brin and Page 1998) and (Kleinberg 1999), algorithms were proposed to derive measures of importance for web pages using Web link structure. Such measures could then be applied to document ranking, greatly improving 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 (Dean and Henzinger 1999), discovering web communities (Kumar et al. 1999), or classifying web pages (Sun et al. 2002).

Classification of web pages using links has already deserved a wide attention from the IR community. For instance, in (Furnkranz 1999), (Glover et al. 2002), (Sun et al. 2002), and (Yang et al. 2002), 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. Particularly, in (Yang et al. 2002) it is shown that the use of terms from linked documents works better when neighboring documents are all in the same class.

Another common strategy consists in estimating categories based on category assignments of already classified neighboring pages. Chakrabarti proposed in (Chakrabarti et al. 1998) an algorithm that uses the known classes of training documents to estimate the class of the neighboring test documents through an iterative process. This idea was further improved in (Oh et al. 2000) and (Angelova and Weikum 2006). In particular, Oh et al used a filtering process to refine the set of linked documents to be used as neighbors. In (Angelova and Weikum 2006), the class estimation was enhanced by taken into consideration some additional evidence such as a distance metric between the classes.

A common problem with these strategies is the scarcity of link information. Such problem has motivated some additional works. For instance, Qi and Davison show how neighboring unlabeled pages can be used to estimate the category of a page (Qi and Davison 2006). In (Shen et al. 2006), the enrichment of the linkage structure with artificial links is proposed. These links are added to non-connected pages that are found associated to each other according to their click-through patterns in a search-engine query log.

Other authors have applied learning algorithms to handle both the text components in web pages and the linkage between them. Joachims et al studied the combination of support vector machine kernel functions representing co-citation and content information (Joachims et al. 2001). Kernel methods were also successfully applied to the domain of patents in (Li et al. 2007). The results results show that a kernel derived from a patent citation network significantly outperforms kernels using no citation information, or citation information considering only neighboring patents.

By using a combination of bibliometric based and text based probabilistic methods, the proposals in (Cohn and Hofmann 2001) and (Fisher and Everson 2003) improved

classification performance over a text-based baseline and show that link information is useful when the document collection has a high link density[2] and most links are of high quality[3]. Two recent strategies have been successfully applied both to citations and web links. The first, proposed in (Zhang et al. 2005), explored a genetic programming strategy to find an effective combination function for classification. In the second, Veloso et al (Veloso et al. 2006) presented a lazy classification approach based on association rules. Both methods were shown to outperform other combination strategies in the literature.

In this work, we also suggest a combination strategy based on the reliability of the classifiers. To know which, between two classifiers, is the more accurate, we test them on a same set of instances. Thus, we can say that we use actual prediction information to refine the original models. In this sense, this algorithm is related to works that propose the use of actual data to refine classification models. For instance, in (Saerens et al. 2002), a simple iterative procedure was presented to adjust the outputs of a classifier with respect to new a priori probabilities estimated from prediction on actual data. Unlike our approach, these work are not focused on combining multiple classifiers.

The main difference between this and previous work is that we study the use of distinct link based similarity measures for document classification in different environments and analyze in detail the reasons for failure and success of each measure in each environment. Further, we perform a user study on misclassification in order to better understand such results. Finally, we present a new study on how to combine the

[2] *Density* in (Fisher and Everson 2003) is a measure based on the sparseness ($\Gamma$) of the collection of linked documents which is defined as the ratio $\Gamma = \frac{|C|}{J \times L}$, where $C$ is the number of edges in the graph, $J$ is the number of vertices(documents) and $L$ is the number of documents that can be cited.

[3] *Quality* is used as property of links that indicates whether links express relatedness between documents or are used for navigation purpose only.

results of text based and bibliometric based classifiers and show that the results of the combinations achieved do not differ much from those obtained using an ideal combiner.

## 3 Bibliometric Similarity Measures

In this section we present the bibliometric similarity measures used with link-based classifiers: Co-citation (Small 1973), Bibliographic coupling (Kessler 1963) and Amsler (Amsler 1972). These three metrics are one-step metrics that only consider directly adjacent neighbor nodes.

Given a document $d$, we define $S_d$ (the *sources* of $d$) as the set formed by all documents that link to $d$. We also define $T_d$ (the *targets* of $d$) as the set formed by all documents $d$ links to. Note that, in general, $T_d$ is a relatively static set whereas $S_d$ can grow over time. Thus, for newly published documents/pages $S_d$ probably will not provide much information. We now describe each link-based bibliometric measure.

The co-citation similarity between two pages $d_1$ and $d_2$ is defined as:

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

Eq. (1) shows that, the more source documents $d_1$ and $d_2$ have in common, the more related they are. This value is normalized by the total set of sources, so that the co-citation similarity varies between 0 and 1. If both $S_{d_1}$ and $S_{d_2}$ are empty, we define the co-citation similarity as zero.

For example, given the documents and links in Figure 1, we have that $S_A = \{D, E, G, H\}$ and $S_B = \{E, F, H\}$, $S_A \cap S_B = \{E, H\}$ and $S_A \cup S_B = \{D, E, F, G, H\}$. Thus $co{-}citation(A, B) = \frac{2}{5}$.

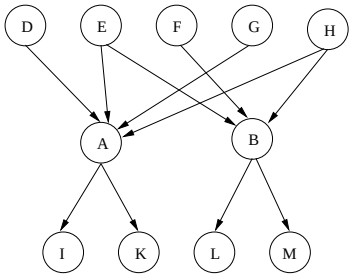



**Fig. 1** Documents $A$ and $B$ with their source documents and target documents.

Bibliographic coupling between two pages $d_1$ and $d_2$ is defined as:

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

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

Consider the example shown in Figure 1. $T_E = \{A, B\}$, and $T_H = \{A, B\}$. So, according to Equation (2), $bibcoupling(E, H) = 1$.

The Amsler similarity between two pages $d_1$ and $d_2$ is defined as:

$$Amsler(d_1, d_2) = \frac{(S_{d_1} \cup T_{d_1}) \cap (S_{d_2} \cup T_{d_2})}{|(S_{d_1} \cup T_{d_1}) \cup (S_{d_2} \cup T_{d_2})|} \tag{3}$$

Eq. (3) tell us that, the more links (either sources or targets documents) $d_1$ 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 source documents or target documents, the similarity is defined as zero.

Once again, considering Figure 1 and documents $A$ and $B$, we have that $(S_A \cup T_A) \cap (S_B \cup T_B) = \{E, H\}$, and, $(S_A \cup T_A) \cup (S_B \cup T_B) = \{D, E, F, G, H, I, J, M\}$, thus, $amsler(A, B) = \frac{2}{8}$.

## 4 Classifiers

For each collection studied we developed several classifiers based on the three bibliometric similarities defined above, which we call *bibliometric based classifiers*. These were compared to traditional text-only classifiers. In both cases, we trained classifiers based on the $k$-Nearest Neighbor ($k$NN) and the Support Vector Machine (SVM) methods. These methods were chosen because experiments in previous work (Joachims 1998; Sebastiani 2002) have shown that they are two of the most successful methods for classifying documents. Besides, Yang et al (Yang and Liu 1999) have shown that the two methods are robust to skewed category distribution, which is a common characteristic in document collections.

### 4.1 The $k$NN Classifier

A $k$NN 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 $k$NN algorithm (Yang 1994), to 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 training document $d_t$ belongs to class $c_i$ and 0 otherwise. The classifier assigns to test document $d$ the class with the highest score. Ties due to the absence of neighbors are handled as described in Section 4.3.

In our experiments we learned $k$NN classifiers for each bibliometric similarity measure by substituting function $similarity(d, d_t)$ in Eq. (4) for the value of the corresponding bibliometric similarity function for the pair $(d, d_t)$ (Equations 1, 2, 3).

Text-based $k$NN classifiers were learned using the cosine measure as the similarity function. In this case, each document $d$ is represented as a vector of term weights, that is, $d = w_{d,0}, w_{d,1}, ..., w_{d,M}$ for a vocabulary of $M$ terms. Thus, the similarity measure corresponds to the cosine of the angle between the two vectors. We use TF-IDF (Salton and Buckley 1988) as the weight of a term $t$ in a document $d$, defined as:

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

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 $k$NN classifiers used.

### 4.2 The SVM Classifier

The SVM classifier (Joachims 1998) works over a vector space where the problem is to find a hyperplane 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 (Chang and Lin 2001). The vector space of our SVM

classifiers corresponds to different sets of vectors, according to the evidence available, as described in the following paragraphs.

Since bibliometric measures are functions that map pairs of documents into real numbers, we can represent the similarity of any pair of documents $(d_i, d_j)$ through a doc-doc matrix $M$ that we refer to as *bibliometric similarity matrix*. Thus, given a similarity matrix $M$, $M_i$ represents the document vector $d_i$ and $M_{ij}$ represents the value of the similarity between $d_i$ and $d_j$. For the bibliometric SVM classifiers, we used the document vectors derived from the corresponding bibliometric similarity matrix.

Similarly, for the text-based SVM classifier we used the document vectors derived from a matrix where lines correspond to documents, columns correspond to terms and the value of each cell is the TF-IDF value of the corresponding pair $(document, term)$.

### 4.3 Documents Containing No Information

Some test documents do not contain enough information for either bibliometric based or text based classifiers. We will refer to this kind of documents as *no-information cases*. In order to minimize classification error, the classifiers always assign the most popular class to these documents and we refer to this assignment strategy as *default classification*[4] from now on.

Regarding text information, this can happen due to the fact that some web pages can be composed only of graphical elements such as images and vector animations or textual content that is completely discarded after the process of feature selection. Regarding link information, no-information documents occur when a document has zero

[4] This is a strategy used to reduce classification error, especially in collections with very skewed class distribution such as those studied here. A good discussion in this matter can be found in (Witten and Frank 2005).

similarity with any other document, when using a bibliometric similarity measure. For a document $d_i$ to have a positive bibliometric similarity to another document $d_j$ there must be at least a third document that is linked to, or by, both $d_i$ and $d_j$. This means that $d_i$ and $d_j$ must have source, or target, documents in common (see Section 3).

## 5 Collections Used in the Experiments

In this section we describe the collections used in our experiments. Unfortunately, there is no widespread accepted benchmark of documents with link information to be used by classification algorithms. Thus, in this paper we use subsets of three distinct collections hosted on the Web. The three collections are presented in the following sections.

### 5.1 The ACM8 Collection

We used a subset of the ACM Digital Library[5], which we obtained from the Association for Computing Machinery (ACM) by agreement. Hereafter, we refer to this subset as the *ACM8 collection*. Note that although the documents are also hosted on the Web, we treat them as scholarly articles. Thus, we only consider information extracted from the documents themselves and not any additional material about them, available in the ACM portal. All individual words in the text contained in the title and abstract, when available[6], were used to index the documents. However, we made no distinction about the place in the document where words occur. We did not use stemming, stopword elimination, nor feature selection in this collection.

[5] `http://portal.acm.org/dl.cfm`
[6] We did not have the complete text of the articles.

Link information was taken from the citations between papers. Many citations in the original ACM Digital Library could not be traced to the corresponding paper for several 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 (Lawrence et al. 1999). High precision and recall in this pre-processing phase is hard to achieve due to problems such as differences in the writing style for names of authors and conferences in the citations. This problem is particularly important in the case of the ACM Digital Library, since most citations were obtained with OCR after scanning, which introduces many errors, making the matching process even harder.

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 matched citations to distinct references. 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 copy of the ACM Digital Library we have is 11.23.

The resulting ACM8 collection is a set of 6,680 documents, labeled under the 8 largest categories of the ACM Digital Library taxonomy. These categories are, in descending order of their sizes: (1) *D-Software* - 1,847 documents, (2) *H-Information Systems* - 1,554 documents, (3) *I-Computing Methodologies* - 1,065 documents, (4) *B-Hardware* - 702 documents, (5) *C-Computer Systems Organization* - 658 documents, (6) *F-Theory of Computation* - 420 documents, (7) *K-Computing Milieux* - 237 documents and (8) *G-Mathematics of Computing* - 197 documents. The categories *A-General Literature*, *E-Data* and *J-Computer Applications* of the ACM taxonomy were not used

because they contain fewer than 20 documents in our subset. Each paper is classified into only one category.

Figure 2 shows the category distributions for the ACM8 collection. Note that the collection has a very skewed distribution, where the two most popular categories represent more than 50% of all documents.

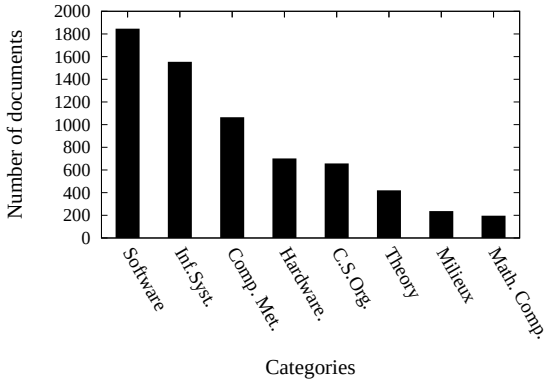
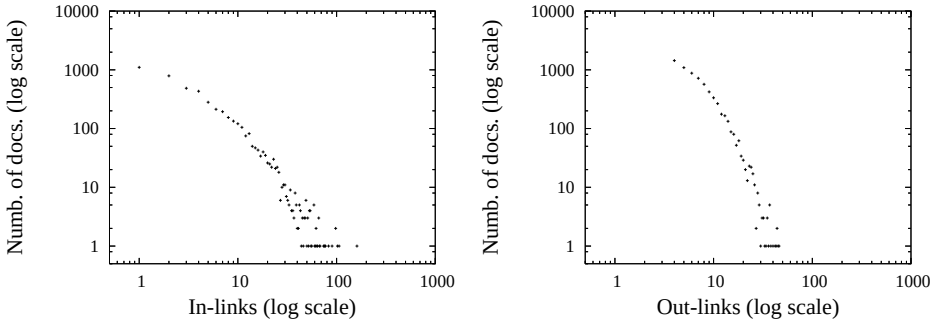



**Fig. 2** Category distribution for the ACM8 collection.

Figure 3 shows the distribution of in-links and out-links for the ACM8 collection. It can be seen that the majority of documents have fewer in-links than out-links.




(a) In-link distribution. (b) Out-link distribution.

**Fig. 3** Link distribution for the ACM8 collection.

Table 1 shows some statistics about links (citations and hyperlinks) and text in the ACM8 collections Links from the ACM8 collection articles to articles outside the ACM8 collection correspond to 77.8% of the links in the collection. Since we have no information about the external documents, in-links can be derived only from internal links, while out-links can be derived from all links. Thus the number of in-links in the ACM8 collection is 11,510, while the number of out-links is almost four times larger.

| Statistics | ACM8 | Cade12 | Wiki8 |
|---|---|---|---|
| Number of documents | 6,680 | 42,391 | 28,044 |
| Internal links | 11,510 | 3,830 | 186,844 |
| Links from external docs. | 0 | 554,592 | 1,584,587 |
| Links to external docs. | 40,387 | 5,894 | 1,223,228 |
| Documents with no in-links | 1,941 | 4,392 | 1,6862 |
| Documents with no out-links | 0 | 40,723 | 84 |
| Average in-links by document | 4.72 | 12.57 | 63.16 |
| Average out-links by document | 7.77 | 0.13 | 50,28 |
| Source of words | Title + Abstract | Title + Page body | Title + Page body |
| Text features | Single words | Single words | Single words |

**Table 1** Statistics for the ACM8, Cade12 and Wiki8 collections.

## 5.2 The Cade12 Collection

We used in our experiments a collection of pages indexed by the Brazilian Web directory Cadê[7], referred to as the *Cade12 collection*. All pages in the Cadê directory were manually classified by human experts. Since they were also indexed by the TodoBR search engine[8], we built the Cade12 collection by obtaining text and links directly from the TodoBR database. Thus, the source of in-links in Cade12 can be a page from outside Cade12. The content of each document in the Cade12 collection is composed of

[7] `http://www.cade.com.br/`

[8] TodoBR is a trademark of Akwan Information Technologies, which was acquired by Google in July 2005.

the text contained in the body and title of the corresponding web page, after discarding HTML tags.

The resulting collection is composed of 42,391 documents, containing a vocabulary of 191,962 distinct words. In our experiments we used information gain (*infogain*) as a feature selection criterion (Sebastiani 2002). Information gain is used to measure the capacity of a feature (term) to separate documents into classes. It is defined as:

$$infogain(t_k, c_i) = \sum_{c_i \in \mathcal{C}} P(c_i) \times \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} P(t, c) \times log \frac{P(t, c)}{P(t) \times P(c)} \qquad (6)$$

where $\mathcal{C}$ is the set of classes and the probabilities are interpreted on an event space of documents. For instance, $P(\bar{t}_k, c_i)$ denotes the probability that, for a random document $x$, term $t_k$ does not occur in $x$ and $x$ belongs to class $c_i$. The probabilities are computed over the training set. Information gain is used as feature selection – only the $m$ terms with the greatest information gain are used as features of the documents, for some arbitrary $m > 0$. Different values of $m$ were tested for the Cade12 collection and $m = 10,000$ was chosen since, in our experiments, it maximizes the micro-average $F_1$ measure for the $k$NN method. The remaining individual words were used as features for the classifiers. We did not filter stopwords nor use stemming in this collection.

The documents were labeled under the 12 first-level classes of the Cadê directory, here listed in decreasing order of their sizes: (1) *Services* - 8,958 documents, (2) *Society* - 7,183 documents, (3) *Recreation* - 5,693 documents, (4) *Computers* - 4,847 documents, (5) *Health* - 3,367 documents, (6) *Education* - 2,977 documents, (7) *Internet* - 2,561 documents, (8) *Culture* - 2,130 documents, (9) *Sports* - 1,942 documents, (10) *News* - 1,135 documents, (11) *Science* - 910 documents and (12) *Shopping* - 688 documents. Figure 4 shows the category distribution for the Cade12 collection. Note that the

collection has a skewed distribution and the three most popular categories represent more than half of all documents.

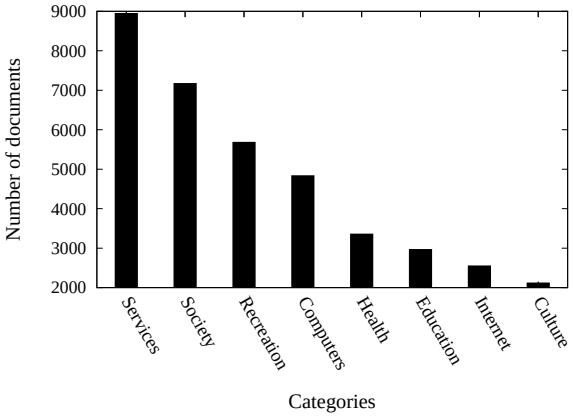


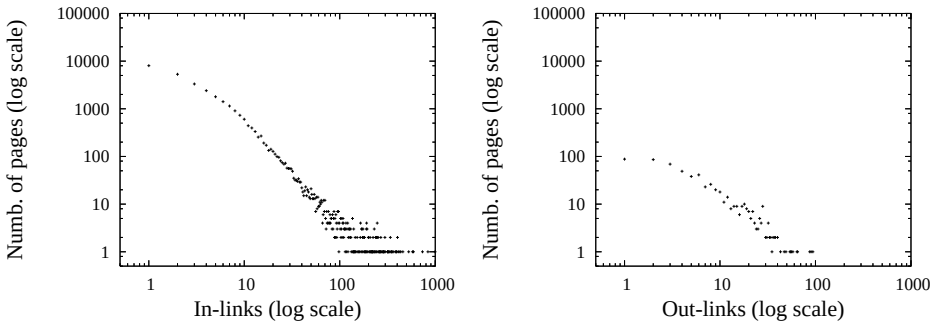

**Fig. 4** Category distribution for the Cade12 collection.

The links of the Cade12 collection were extracted from the set of 40,871,504 links of the TodoBR database. As observed in (Calado et al. 2006), 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 at the time it was crawled.

Note 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 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. As consequence of this process about 10% of the documents resulted with no in-link information.

Figure 5 presents the distribution of in-links and out-links in the Cade12 collection. Note that most pages have no out-links at all, but the majority does have in-links.




(a) In-link distribution. (b) Out-link distribution.

**Fig. 5** Link distribution for the Cade12 collection.

### 5.3 The Wiki8 Collection

Finally, we also used a subset of the English version of Wikipedia, hereafter referred to as the *Wiki8 collection*. It was extracted from the Wikipedia dump file, freely available on the Web[9]. The Wiki8 collection was obtained by first selecting Wikipedia categories of general nature, such that their topics could be easily assessed by the human judges that participated in the user study (described in Section 6.4.2). We chose the 8 following categories: (1) *History* - 17,782 documents, (2) *Politics* - 3,848 documents, (3) *Chemistry* - 1,841 documents, (4) *Philosophy* - 1,323 documents (5) *Biology* - 1,287

[9] `http://download.wikimedia.org/enwiki/2006816`

documents, (6) *Mathematics* - 976 documents, (7) *Astronomy* - 637 documents and (8) *Computer Sciences* - 350 documents. By gathering all the documents in each category we built the new collection with 28,044 documents. As shown in Figure 6, the category distribution in the Wiki8 collection is also very skewed, as more than half of the documents belong to *History*, the most popular class.

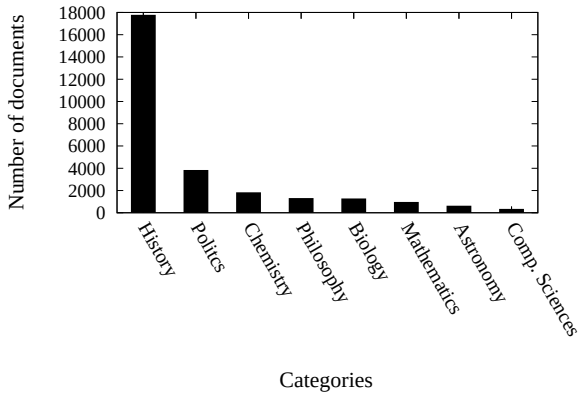
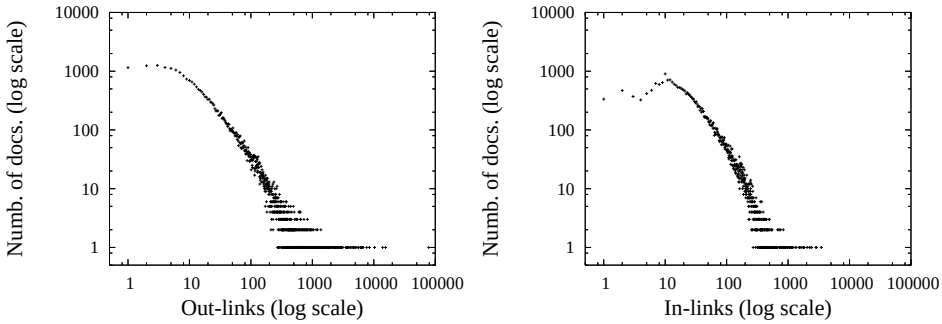



**Fig. 6** Category distribution for the Wiki8 collection.

The text of each Wiki8 document is derived from the text of the corresponding Wikipedia article, discarding HTML tags. We also removed meta-information about the class of the document. The resulting collection contains a total of 101,563 of individual words derived from title and body of documents. We used only 10,000 words with the best infogain. We did not make a distinction about the place in the documents where words occur. We did not filter stopwords and did not use stemming. We removed all links to category pages in the Wiki8 collection, since they would make classification obvious.

Figure 7 presents the distribution of in-links and out-links in the Wiki8 collection after removing category links. There are more in-links than out-links in the Wiki8

collection, but about 6% of the documents do not have in-links and 0.3% do not have out-links.




(a) In-link distribution. (b) Out-link distribution.

**Fig. 7** Link distribution for the Wiki8 collection.

## 6 Experimental Results

In this section we describe our experiments to assert the usefulness of bibliometric measures in the task of classifying documents in the three collections presented in Section 5. We used stratified ten-fold cross validation (Mitchell 1997) in all classification experiments. This consists in dividing a pre-classified collection of documents into 10 disjoint subsets of equal size. A classifier is then trained and tested 10 times, using each of the 10 subsets in turn as the test set, and using all the remaining data as the training set, trying to keep approximately the same proportions of labels as the original dataset. Stratified cross-validation is a common strategy used to compare the effectiveness of classifiers in many contexts (Mitchell 1997; Sebastiani 2002; Witten and Frank 2005).

For all comparisons reported in this work, we used the Wilcoxon signed-rank test (Wilcoxon 1945) for determining if the difference in performance was statistically significant. This is a nonparametric paired test which does not assume any particular

distribution of the tested values. All reported results achieved at least a 5% level of significance, except where explicitly stated otherwise.

In our experiments we used the same test and train sets for all classifiers in all 10 runs of cross validation. At the end of each run we computed, for each classifier, its micro-averaged and macro-averaged $F_1$ scores. The $F_1$ measure is defined as:

$$F_1 = \frac{2rp}{r+p}$$

where $p$ is the precision of the classifier and $r$ is its recall.

Micro-averaged $F_1$ corresponds to a global $F_1$ value obtained over all the $n \times m$ binary decisions, where $n$ is the number of test documents and $m$ is the number of classes. Macro-averaged $F_1$ is computed by first obtaining the $F_1$ values for each class individually and then averaging over all classes (Yang and Liu 1999). The final results of each experiment represent the average of the ten runs for both measures.

For each collection we used the text based classification of each method as the baseline for the method. Thus, the results of the $k$NN classifier using the cosine measure and the results of the SVM classifier with TF-IDF were taken as baselines.

In the following sections, we start by showing the experiments comparing classifiers based only on link information with text based classifiers, then we study the combination of link and text information in the classification task. Finally, we present experiments to better understand the failures that occurred in the classification process.

### 6.1 Experiments with Bibliometric Classifiers

Table 2 presents the micro-averaged and macro-averaged $F_1$ values for the bibliometric and text based classifiers on the ACM8 collection. The last two columns of the table show the percentage of gain over the text based classifier.

| Method | Similarity | $micF_1$ | $macF_1$ | Gains over text-based classifier | |
|---|---|---|---|---|---|
| | | | | $micF_1$ | $macF_1$ |
| $k$NN | 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** |
| | Text-cosine | 76.95 | 70.57 | – | – |
| SVM | Co-citation | 59.33 | 49.98 | -26.17 | -34.21 |
| | Bib-coupling | 80.72 | 74.59 | 0,4 | -0.18 |
| | Amsler | **83.08** | **77.08** | **3.37** | **1.46** |
| | Text-TF-IDF | 80.37 | 75.97 | – | – |

**Table 2** Macro-averaged and micro-average $F_1$ results for $k$NN and SVM classifiers applied over the ACM8 collection. Notice that the following differences between methods are not significant: Bib-coupling vs. TF-IDF (SVM), Amsler vs. TF-IDF (SVM with MacF1), and Bib-coupling (kNN) vs. Amsler (SVM). The best values for each classification method are shown in boldface.

Classifiers using the Amsler similarity were the best performers both for $k$NN and SVM 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. This is because there are many pairs of documents that have at least one out-link in common. In fact, 97% of the documents have bibliographic coupling with at least some other document in the collection. This means that not only there are many out-links in ACM8, but cited (target) documents tend to be cited by two or more documents. Also, when using the $k$NN classifier with the Amsler measure, most test documents have co-occurrent target documents in the training set. In fact, only 74 no-information cases were found. Thus, $k$NN rarely used the default classification in ACM8 as a means to decide the class of the test documents.

Both text based classifiers ($k$NN and SVM) presented good performances. This means that, despite being short, the text in documents of the ACM8 collection is not noisy.

Co-citation-based classifiers presented the worst results. Since there are few documents that share the same in-links and co-citation is a measure of the number of in-links two documents have in common, it is not sufficiently precise for the classifier to decide the class of a test document. In fact, 29% of the documents do not have co-citations. In the case of ACM8, this can be justified by the small number of in-links in the collection (fewer than twice the number of documents). For instance, 85% of the documents that $k$NN with the co-citation measure failed to classify have fewer than 4 in-links. On the other hand, of the 61.75% of documents that $k$NN with the co-citation measure correctly classified, only 28% of them have fewer than two in-links.

Table 3 presents results for the same set of experiments using $k$NN and SVM classifiers with bibliometric and text-based information applied to the Cade12 collection. In this collection, 70.49% of the documents are co-cited with other documents. Classifiers using the Amsler similarity or co-citation similarity achieved, gains of up to 69.8%.

| Method | Similarity | $micF_1$ | $macF_1$ | Gains (%) over text based classifier | |
|---|---|---|---|---|---|
| | | | | $micF_1$ | $macF_1$ |
| $k$NN | Co-citation | 68.51 | **75.60** | 36.9 | **69.8** |
| | Bib-coupling | 22.09 | 5.39 | -55.8 | -87.9 |
| | Amsler | **68.56** | 75.53 | **37.0** | 69.7 |
| | 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 | 51.47 |
| | Text-TFIDF | 54.18 | 49.38 | – | – |

**Table 3** Macro-averaged and micro-averaged $F_1$ results for $k$NN and SVM classifier applied over the Cade12 collection. We note that the differences between co-citation and Amsler are not significant, except for SVM with $macF_1$. Differences between kNN and SVM are not significant, except in the cases of text classifiers. The Best values for each classification method are shown in boldface.

Bib-coupling-based classifiers presented the worst results among the link based classifiers. This is because only 1% of the documents have at least one target document which is also a target of another document in the collection. In spite of this scarcity, all the classifiers achieved values of micro-average $F_1$ greater than 22% due to the default classification. This strategy works because of the large number of documents that belong to the most popular class. The small number of documents with bib-coupling values are due to the rareness of out-links in the collection.

Although the $k$NN classifier using the co-citation measure performed better than using the bib-coupling measure, and better than the text based $k$NN, about 30% of the documents were classified using the default classification. Thus, in order to make clear the true contribution of bibliometric information for this collection, we conducted an experiment removing the documents for which the classifier applied the default classification. Since the results between $k$NN and SVM classifiers presented on Table 3 are only slightly different, we used only the $k$NN classifier, which presented better performance. The results are shown in Table 4. Similar experiments were not conducted on the ACM8 and Wiki8 collections because the no-information cases in these collections are rare, corresponding to fewer than 2% of the documents.

| $k$NN with co-citation | $micF_1$ | $macF_1$ |
|---|---|---|
| Using Default Classification | 68.51 | 75.60 |
| Not Using Default Classification | **85.29** | **80.73** |

**Table 4** Results for the $k$NN classifier in Cade12 when using default classification (considering all documents) and considering only documents that are not no-information documents. The Best values are shown in boldface.

The difference between the two results shows that the lower values for macro-averaged and micro-average $F_1$ obtained in the first experiment involving all documents are mainly due to the lack of link information. In fact, whenever co-citation

information is available, its quality can be considered good for classification in the Cade12 collection. Only about 15% of the classification failures in the collection are due to wrong conclusions extracted from the co-citation measure itself. For example, one of the documents has class label *Society* but $k$NN assigned label *Recreation* to it because, among the $k$ documents that are most related to it by co-citation, 68.3% of them have class label *Recreation* and 31.7% have class label *Society*.

Table 5 presents results for both $k$NN and SVM classifiers on the Wiki8 collection. Bib-coupling based classifiers presented a performance very close to the Amsler based classifiers. In spite of the good performance of the $k$NN classifiers with the text cosine and co-citation measures, they present gains of up to 20.16%. Although there are more in-links than out-links in the Wiki8 collection, co-occurrent target documents are more evenly distributed over the collection than co-occurrent source documents. In fact, only 1% of the documents have no target document in common with any other document, while 12% of the documents do not have source documents in common with any other document. Since the bib-coupling measure is directly related to the number of target documents that two pages have in common, classifiers using this measure produce the best results.

| Method | Similarity | $micF_1$ | $macF_1$ | Gains (%) over text based classifier | |
|---|---|---|---|---|---|
| | | | | $micF_1$ | $macF_1$ |
| $k$NN | Co-citation | 81.3 | 68.43 | 0.5 | -0.1 |
| | Bib-coupling | 86.95 | **82.31** | 7.51 | **20.16** |
| | Amsler | **87.73** | 82.05 | **8.48** | 19.78 |
| | Text-cosine | 80.87 | 68.50 | – | – |
| SVM | Co-citation | 74.68 | 60.09 | -15.4 | -27.6 |
| | Bib-coupling | 86.07 | 80.61 | -2.5 | -2.9 |
| | Amsler | 85.66 | 80.84 | -3.0 | -2.5 |
| | Text-TFIDF | **88.27** | **82.99** | – | – |

**Table 5** Macro-averaged and micro-average $F_1$ results for $k$NN and SVM classifiers applied over the Wiki8 collection. Notice that the following differences between methods are not significant: Co-citation vs. Cosine (kNN), Bib-coupling vs. Amsler (kNN for macF1), Bib-coupling vs. TF-IDF (SVM), and Bib-coupling vs. Amsler (SVM). The Best values for each classification method are shown in boldface.

Since only 1% of the documents do not share any target document, the chances for a document to not have bib-coupling similarity to any training document is also small. In fact, only 0.03% of the test documents are no-information cases. So, as is in the case of the ACM8 collection, almost all mistakes and hits are consequence of the usage of the classification method used and not due to the default classification.

Note that the text based classifiers presented a much better performance in the Wiki8 collection, when compared to their performance in the other two collections. This is due to the high specificity of text information within each class, in which there are many terms that occur frequently in one class and are rare in other classes. For example, terms like *biology*, *cell*, and *cells* occurred at least in 30% of the documents of the class *Biology* and are almost nonexistent in the remaining classes. We can also find sets of discriminative terms like these for the other classes of the collection.

The quality of the text cosine measure in the Wiki8 collection is even more evident when we compare it to the quality of the cosine measure in the other two collections. This comparison was performed by computing the infogain of the terms in the three collections. For each collection we ranked the terms in descending order of their infogain values and computed the mean infogain of the top $k$ terms. Table 6 shows the values for $k$ equal to 100, 1 000 and 10 000 in each collection. We note that the mean values of infogain for the Wiki8 collection is greater than those of the other collections in all cases.

| Collection | Average Infogain for $k$ Best Terms | | |
|---|---|---|---|
| | $k = 100$ | $k = 1000$ | $k = 10000$ |
| Wiki8 | **0.038** | **0.013** | **0.0033** |
| ACM88 | 0.020 | 0.006 | 0.000126 |
| Cade12 | 0.012 | 0.0049 | 0.000125 |

**Table 6** Average infogain of the $k$ terms with best infogain in each collection.

In sum, our experiments show that in all three scenarios the classifiers based on bibliometric measures provided accurate results. Results indicate that both the density and distribution of the links found in the three collections and the information extracted from links may play an important role in classification tasks.

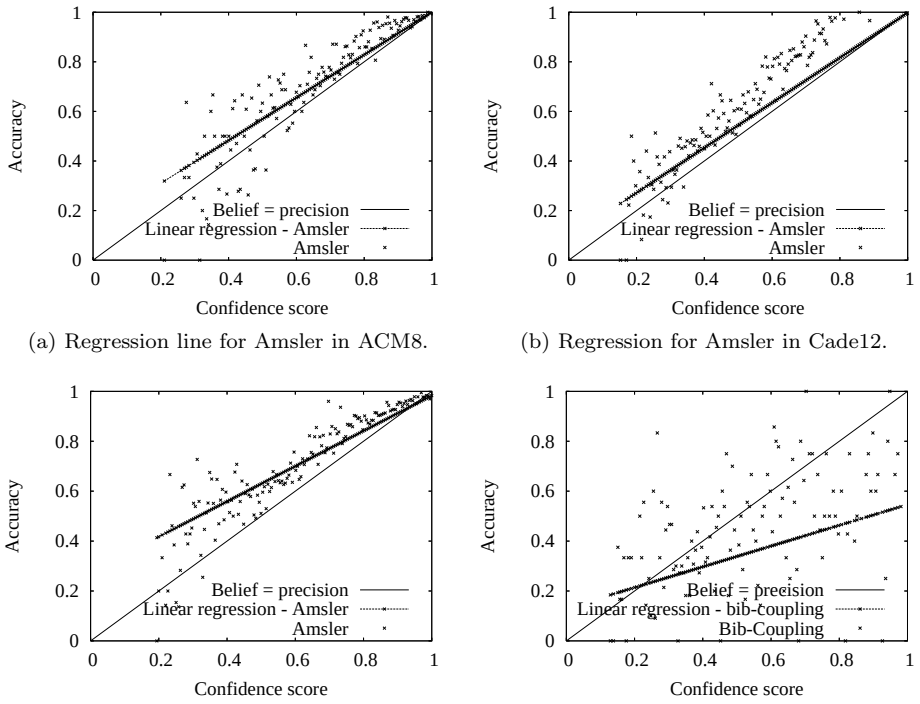
### 6.2 Combining Bibliometric and Textual Information

In this section we discuss how the bibliometric and textual information can be combined to produce systems that automatically discover the class of documents based on these two sources of information.

Classifiers using bibliometric similarity measures were shown to be effective, as observed in Section 6.1. However, another interesting possibility is to combine the outputs of these classifiers with text based classifiers to improve classification performance. To this end, it is important that the scores assigned by each classifier are reliable. An ideal classifier, regarding reliability, should provide belief estimates exactly proportional to its 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 correctly classify $p \times |\mathcal{D}_p|$ documents of the set $\mathcal{D}_p$. In this section we show that the reliability of classifiers is useful for combining their results in order to improve document classification.

In spite of not being ideal classifiers, the $k$NN classifiers using bibliometric measures do have the property of providing belief estimates proportional to their accuracy in ACM8, Cade12 and Wiki8 collections. Figures 8(a), 8(b) and 8(c) show the accuracy values obtained for belief degrees estimated by the $k$NN classifier, using the Amsler similarity measure. In all figures, the dashed lines are derived by linear regression

applied over the belief degree points, and the solid lines correspond to an ideal classifier for which the belief degree would correspond exactly to the accuracy obtained.




(a) Regression line for Amsler in ACM8.

(b) Regression for Amsler in Cade12.

(c) Regression for Amsler in Wiki8.

(d) Regression for Bib-coupling in Cade12.

**Fig. 8** Accuracy per belief degree. Graphics (a), (b) and (c) show the regression line for the Amsler similarity measure in ACM8, Cade12 and Wiki8 collections, respectively. Graphic (d) shows the regression line for bib-coupling in Cade12.

In all first three plots the regression lines are very similar to the line representing the reliability of an ideal classifier. It means that the values provided as belief degrees approximately correspond 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. Similar figures were obtained for the $k$NN classifiers using the other similarity measures in the three collections, which we do not include here to avoid repetition of arguments. The only exception occurs with the $k$NN classifier based on the bib-coupling

measure in the Cade12 collection, where the regression line clearly differs from the ideal line, as shown in Figure 8(d). This occurs because there are only few documents that have bib-coupling similarity to some other document in the collection, as discussed in Section 6.1.

Given that our bibliometric based classifiers are reliable, we now describe a method for combining them with other classifiers. Combination of results of classifiers is a well known method to boost classification performance. It is especially 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.

However, if one of the classifiers to be combined presents high accuracy and provides reliable estimations it is possible to use it as a guide in the combination process. 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). In such cases, it would be better to use the classification decision provided by the second classifier. This idea is formally presented in Figure 9.

The algorithm in Figure 9 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 1-10). For this, it obtains regression lines for $A$ (lines 3 and 5) and $B$ (lines 4 and 6).

It then finds the point $p$ where the lines cross (lines 7-10) and uses this point to determine which classifier can provide the best decisions (lines 11-13). In sum, decisions from classifier $A$ are preferable if it yields belief estimations greater than $p$.

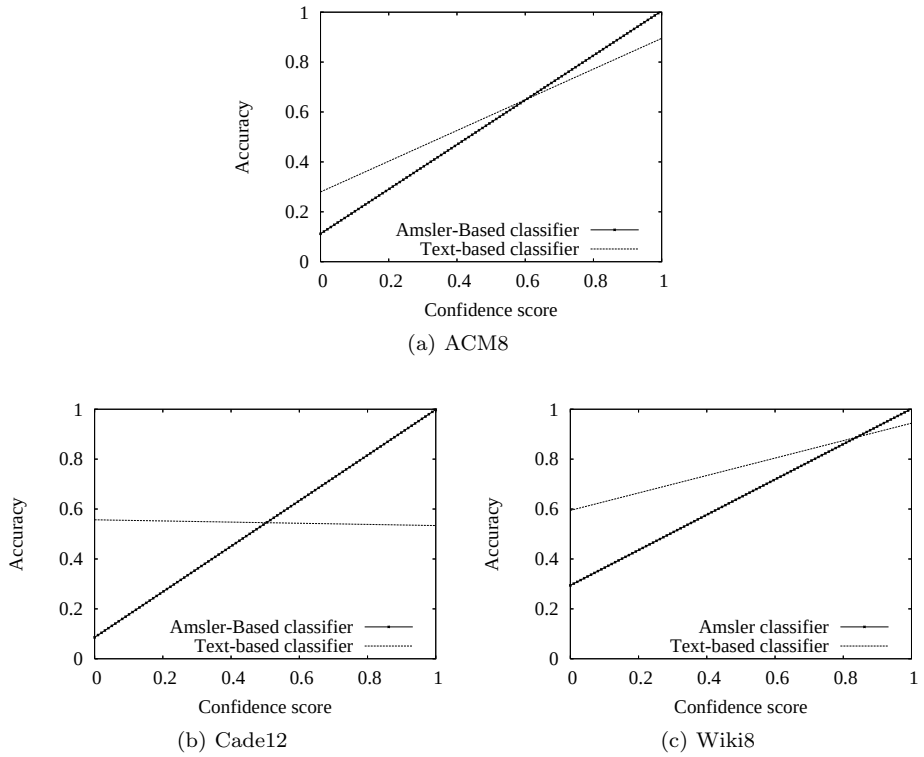
```
1 Let A be the most reliable classifier to be combined;
2 Let B be the least reliable classifier to be combined;
3 Let A_tr be a set of points {c_Ai, y_Ai}, where c_Ai represents the confidence of
      A in the classification given for document i in the training collection (
      0 ≤ c_Ai ≤ 1), if the classification provided by A for document i is
      correct, then y_Ai is 1 and is 0 otherwise;
4 Let B_tr be a set of points {c_Bi, y_Bi}, where y_Bi is 1 if the classification
      provided by B for document 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) p = 0;
9       else p = 1; }
10   else p = (b−d)/(c−a);
11 for each document i in the test collection {
12   if (c_Ai > p) classification of document i is given by A;
13      else classification of document i is given by B; }
```

**Fig. 9** Combining the results of the classifiers.

We applied the algorithm of Figure 9 to the three collections we studied. In each collection we used the $k$NN classifier based on the Amsler similarity measure as the first classifier to be combined because the Amsler measure presented the best results over all other similarity measures in all three collections (see Section 6.1). As the second classifier to be combined we used the text based classifier that performed better in each collection.

Figure 10 shows the regression lines obtained by applying the algorithm of Figure 9 to each collection. The dashed and solid lines on the figures correspond to the lines computed by the linear functions derived from line 5 of the algorithm. For all three collections the classifiers based on the Amsler measure were used as guides since they are more reliable than text based classifiers. For this reason we refer to this strategy as *link combination*.

For the ACM8 collection we can see that the regression lines for the Amsler based classifier and the best text classifier are very similar. This means that for any belief degree of the Amsler classifier the performances of both Amsler based and text based




(a) ACM8

(b) Cade12

(c) Wiki8

**Fig. 10** Regression lines for belief degrees of Amsler based $k$NN classifier and for belief degrees of TF-IDF based classifier in the three collections.

classifiers are very similar. Consequently, in this case the link combination is not able to present much gain over the Amsler classifier as will be seen later in this section.

As Figure 10(b) shows, the regression line for the text based classifier for the Cade12 collection is close to a constant function. This is a consequence of the poor quality of text information in the Cade12 collection. On the other hand, the regression line for the $k$NN classifier based on the Amsler measure is very similar to an ideal classifier. Note that the $k$NN classifier based on the Amsler measure is reliable in spite of the many no-information cases present in the Cade12 collection, as discussed in Section 6.1. This means that the method and the default classification did together a good job with respect to both reliability and accuracy of the classifier. So the belief degree of the

Amsler based classifier can be used to drive the combination of results. Both lines cross each other at point $p$ corresponding to a belief degree of 50%. If the belief degree of the Amsler based classifier falls at $p$ or above, the class it indicates is adopted. If this is not the case, the class pointed by the text based classifier is preferred. Contrary to what happened in the ACM8 collection, link combination in this case is expected to perform better than text and Amsler based classifiers considered in isolation.

Figure 10(c) shows the regression lines for the Wiki8 collection. As can be seen, the $k$NN classifier using the Amsler measure is more reliable than the text based classifier for this collection also, since its belief degree values are similar to its accuracy values. However, the regression lines of the two classifiers cross each other for values of belief degree superior to 0.85. This means that the link combination method will take the output of the text based classifiers for most of the belief degrees. Also, the accuracy for both classifiers is similar for belief degrees superior to 0.85, thus link combination for this collection is expected to perform only slightly better than the text based classifier alone.

Table 7 presents the results of the link combination strategy for the three collections. For comparison purposes, the results shown in Tables 2, 3 and 5 for the Amsler based classifiers are shown in the table.

Table 7 also shows the results for the *Bayesian combination* method described in (Calado et al. 2006). This method was used originally with a subset of the Cadê collection different from the Cade12 collection used here and presented good improvements over the text and link based classifiers in isolation. Thus, we use this method as the baseline for comparison with the link combination method just described. The Bayesian method uses a Bayesian network model to derive the probability $P(f_j|c)$ that a test document $j$ belongs to class $c$ (Turtle and Croft 1991). This probability is used

| Collection | Methods | $micF_1$ | $macF_1$ | Gains over link classifier (%) | |
|---|---|---|---|---|---|
| | | | | $micF_1$ | $macF_1$ |
| ACM8 | $k$NN-Amsler | 84.43 | 79.41 | – | – |
| | SVM-TFIDF | 80.37 | 75.97 | -4.8 | -4.33 |
| | Bayesian Comb. | **87.04** | **82.76** | **3.0** | **4.2** |
| | Link Comb. | 85.76 | 81.37 | 1.6 | 2.46 |
| Cade12 | $k$NN-Amsler | 68.56 | 75.53 | – | – |
| | SVM-TFIDF | 54.18 | 49.38 | -20.97 | -34.62 |
| | Bayesian Comb. | 76.51 | 79.29 | 11.6 | 4.97 |
| | Link Comb. | **78.04** | **80.39** | **13.82** | **6.43** |
| Wiki8 | SVM-TFIDF | 88.27 | 82.99 | – | – |
| | $k$NN-Amsler | 87.15 | 82.05 | -1.26 | -1.13 |
| | Bayesian Comb. | **90.75** | **87.28** | **2.8** | **5.16** |
| | Link Comb. | 90.44 | 86.44 | 2.45 | 4.15 |

**Table 7** Macro-averaged and micro-average $F_1$ results for combining approaches in the ACM8, Cade12 and Wiki8 collections. Differences between Bayesian Comb. and Link Comb. are not statistically significant. The best values for each collection are shown in boldface.

to directly combine the belief degrees provided by the text based and the citation-link based classifiers, and is defined as:

$$P(f_j|c) = \eta\Big[1 - \big(1 - W_t\ P(t_j|c)\big)\big(1 - W_l\ P(l_j|c)\big)\Big] \tag{7}$$

where $\eta$ 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 $\frac{W_l}{W_t} = 1.1$ for the ACM8 collection, $\frac{W_l}{W_t} = 0.20$ for the Cade12 collection and $\frac{W_l}{W_t} = 1.1$ for the Wiki8 collection. By employing these weight ratios the *Bayesian combination* achieved its best performance in the experiments.

As we can see in Table 7, the combination methods presented similar results. As we stated before, for the case of the ACM8, the regression lines of both text and Amsler based classifier are similar. Thus, link combination could not improve much by using the output of text classifier for belief degrees of the Amsler classifier smaller than about 0.7, which is the belief degree where the two lines cross.

For the Wiki8 collection, the accuracy of the text based classifier is superior to the Amsler based classifier for all belief degrees inferior to approximately 0.85. Also the accuracy above this point is very similar for both classifiers, thus link combination could not improve much the result by choosing the output of the Amsler classifier for belief degrees superior to 0.85.

The gains obtained from any combination strategies seem, at first, quite small in both ACM8 and Wiki8 collections. However, let us suppose that we had a perfect combination method that would be able to choose between the two classifiers the one which assigned the right class, whenever one of them gives a right assignment. The $macF_1$ and $micF_1$ average values for such perfect combiner can be obtained with 10-fold cross validation, using the same folds that were used in all the other experiments for each collection.

When comparing this perfect combiner to the results obtained on each collection we realize that the possible improvements in results are not so high. Table 8 shows the values of $micF_1$ and $macF_1$ for the perfect combiner and its gains over the best combination method obtained for each collection. The results of the perfect combiner correspond to the upper limits for the combination of the results of classifiers. As we can see, there is room for enhancement, but the possible gains over the ones obtained would be small, if we consider the optimal case.

| Collection | Methods | $micF_1$ | $macF_1$ | Gains (%) over best link based classifier | |
|---|---|---|---|---|---|
| | | | | $micF_1$ | $macF_1$ |
| ACM8 | Bayesian Comb. | 87.04 | 82.76 | – | – |
| | Perfect Comb. | 91.24 | 88.50 | 4.8 | 9.6 |
| Cade12 | Link Comb. | 78.04 | 80.39 | – | – |
| | Perfect Comb. | 83.99 | 86.58 | 7.62 | 7.7 |
| Wiki8 | Bayesian Comb. | 90.75 | 87.28 | – | – |
| | Perfect Comb. | 94.22 | 92.32 | 3.8 | 5.7 |

**Table 8** Micro-averaged and macro-averaged $F_1$ results for the perfect combiner and the best combination method (Bayesian or Link Comb.) used in each collection.

### 6.3 Discussion of the Results

The documents of the ACM8 collection are more coherent in presentation, style, and vocabulary. Also, its citations are used in a more rigorous way. Thus, the ACM8 collection can be considered of a technical nature. On the other hand, the Cade12 collection presents pages characterized by great freedom regarding structure, style, and content. Its links may be used for the same functionalities as citations in the ACM8 collection, but they are also used for additional functionalities not necessarily associated with topic similarity. Further, the lack of a systematic revision process leads to low quality and unreliable textual content. Finally, the Wiki8 collection presents several characteristics we would expect from an encyclopedia, such as a regular use of citations to articles, not only related by content, but also by other contextual aspects such as dates and places. The existence of an active revision process contributes for high quality textual content and some coherence regarding style, despite the possibly large number of reviewers.

As expected, text classifiers were more effective when used with the Wiki8 and ACM8 collections than with the Cade12 collection, which reflects the observed quality difference in textual content. As a consequence, the impact of the appropriate bibliometric classifier on combined classification was larger for the Cade12 collection due to

the poor performance presented by text classifiers. Note that even for the web pages in the Cade12 collection, where links may be used for much different functionality, the in-links were useful for determining topic relatedness. Probably, the noise introduced by additional functionalities had small impact on the creation of spurious relationships between pairs of source pages. The appropriate bibliometric classifiers also performed better than the text based classifier for more regular collections (ACM8 and Wiki8) despite the good quality of text information. However, in these collections the impact of bibliometric classification over the textual one was smaller.

Another interesting aspect observed in our study was the different distributions of in-links and out-links in the collections and the impact they have in bibliometric classifiers. In fact, the most important difference between the Wiki8 and ACM8 collections was that while the Wiki8 collection derives good classification results from using both in-links and out-links, for the ACM8 collection the best results were achieved by using out-links. Unlike the ACM8 collection, in-links yielded the best results in the Cade12 collection. Also, as the experiments show, the number of sources or targets is a necessary but not sufficient condition to determine the quality of bibliometric information for classification. It is also necessary to have an appropriate distribution of the co-occurrence of sources or targets in the collection.

In particular, the prior observation related to the ACM8 collection should be taken carefully, given the small number of links in our sample. Note that this scarcity of links is a consequence of problems in the link extraction process and should not be taken as a general characteristic of a digital library of scientific papers. Also note that the described results represent experiments with taxonomies composed by a small number of broad classes.

## 6.4 Further Understanding the Classification Failures

In this section we investigate the possible reasons for the classification failures produced by the bibliometric classifiers. We performed two types of studies to evaluate the origins and meaning of the failures produced. First, we use information available in the ACM8 collection to study the failures that are consequence of documents containing multiple classes. Second, we perform a more comprehensive study with users to understand the failures produced in the three experimented collections.

### *6.4.1 Possible Multi-classification Cases*

Since the $k$NN classifier using the Amsler similarity measure was the best bibliometric based classifier, we decide to further investigate its cases of misclassification. We found that in 58% of the failures, the class assigned by the documents' authors appears as the second most probable class assigned by the classifier in ACM8. Although all documents of the ACM8 collection were assigned to only one first level class of the ACM hierarchy by their authors, we intended to investigate if some of the above cases could be considered correct in a multi-classification setting, as follows.

In the ACM computing classification system tree (ACM 1998), the associations between classes are declared explicitly. For instance, Figure 11 shows an entry in the classification system tree describing the subclass *I.7 - Document and Text Processing*. The labels appearing on the right of the subclass title (*H.4* and *H.5*) indicate that a document classified under the subclass *I.7* is also related to the subclasses *H.4* and *H.5*. As a consequence, a document classified under the class *I.7* (or its subclasses) might also be classified under the classes *H.4* and *H.5* (or its subclasses).

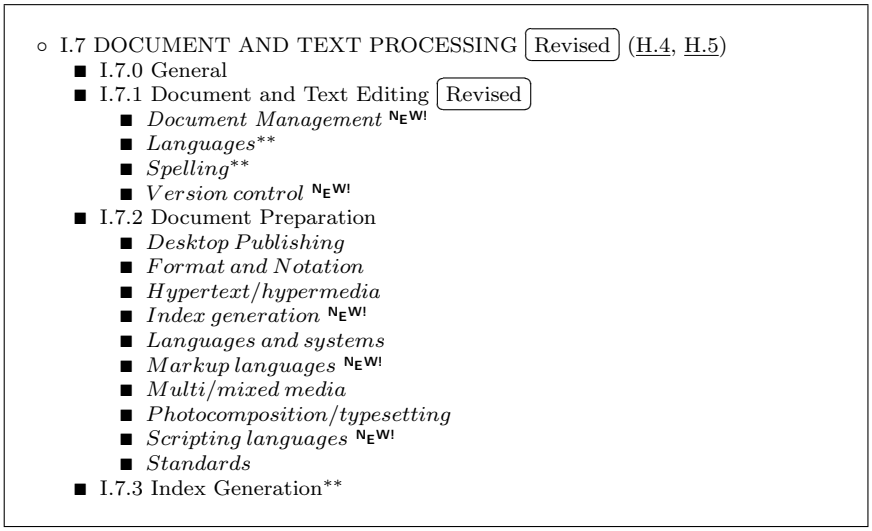



**Fig. 11** Part of the ACM classification tree showing relations among subclasses of different first-level classes.

To find the proportion of misclassification cases that could be considered correct assignments in a multi-classification setting, we have to determine the misclassified documents which could be assigned to multiple classes, among them, the one chosen by the $k$NN classifier. Thus, given a test document $d_t$ for which the $k$NN classifier failed, let $L_{kNN}$ be the list of the the subclasses of the $k$ most similar documents to $d_t$ and let $L_{auth}$ be the list of subclasses of $d_t$ assigned by its authors. By inspecting both lists, we can find pairs of subclasses ($c_i$,$c_j$), where $c_i \in L_{kNN}$ and $c_j \in L_{auth}$, such that $c_i$ and $c_j$ or some of their ancestors are explicitly related in the ACM hierarchy. Once we find these pairs, we select $c_i$ as a potential class of $d_t$ if its first-level ancestor was assigned by the $k$NN classifier and its occurrence count in $L_{kNN}$ is greater than a certain threshold $f$, determined experimentally after sampling some documents. In our experiments, we used $f = 3$.

Table 9 shows an example of a misclassified document that was assigned to subclass I.7.2 by its author ($L_{auth} = \{I.7.2\}$ in this case). The second column shows $L_{kNN}$, the subclasses of the $k$ nearest neighbors of this document. The numbers in parentheses correspond to the occurrence of each subclass. The subclasses in boldface occurred more than three times in $L_{kNN}$ and have as ancestor class $H.5$ that is related to subclass $I.7$. This class, by its turn, is ancestor of the class assigned by the author of the document (see Figure 11). Thus, if the $k$NN classifier assigns class $H$ to this document, this should be considered a correct decision in a multi-classification setting.

| Author assigned | Subclasses in the $k$ most similar documents |
|---|---|
| I.7.2 | **H.5.1** (18), **H.5.2** (10), H.2.4 (7), H.2.1 (5), **H.5.3** (5), **H.5.4** (5), H.3.1 (4), H.2.8 (3), H.3.4 (3), H.3.3 (3), H.1.2 (3), H.2.3 (2), C.0 (1), H.3.7 (1), H.4.3 (1), D.2.13 (1), D.2.6 (1), H.3.2 (1), H.3.5 (1) |

**Table 9** Example of the detection of a candidate for multi-classification. Subclasses in boldface occurred more than 3 times in the list of the $k$ nearest neighbors of the test document.

Table 10 summarizes the cases of misclassification that could be considered correct decisions in a multi-classification setting. For obtaining this data, we used the $k$NN classifier with the Amsler similarity measure. Additionally, we manually checked them to confirm that the document could be really considered as pertaining to both classes. The second column of the table contains the total of misclassified documents per class. The third column contains the number of failures that were considered multi-classification cases. The fourth column contains the percentages of these cases. As we can see, 24% of the misclassifications should be considered correct decisions if we had used multi-classification.

| Class | $k$NN Failures | | |
|---|---|---|---|
| | Total failures | Multi-classification | |
| | | Total | % |
| B | 123 | 43 | 34.95 |
| C | 168 | 50 | 29.76 |
| D | 175 | 55 | 42.8 |
| F | 159 | 34 | 31.42 |
| G | 71 | 12 | 16.9 |
| H | 97 | 18 | 18.56 |
| I | 107 | 19 | 17.76 |
| K | 72 | 2 | 2.78 |
| Multi-classification Average | | 29.12 | 24.36 |

**Table 10** The number of $k$NN classification failures by class and the number and percentage of these failures that can be considered multi-classification cases.

*6.4.2 User Study*

Motivated by the difficulty in improving classification results in our test collections through our combination methods, even when bibliometric relations exist, we decided to perform a user study to investigate the difficulty of the misclassification cases. In what follows, we refer to classes originally assigned to documents of the collections by a specialist or by the document's author as the *correct class* of the document. Whenever the classifier or a volunteer who took part in the experiment assigns to a document a class distinct of its original one, we say that a *wrong* class was assigned to it.

When a bibliometric based $k$NN classifier assigns a wrong class to a test document, it does so because most of the training documents related to the test document by the bibliometric measure belongs to a class other than the correct one. By inspecting some of these cases, we suspected that even humans would have difficulty in classifying them. This assumption is reinforced by studies on inter-indexing inconsistencies in intellectual indexing research. For instance, Qin has investigated dissimilarities between citation-semantic and analytic indexing, emphasizing the high degree of variation regarding the keywords and semantic categories used to describe document contents (Qin 2000).

To test this assumption we conducted another experiment in order to study human classification of those unsuccessful cases. Since we are interested in misclassification cases related to bibliometric information we removed all the no-information documents. We then applied the $k$NN classifier with the Amsler similarity measure to our test collections using ten fold cross-validation. For each collection, we grouped the classifier results by the corresponding categories, such that each class could be considered as a stratum from where we derived the samples. The size of the sample for each class was determined using the proportion of hits and failures of the classifier such that samples represent the original collection with 95% of confidence (Cochran 1977). We then obtained proportional samples of random cases that were hits and failures of the classifier for each class and finally collected all the obtained documents to form a single sample for each collection.

Given the large number of volunteers necessary for classifying the documents with 95% confidence, we used two distinct samples from the three collections for a total of $1,234$ documents. In the first one, we sampled classifier errors and hits from the classes easiest and hardest to classify. Thus, these classes represent the lower and upper performance bounds achieved by our best method. The sample was composed by 204 ACM8 documents, 239 Cade12 pages, and 172 Wiki8 articles. Table 11 shows the classes used in this sample. The second sample consisted only of errors of the classifier, comprising 214 ACM8 documents, 323 Cade12 pages, and 82 Wiki8 articles.

| Collection | Highest accuracy (#docs) | Lowest accuracy (#docs) |
|---|---|---|
| ACM8 | D (90) | K (114) |
| Cade12 | Sports (73) | Culture (166) |
| Wiki8 | Politics (139) | Math (33) |

**Table 11** Classes with highest and lowest accuracy in each test collection. The number of documents for each class is shown between parenthesis.

For each sample the documents were distributed in pools in such way that each one was evaluated by two distinct volunteers, for a total of 32 graduate and undergraduate Computer Science students. Note that documents from the ACM8 collection were assigned only to Computer Science graduate students given the expertise required to distinguish among Computer Science topics.

For the two samples of the ACM8 collection people could analyze the title, authors, keywords, abstract (when available), the conference name, and the links' text. Evaluators of samples of the Cade12 collection had access to the full page (which included images and photos). For the the Wiki8 collection samples, however, the only information available was the raw text of each document. For each document, we asked the volunteers to choose the classes they judge most appropriate. We also asked them to rank the classes, such that we could know which one would be chosen in a unique label classification setting.

The experiment just described were conducted with the following specific objectives:

1. To compare the difficulty volunteers have in classifying any document in our test collections to the difficulty in classifying only documents previously misclassified by our best automatic method.
2. To investigate consensus among the volunteers regarding their classification.

Table 12 shows the results regarding our first objective. It shows the difference between classification made by volunteers in the two samples. Table 12(a) shows results for the sample that contains both errors and hits of the automatic classifier. We can see that in the majority of the human classifications for this sample the correct class appears as the first class assigned by the volunteers. The percentage values are even

greater for the classification cases where the correct class appears in any order in the rank of classes produced by a volunteer.

(a) Sample including classifier errors and hits

| Choice regarding the correct class | ACM8 | Cade12 | Wiki8 |
|---|---|---|---|
| The first class is the correct one | 67.5% | 55.7% | 73.8% |
| The correct class is among the chosen classes | 75.0% | 82.6% | 86.2% |

(b) Sample containing only classifier errors

| Choice regarding the correct class | ACM8 | Cade12 | Wiki8 |
|---|---|---|---|
| The first class is the correct one | 38.3% | 43.3% | 41.5% |
| The correct class is among the chosen classes | 59.1% | 72.3% | 81.7% |

**Table 12** Results of classifications made by volunteers.

Table 12(b) shows the results for the second sample composed by only errors of the automatic classifier. We have that the percentage of classifications that assigned the correct class as the first class are notably smaller than those for the first sample, not reaching even 45%. This confirms our expectation about the difficulty of classifying the sample composed only by documents that were not classified correctly by the automatic classifier. However, the percentages of classifications that volunteers chose the correct class as one of the possible classes of the document are significant and suggests that many documents that are misclassified, specially in the Cade12 collection, might belong to multiple classes. Nevertheless, these percentages are still inferior than the corresponding ones for the first sample. This lead us to conclude that the sample with only errors of the automatic classifier is harder for human judges even when human classifiers perform multi-classification.

Table 13 shows the results regarding our second objective, which was to analyze the consensus between volunteers. We note that there is more consensus between the

two volunteers[10], in Table 13(a) where most of the documents were classified correctly by the automatic method. The majority of the documents received the same first class from both volunteers but this is not the case for the sample containing only classifier errors. In both samples, few documents achieved consensus about the correct class being the most appropriate class (first option). However, this kind of consensus is even less frequent in the sample composed only by classifier errors. Thus, the experiment shows that consensus in much harder to be achieved for documents that the automatic classifier failed to classify.

(a) Sample including classifier errors and hits

| Consensus about... | ACM8 | Cade12 | Wiki8 |
|---|---|---|---|
| The same classes as the first option | 54.7% | 52.5% | 76.3% |
| The correct class as the first option | 47.8% | 39.2% | 63.1% |

(b) Sample containing only classifier errors

| Consensus about... | ACM8 | Cade12 | Wiki8 |
|---|---|---|---|
| The same class as the first option | 35.0% | 38.1% | 39.0% |
| The correct class as the first option | 21.5% | 28.8% | 34.1% |

**Table 13** Percentage of documents for which users reached consensus in the test collections.

We also investigated the opinion of the users for the documents that we denominate *hard decisions* of the classifier. A hard decision corresponds to misclassified documents for which the classifier assigns the correct class as the second choice and the probability difference between the first and second choices was very small (less or equal to 0.2 in our experiments). The second line of Table 14 shows that the majority of the documents that are hard decision cases were misclassified or received a two-class vote by at least one human evaluator. Also, only a few hard decision cases were correctly classified

[10] Remember that each document in both samples occurs in two distinct pools in the experiment and thus was evaluated by two distinct volunteers.

by all volunteers. Thus hard decision cases are really very difficult even for human classification.

| Hard decisions | ACM8 | Cade12 | Wiki8 |
|---|---|---|---|
| Percentage of hard decision cases | 13.08 | 23.83 | 26.83 |
| Percentage of hard decision documents wrongly classified or that received more than one class | 71.4 | 72.72 | 54.54 |
| Percentage of hard decision documents correctly classified by all volunteers | 25.0 | 19.48 | 31.81 |

**Table 14** Human classification of documents that correspond to hard decisions for the automatic classifier.

The above results and observations indicate that the failures of the classifier based on bibliometric measures are difficult cases. Even human classification did not achieve much success. Further, consensus on the correct class is very rare among human evaluators and the hard decision cases for the classifier are even harder ones to correctly classify.

## 7 Conclusions

In this work we studied the usage of classifiers based on bibliometric similarity measures for classifying web collections. As case studies we chose subcollections of three important and popular collections: a digital library of scientific articles, a directory of web pages and an on-line encyclopedia. We compared the performance of bibliometric-based classifiers and text-based classifiers. Experiments have shown that bibliometric based classifiers performed better than text based classifiers in two of the collections studied and presented results only marginally inferior to text based classifier in the collection derived from the encyclopedia.

Extensive experimentation and analytical studies were conducted to better understand the characteristics of link distribution among documents that affect the per-

formance of bibliometric based classifiers which provided a deeper understanding on how this information can be explored, as well as on its limitations. We concluded that distribution of the co-occurrence of source documents and target documents among documents and classes have great impact on the performance of bibliometric classifiers and are responsible for the existence of the bibliometric information. For example, in the Cade12 collection, the distribution of links is in a such way that the great majority of documents do not have target documents in common with any other document and about 70% of the documents are co-cited with other documents. Thus bibliometric based classifiers have a limited accuracy in this collection dictated by the lack of information. Despite this, our experiments show that classifiers using co-citation similarity measure still performed much better than text based classifiers in Cade12.

Another factor affects the bibliometric based classifiers. It is what we called the coherence between the class of the document and the classes of its nearest neighbors. When linkage information is not scarce, the classifier fails if the most common class among the nearest neighbors of the document is distinct from the true class of the document. In some collections this may occur because the document naturally belongs to more than one class. This happens, for example in the ACM8 collection and we conducted an experiment that allowed us to conclude that many misclassified documents are indeed documents that could also be assigned to the class indicated by the bibliometric classifier. We hypothesized that most of the failures could be also difficult for humans to classify correctly. A user study showed that most of the failure cases are really hard to solve and consensus about the correct class of documents is hard to achieve.

Textual content and bibliometric relations are complementary sources of information. In this work, we also take advantage of this fact and used a procedure to com-

bine the results of text based and bibliometric 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 $F_1$ of up to 13.8% in a web directory, although gains were much less significant in the digital library and encyclopedia datasets used. We also presented for each collection the maximum values of $macF_1$ and $micF_1$ that could be achieved if a perfect combination method could be used.

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 bibliometric based classification to automatically expand web directories. Finally, we intend to investigate ways of combining different evidences in a same classification method and compare this approach to the combination of classifier results that we used in this work.

## 8 Acknowledgments

This work was supported by the Brazilian National Institute of Science and Technology for the Web (Grant MCT/CNPq 573871/2008-6), Project FCT IR-BASE (Grant POSC/EIA/58194/2004), Project InfoWeb (MCT/CNPq/CT-INFO 550874/2007-0), Project InWeb (Grant 573871/2008-6 CNPq) Project 5S-VQ (Grant MCT/CNPq/-CT-INFO 55.1013/2005-2), CNPq Grant 305237/02-0 (Nivio Ziviani), CNPq Grant 302209/2007-7 (Edleno S. de Moura) and CNPq Grant 301043/2006-0 (Marcos André Gonçalves), Project SIRIAA (Grant MCT/CNPq/CT-Amazônia 55.3126/2005-9).

## References

ACM (1998) The acm computing classification system - 1998 version. http://www.acm.org/class/1998/ccs98.html

Almind TC, Ingwersen P (1997) Informetric analyses on the World Wide Web: Methodological approaches to "webometrics". Journal of Documentation 53(4):4004–426

Amsler R (1972) Application of citation-based automatic classification. Tech. rep., The University of Texas at Austin, Linguistics Research Center

Angelova R, Weikum G (2006) Graph-based text classification: learn from your neighbors. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 485–492

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

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. In: Proceedings of the 7th International World Wide Web Conference (WWW98), pp 107–117

Calado P, Cristo M, Moura E, Ziviani N, Ribeiro-Neto B, Gonçalves MA (2003) Combining link-based and content-based methods for web document classification. In: Proceedings of the 12th International Conference on Information and Knowledge Management, New Orleans, LA, USA, pp 394–401

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

Chakrabarti S, Dom B, Indyk P (1998) Enhanced hypertext categorization using hyperlinks. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp 307–318

Chang C, Lin CJ (2001) Libsvm: a library for support vector machines

Cochran WG (1977) Sampling Techniques, 2nd edn. John Wiley & Sons

Cohn D, Hofmann T (2001) The missing link - a probabilistic model of document content and hypertext connectivity. In: Leen TK, Dietterich TG, Tresp V (eds) Advances in Neural Information Processing Systems 13, MIT Press, pp 430–436

Couto T, Cristo M, Gonçalves MA, Calado P, Ziviani N, Moura E, Ribeiro-Neto B (2006) A comparative study of citations and links in document classification. In: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital libraries, pp 75–84

Cristo M, Calado P, Moura E, Nivio Ziviani BRN (2003) Link information as a similarity measure in web classification. In: 10th Symposium On String Processing and Information Retrieval SPIRE 2003, Lecture Notes in Computer Science, vol 2857, pp 43–55

Dean J, Henzinger MR (1999) Finding related pages in the World Wide Web. Computer Networks 31(11–16):1467–1479, also in Proceedings of the 8th International World Wide Web Conference (WWW99)

Egghe L, Rousseau R (1990) Introduction to informetrics: quantitative methods in library, documentation and information science. Elsevier Science Publishers, North-Holland, Amsterdam, The Netherlands

Fisher M, Everson R (2003) When are links useful? Experiments in text classification. In: Proceedings of the 25th European Conference on Information Retrieval Research, pp 41–56

Furnkranz J (1999) Exploiting structural information for text classification on the WWW. In: Proceedings of the 3rd Symposium on Intelligent Data Analysis (IDA99), pp 487–498

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

Glover EJ, Tsioutsiouliklis K, Lawrence S, Pennock DM, Flake GW (2002) Using Web structure for classifying and describing Web pages. In: Proceedings of the 11th International World Wide Web Conference (WWW02)

Gövert N, Lalmas M, Fuhr N (1999) A probabilistic description-oriented approach for categorizing web documents. In: Proceedings of the 8th International Conference on Information and Knowledge Management, Kansas City, MO, USA, pp 475–482

Hawking D, Craswell N (2001) Overview of TREC-2001 Web track. In: The Tenth Text REtrieval Conference (TREC-2001), Gaithersburg, MD, USA, pp 61–67

Joachims T (1998) Text categorization with support vector machines: learning with many relevant features. In: Proceedings of ECML-98, 10th European Conference on Machine Learning, Chemnitz, Germany, pp 137–142

Joachims T, Cristianini N, Shawe-Taylor J (2001) Composite kernels for hypertext categorisation. In: Proceedings of the 18th International Conference on Machine Learning, ICML-01, pp 250–257

Kessler MM (1963) Bibliographic coupling between scientific papers. American Documentation 14(1):10–25

Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. Journal of the ACM 46(5):604–632

Kumar R, Raghavan P, Rajagopalan S, Tomkins A (1999) Trawling the Web for emerging cyber-communities. Computer Networks 31(11–16):1481–1493, also in Proceedings of the 8th International World Wide Web Conference (WWW99)

Larson RR (1996) Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace. In: Annual Meeting of the American Society for Information Science, Baltimore, MD, USA, pp 71–78

Lawrence S, Giles CL, Bollacker KD (1999) Autonomous citation matching. In: Etzioni O, Müller JP, Bradshaw JM (eds) Proceedings of the Third Annual Conference on Autonomous Agents (AGENTS-99), ACM Press, pp 392–393

Li X, Chen H, Zhang Z, Li J (2007) Automatic patent classification using citation network information: an experimental study in nanotechnology. In: Proceedings of the ACM IEEE Joint Conference on Digital Libraries, pp 419–427

Marshakova IV (1973) A system of document connection based on refernces. Scientific and Technical Information Serial of VINITI 6(2):3–8

Mitchell T (1997) Machine Learning. McGraw-Hill

Moed HF (2005) Citation Analysis in Research Evaluation (Information Science & Knowledge Management). Springer-Verlag New York, Inc., Secaucus, NJ, USA

Oh HJ, Myaeng SH, Lee MH (2000) 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, pp 264–271

Qi X, Davison BD (2006) Knowing a web page by the company it keeps. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, pp

228–237

Qin J (2000) Semantic similarities between a keyword database and a controlled vocabulary database: An investigation in the antibiotic resistance literature. Journal of the American Society for Information Science 51(2):166–180

Saerens M, Latinne P, Decaestecker C (2002) Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. Neural Comput 14(1):21–41, DOI http://dx.doi.org/10.1162/089976602753284446

Salton G (1963) Associative document retrieval techniques using bibliographic information. Journal of the ACM 10(4):440–457

Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5):513–523

Sebastiani F (2002) Machine learning in automated text categorization. ACM Computing Surveys 34(1):1–47

Shen D, Sun JT, Yang Q, Chen Z (2006) A comparison of implicit and explicit links for web page classification. In: Proceedings of the 15th international conference on World Wide Web, New York, NY, USA, pp 643–650

Slattery S, Mitchell T (2000) Discovering test set regularities in relational domains. In: Proceedings of the 17th International Conference on Machine Learning, Stanford, CA, USA, pp 895–902

Small HG (1973) 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

Smith AG (2004) Web links as analogues of citations. Information Research 9(4)

Sun A, Lim EP, Ng WK (2002) Web classification using support vector machine. In: Proceedings of the Fourth International Workshop on Web Information and Data Management, pp 96–99

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

Turtle H, Croft WB (1991) Evaluation of an inference network-based retrieval model. ACM Transactions on Information Systems 9(3):187–222

Veloso A, Wagner Meira J, Cristo M, Gonçalves M, Zaki M (2006) Multi-evidence, multi-criteria, lazy associative document classification. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, pp 218–227

Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics Buletin 1(6):80–83

Witten IH, Frank E (2005) Data Mining, Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann Publishers, San Francisco, CA, USA

Yang Y (1994) 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, pp 13–22

Yang Y, Liu X (1999) A re-examination of text categorization methods. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, pp 42–49

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

Zhang B, Chen Y, Fan W, Fox EA, Goncalves M, Cristo M, Calado P (2005) Intelligent GP fusion from multiple sources for text classification. In: Proceedings of the 14th ACM international Conference on Information and Knowledge Management, ACM Press, Bremen, Germany