Web Page Classification: Features and Algorithms*

Xiaoguang Qi and Brian D. Davison
Department of Computer Science & Engineering
Lehigh University

June 2007

Abstract

Classification of web page content is essential to many tasks in web information retrieval such as maintaining web directories and focused crawling. The uncontrolled nature of web content presents additional challenges to web page classification as compared to traditional text classification, but the interconnected nature of hypertext also provides features that can assist the process.

As we review work in web page classification, we note the importance of these web-specific features and algorithms, describe state-of-the-art practices, and track the underlying assumptions behind the use of information from neighboring pages.

1 Introduction

Classification plays a vital role in many information management and retrieval tasks. On the Web, classification of page content is essential to focused crawling, to the assisted development of web directories, to topic-specific web link analysis, and to analysis of the topical structure of the Web. Web page classification can also help improve the quality of web search.

In this survey we examine the space of web classification approaches to find new areas for research, as well as to collect the latest practices to inform future classifier implementations. Surveys in web page classification typically lack a detailed discussion of the utilization of web-specific features. In this survey, we carefully review the web-specific features and algorithms that have been explored and found to be useful for web page classification. The contributions of this survey are:

- a detailed review of useful web-specific features for classification;
- an enumeration of the major applications for web classification; and,
- a discussion of future research directions.

The rest of this paper is organized as follows: the background of web classification and related work are introduced in Section 2; features and algorithms used in classification are reviewed in Section 3 and Section 4, respectively; we discuss several related issues in Section 5, point out some interesting direction in Section 6, and conclude in Section 6.

*Technical Report LU-CSE-07-010, Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA, 18015.

2 Background and related work

Before reviewing web classification research, we first introduce the problem, motivate it with applications, and consider related surveys in web classification.

2.1 Problem definition

Web page classification, also known as web page categorization, is the process of assigning a web page to one or more predefined category labels. Classification is often posed as a supervised learning problem (Mitchell 1997) in which a set of labeled data is used to train a classifier which can be applied to label future examples.

The general problem of web page classification can be divided into multiple sub-problems: subject classification, functional classification, sentiment classification, and other types of classification. Subject classification is concerned about the subject or topic of a web page. For example, judging whether a page is about "arts", "business" or "sports" is an instance of subject classification. Functional classification cares about the role that the web page plays. For example, deciding a page to be a "personal homepage", "course page" or "admission page" is an instance of functional classification. Sentiment classification focuses on the opinion that is presented in a web page, i.e., the author's attitude about some particular topic. Other types of classification include genre classification (e.g., (zu Eissen and Stein 2004)), search engine spam classification (e.g., (Gyöngyi and Garcia-Molina 2005b; Castillo, Donato, Gionis, Murdock, and Silvestri 2007)) and so on. This survey focuses on subject and functional classification.

Based on the number of classes in the problem, classification can be divided into binary classification and multi-class classification, where binary classification categorizes instances into exactly one of two classes (as in Figure 1(a)); multi-class classification deals with more than two classes. Based on the number of classes that can be assigned to an instance, classification can be divided into single-label classification and multi-label classification. In single-label classification, one and only one class label is to be assigned to each instance, while in multi-label classification, more than one class can be assigned to an instance. If a problem is multi-class, say four-class classification, it means four classes are involved, say Arts, Business, Computers, and Sports. It can be either single-label, where exactly one class label can be assigned to an instance (as in Figure 1(b)), or multi-label, where an instance can belong to any one, two, or all of the classes (as in Figure 1(c)). Based on the type of class assignment, classification can be divided into hard classification and soft classification. In hard classification, an instance can either be or not be in a particular class, without an intermediate state; while in soft classification, an instance can be predicted to be in some class with some likelihood (often a probability distribution across all classes, as in Figure 1(d)).

Based on the organization of categories, web page classification can also be divided into flat classification and hierarchical classification. In flat classification, categories are considered parallel, i.e., one category does not supersede another. While in hierarchical classification, the categories are organized in a hierarchical tree-like structure, in which each category may have a number of subcategories. An illustration is shown in Figure 2. Section 4 will address the issue of hierarchical classification further.

2.2 Applications of web classification

As briefly introduced in Section 1, classification of web content is essential to many information retrieval tasks. Here, we present a number of such tasks.

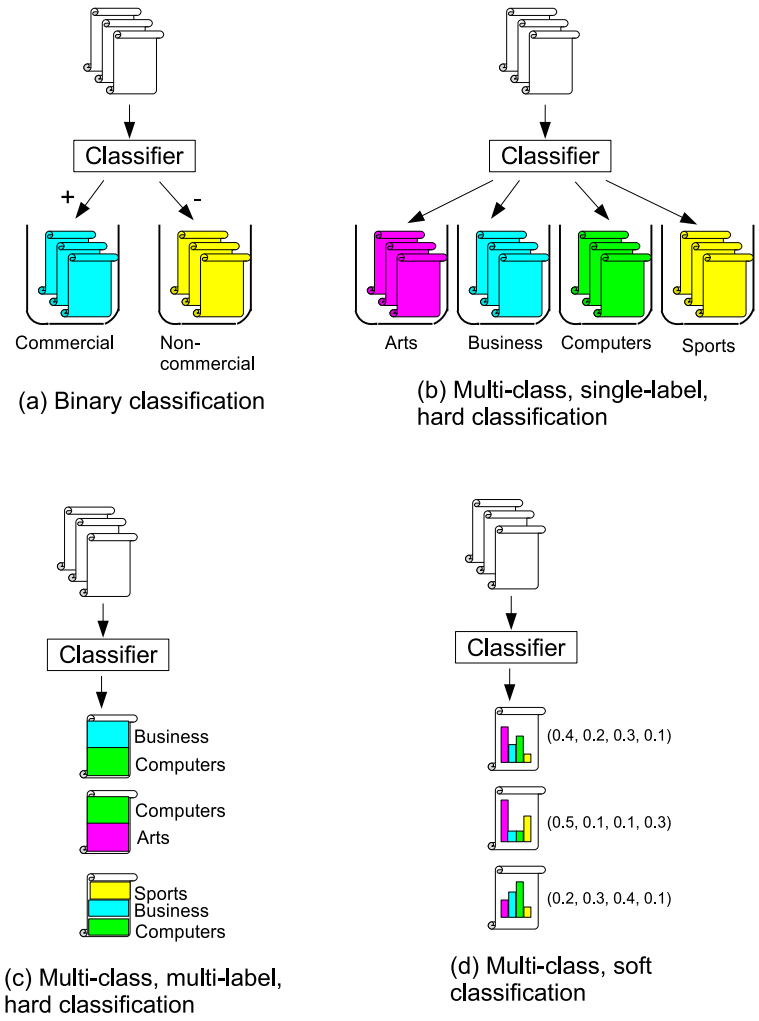

Figure 1: Types of classification

2.2.1 Constructing, maintaining or expanding web directories (web hierarchies)

Web directories, such as those provided by Yahoo! (2007) and the dmoz Open Directory Project (ODP) (2007), provide an efficient way to browse for information within a predefined set of categories. Currently, these directories are mainly constructed and maintained by editors, requiring extensive human effort. As of July 2006, it was reported (Corporation 2007) that there are 73,354 editors involved in the dmoz ODP. As the Web changes and continues to grow, this manual approach will become less effective. One could easily imagine building classifiers to help update and expand such directories. For example, Huang et al. (2004a, 2004b) propose an approach to automatic creation of classifiers from web corpora based on user-defined hierarchies. Further more, with advanced classification techniques, customized (or even dynamic) views of web directories can be generated automatically. There appears to be room for further interesting work along this direction.

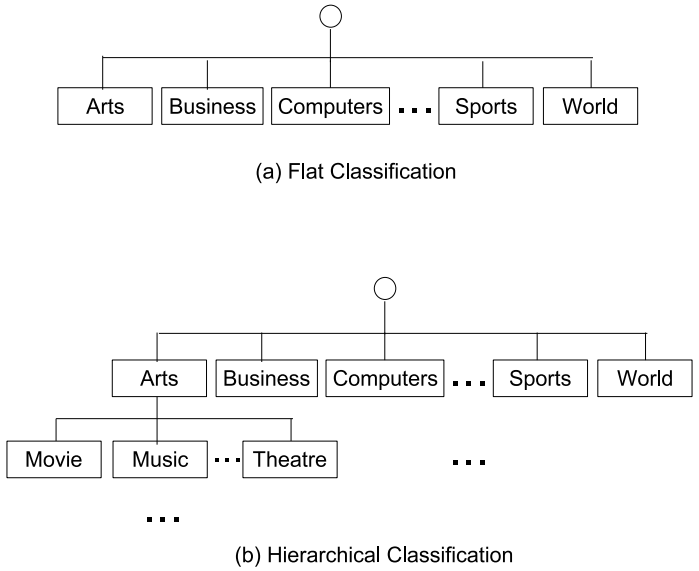

Figure 2: Flat classification and hierarchical classification.

2.2.2 Improving quality of search results

Query ambiguity is among the problems that undermine the quality of search results. For example, the query term "bank" could mean the border of a water area or a financial establishment. Various approaches have been proposed to improve retrieval quality by disambiguating query terms. Chekuri et al. (Chekuri et al. 1997) studied automatic web page classification in order to increase the precision of web search. A statistical classifier, trained on existing web directories, is applied to new web pages and produces an ordered list of categories in which the web page could be placed. At query time the user is asked to specify one or more desired categories so that only the results in those categories are returned, or the search engine returns a list of categories under which the pages would fall. This approach works when the user is looking for a known item. In such a case, it is not difficult to specify the preferred categories. However, there are situations in which the user is less certain about what documents will match, for which the above approach does not help much.

Search results are usually presented in a ranked list. However, presenting categorized, or clustered, results could be more useful to users. An approach proposed by Chen and Dumais (2000) classifies search results into a predefined hierarchical structure and presents the categorized view of the results to the user. Their user study demonstrated that the category interface is liked by the users better than the result list interface, and is more effective for users to find the desired information. Compared to the approach suggested by Chekuri et al., this approach is less efficient at query time because it categorizes web pages on-the-fly. However, it does not require the user to specify desired categories; therefore, it is more helpful when the user does not know the query terms well. Similarly, Käki (2005) also proposed to present a categorized view of search results to users. Experiments showed that the categorized view is beneficial for the users, especially when the ranking of results is not satisfying.

In 1998, Page and Brin developed the link-based ranking algorithm called PageRank (1998). PageRank calculates the authoritativeness of web pages based on a graph constructed by web pages and their hyperlinks, without considering the topic of each page. Since then, much research has been explored to differentiate authorities of different topics. Haveliwala (2002) proposed Topic-sensitive PageRank, which performs multiple PageRank calculations, one for

each topic. When computing the PageRank score for each category, the random surfer jumps to a page in that category at random rather than just any web page. This has the effect of biasing the PageRank to that topic. This approach needs a set of pages that are accurately classified. Nie et al. (2006) proposed another web ranking algorithm that considers the topics of web pages. In that work, the contribution that each category has to the authority of a web pages is distinguished by means of soft classification, in which a probability distribution is given for a web page being in each category. In order to answer the question "to what granularity of topic the computation of biased page ranks make sense," Kohlschutter et al. (2007) conducted analysis on ODP categories, and showed that ranking performance increases with the ODP level up to a certain point. It seems further research along this direction is quite promising.

2.2.3 Helping question answering systems

A question answering system may use classification techniques to improve its quality of answers. Yang and Chua (Yang and Chua 2004a; Yang and Chua 2004b) suggested finding answers to list questions (where a set of distinct entities are expected, e.g., "name all the countries in Europe") through web page functional classification. Given a list question, a number of queries are formulated and sent to search engines. The web pages in the results are retrieved and then classified by decision tree classifiers into one of the four categories: collection pages (containing a list of items), topic pages (representing an answer instance), relevant pages (supporting an answer instance), and irrelevant pages. In order to increase coverage, more topic pages are included by following the outgoing links of the collection pages. After that, topic pages are clustered, from which answers are extracted.

There have additionally been a number of approaches to improving quality of answers by means of question classification (Harabagiu, Pasca, and Maiorano 2000; Hermjakob 2001; Kwok, Etzioni, and Weld 2001; Zhang and Lee 2003) which are beyond the scope of this survey.

One interesting question that previous publications do not answer is how useful web page subject classification is in question answering systems. In Section 2.2.2, we reviewed a number of approaches that use the topical information of web pages to improve the performance of web search. Similarly, by determining the category of expected answers of a question and classifying the web pages that may contain candidate answers, a question answering system could benefit in terms of both accuracy and efficiency.

2.2.4 Building efficient focused crawlers or vertical (domain-specific) search engines

When only domain-specific queries are expected, performing a full crawl is usually inefficient. Chakrabarti et al. (Chakrabarti et al. 1999) proposed an approach called focused crawling, in which only documents relevant to a predefined set of topics are of interest. In this approach, a classifier is used to evaluate the relevance of a web page to the given topics so as to provide evidence for the crawl boundary.

2.2.5 Other applications

Besides the applications discussed above, web page classification is also useful in web content filtering (Hammami, Chahir, and Chen 2003; Chen, Wu, Zhu, and Hu 2006), assisted web browsing (Armstrong, Freitag, Joachims, and Mitchell 1995; Pazzani, Muramatsu, and Billsus 1996; Joachims, Freitag, and Mitchell 1997) and in knowledge base construction (Craven, DiPasquo, Freitag, McCallum, Mitchell, Nigam, and Slattery 1998).

2.3 The difference between web classification and text classification

The more general problem of text classification (Sebastiani 1999; Aas and Eikvil 1999; Tan 1999; Tong and Koller 2001; Sebastiani 2002; Cardoso-Cachopo and Oliveira 2003; Bennett, Dumais, and Horvitz 2005) is beyond the scope of this article. Compared with standard text classification, classification of web content is different in the following aspects. First, traditional text classification is typically performed on "structured corpora with well-controlled authoring styles" (Chekuri, Goldwasser, Raghavan, and Upfal 1997), while web collections do not have such a property. Second, web pages are semi-structured documents in HTML, so that they may be rendered visually for users. Although other document collections may have embedded information for rendering and/or a semi-structured format, such markup is typically stripped for classification purposes. Finally, web documents exist within a hypertext, with connections to and from other documents. While not unique to the web (consider for example the network of scholarly citations), this feature is central to the definition of the web, and is not present in typical text classification problems. Therefore, web classification is not only important, but distinguished from traditional text classification, and thus deserving of the focused review found in this article.

2.4 Related surveys

Although there are surveys on textual classification that mention web content, they lack an analysis of features specific to the web. Sebastiani (2002) mainly focused on traditional textual classification. Chakrabarti (2000) and Kosala and Blockeel (2000) reviewed web mining research in general as opposed to concentrating on classification. Mladenic (1999) reviewed a number of text-learning intelligent agents, some of which are web-specific. However, her focus was on document representation and feature selection. Getoor and Diehl (2005) reviewed data mining techniques which explicitly consider links among objects, with web classification being one of such areas. Fürnkranz (2005) reviews various aspects of web mining, including a brief discussion on the use of link structure to improve web classification. Closer to the present article is the work by Choi and Yao (2005) which described the state of the art techniques and subsystems used to build automatic web page classification systems.

This survey updates and expands on prior work by considering web-specific features and algorithms in web page classification.

3 Features

In this section, we review the types of features found to be useful in web page classification research.

Written in HTML, web pages contain additional information, such as HTML tags, hyperlinks and anchor text (the text to be clicked on to activate and follow a hyperlink to another web page, placed between HTML <A> and </A> tags), other than the textual content visible in a web browser. These features can be divided into two broad classes: on-page features, which are directly located on the page to be classified, and features of neighbors, which are found on the pages related in some way with the page to be classified.

3.1 Using on-page features

3.1.1 Textual content and tags

Directly located on the page, the textual content is the most straightforward feature that one may consider to use. However, due to the variety of uncontrolled noise in web pages, directly using a bag-of-words representation for all terms may not achieve top performance. Researchers

have tried various methods to make better use of the textual features. One popular method is feature selection, which we cover in Section 4. N-gram representation is another method that is found to be useful. Mladenic (1998) suggested an approach to automatic web page classification based on the Yahoo! hierarchy. In this approach, each document is represented by a vector of features, which includes not only single terms, but also up to 5 consecutive words. The advantage of using n-gram representation is that it is able to capture the concepts expressed by a sequence of terms (phrases), which are unlikely to be characterized using single terms. Imagine a scenario of two different documents. One document contains the phrase "New York". The other contains the terms "new" and "york", but the two terms appear far apart. A standard bag-of-words representation cannot distinguish them, while a 2-gram representation can. However, an n-gram approach has a significant drawback; it usually generates a space with much higher dimensionality than the bag-of-words representation does. Therefore, it is usually performed in combination with feature selection.

One obvious feature that appears in HTML documents but not in plain text documents is HTML tags. It has been demonstrated that using information derived from tags can boost the classifier's performance. Golub and Ardo (2005) derived significance indicators for textual content in different tags. In their work, four elements from the web page are used: title, headings, metadata, and main text. They showed that the best result is achieved from a well-tuned linear combination of the four elements. This approach only distinguished the four types of elements while mixing the significance of other tags. Kwon and Lee (2000, 2003) proposed classifying web pages using a modified k-Nearest Neighbor algorithm, in which terms within different tags are given different weights. They divided all the HTML tags into three groups and assigned each group an arbitrary weight.

Thus, utilizing tags can take advantage of the structural information embedded in the HTML files, which is usually ignored by plain text approaches. However, since most HTML tags are oriented toward representation rather than semantics, web page authors may generate different but conceptually equivalent tag structures. Therefore, using HTML tagging information in web classification may suffer from the inconsistent formation of HTML documents.

Good quality document summarization can accurately represent the major topic of a web page. Shen et al. (2004) proposed an approach to classifying web pages through summarization. They showed that classifying web pages on their summaries is able to improve the accuracy by around 10% as compared with content based classifiers.

Rather than deriving information from the page content, Kan and Thi (Kan 2004; Kan and Thi 2005) demonstrated that a web page can be classified based on its URL. While not of ideal accuracy, this approach eliminates the necessity of downloading the page. Therefore, it is especially useful when the page content is not available or time/space efficiency is strictly emphasized.

3.1.2 Visual analysis

Each web page has two representations, if not more. One is the text representation written in HTML. The other one is the visual representation rendered by a web browser. They provide different views of a page. Most approaches focus on the text representation while ignoring the visual information. Yet the visual representation is useful as well.

A web page classification approach based on visual analysis was proposed by Kovacevic et al. (2004), in which each web page is represented as a hierarchical "visual adjacency multigraph." In the graph, each node represents an HTML object and each edge represents the spatial relation in the visual representation. Based on the result of visual analysis, heuristic rules are applied to recognize multiple logical areas, which correspond to different meaningful parts of the page. They compared the approach to a standard bag-of-words approach and demonstrated great improvement. In a complementary fashion, a number of visual features, as well as textual features, were used in the web page classification work by Asirvatham and Ravi (2001). Based

on their observation that research pages contain more synthetic images, the histogram of the images on the page is used to differentiate between natural images and synthetic images to help classification of research pages.

Although the visual layout of a page relies on the tags, using visual information of the rendered page is arguably more generic than analyzing document structure focusing on HTML tags (Kovacevic, Diligenti, Gori, and Milutinovic 2004). The reason is that different tagging may have the same rendering effect. In other words, sometimes one can change the tags without affecting the visual representation. Based on the assumption that most web pages are built for human eyes, it makes more sense to use visual information rather than intrinsic tags.

On-page features are useful but they provide information only from the viewpoint of the page creator. Sometimes it is necessary to use features that do not reside on the page. We discuss this issue in the following subsection.

3.2 Using features of neighbors

3.2.1 Motivation

Although web pages contain useful features as discussed above, in a particular web page these features are sometimes missing, misleading, or unrecognizable for various reasons. For example, some web pages contain large images or flash objects but little textual content, such as in the example shown in Figure 3. In such cases, it is difficult for classifiers to make reasonable judgments based on features on the page.

In order to address this problem, features can be extracted from neighboring pages that are related in some way to the page to be classified to supply supplementary information for categorization. There are a variety of ways to derive such connections among pages. One obvious connection is the hyperlink. Since most existing work that utilizes features of neighbors is based on hyperlink connection, in the following, we focus on hyperlinks connection. However, other types of connections can also be derived; and some of them have been shown to be useful for web page classification. These types of connections are discussed in Section 3.2.5.

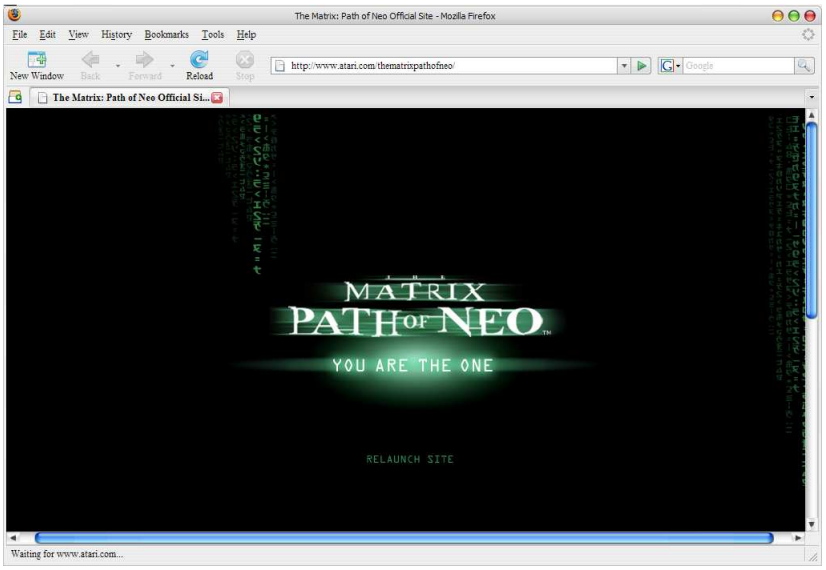

Figure 3: An example web page which has few useful on-page features.

3.2.2 Underlying assumptions

When exploring the features of neighbors, some assumptions are implicitly made in existing work. Usually, it is assumed that if pages p_a and p_b belong to the same category, pages neighboring them in the web graph share some common characteristics. This assumption does not require that the neighboring pages belong to the same category as p_a and p_b do. In the following, we refer to this thesis as the *weak assumption*. The weak assumption works in both subject classification and functional classification. Under the weak assumption, a classifier can be derived from the features of the neighboring pages of training examples, and used to predict the categories of testing examples based on the features of their neighbors.

In subject classification, a stronger assumption is often made—that a page is much more likely to be surrounded by pages of the same category. In other words, the presence of many "sports" pages in the neighborhood of p_a increases the probability of p_a being in "sports". We term this the *strong assumption*. The strong assumption requires a strong correlation between links and topics of web pages. Davison (Davison 2000) showed that linked pages were more likely to have terms in common. Chakrabarti et al. (2002) studied the topical structure of the Web and showed that pages tend to link to pages on the same topic. Similarly, Menczer (2005) also showed a strong correlation between links and content of web pages. The strong assumption has been shown to work well in subject classification on broad (i.e., high-level) categories. However, evidence for its validity on fine-grained categories is lacking. Furthermore, it seems unlikely that the strong assumption works in function classification. Under the strong assumption, one might build statistical classifiers to predict the category of the page in question simply by taking the majority class of its neighboring pages.

3.2.3 Neighbor selection

Another question when using features from neighbors is that of which neighbors to examine. Existing research mainly focuses on pages within two steps of the page to be classified. At a distance no greater than two, there are six types of neighboring pages according to their hyperlink relationship with the page in question: parent, child, sibling, spouse, grandparent and grandchild, as illustrated in Figure 4. The effect and contribution of the first four types of neighbors have been studied in existing research. Although grandparent pages and grandchild pages have also been used, their individual contributions have not yet been specifically studied. In the following, we group the research in this direction according to the neighbors that are used.

In general, directly incorporating text from parent and child pages into the target page does more harm than good because parent and child pages are likely to have different topics than the target page (Chakrabarti, Dom, and Indyk 1998; Ghani, Slattery, and Yang 2001; Yang, Slattery, and Ghani 2002). This, however, does not mean that parent and child pages are useless. The noise from neighbors can be greatly reduced by at least two means: using an appropriate *subset* of neighbors, and using an appropriate *portion* of the content on neighboring pages. Both methods have been shown to be helpful.

Using a subset of parent and child pages can reduce the influence from pages on different topics than the target page. For example, while utilizing parent and child pages, Oh et al. (2000) require the content of neighbors to be sufficiently similar to the target page. Using a portion of content on parent and child pages, especially the content close enough to the hyperlink that points to the target page, can reduce the influence from the irrelevant part of neighboring pages. Usually, title, anchor text, and the surrounding text of anchor text on the parent pages are found to be useful. This family of approaches takes advantage from both hyperlinks and HTML structure information. Below, we review some existing approaches of this type.

Attardi et al. (1999) proposed to use the title, anchor text, and a portion of text surrounding the anchor text on parent pages to help determine the target page's topic, and showed promising

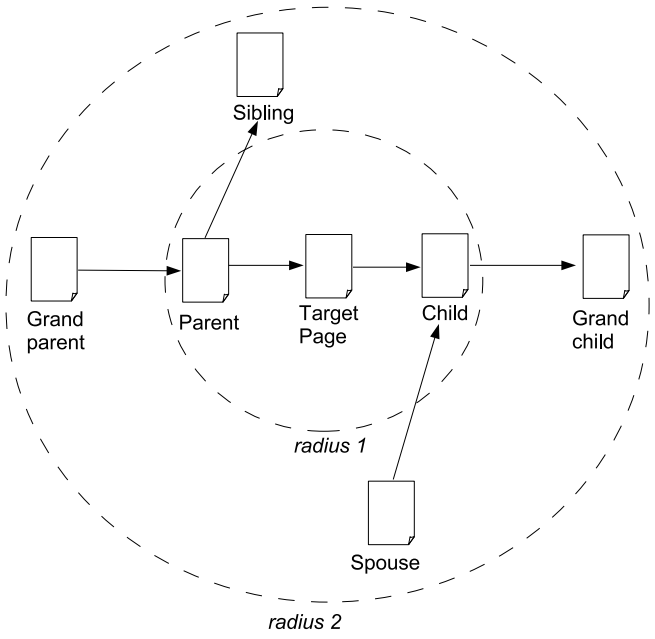

Figure 4: Neighbors within radius of two.

result. Fürnkranz (1999) used features on parent pages like anchor text, the neighborhood of anchor text, and the headings that precede the link, and showed improvement over the classifier that uses text on the target page alone. In later work (2001), an interesting approach was proposed by Fürnkranz in which text on the parent pages surrounding the link is used to train a classifier instead of text on the target page. As a result, a target page will be assigned multiple labels by such a classifier, one for each incoming link. These labels are then combined by some voting scheme to form the final prediction of the target page's class. Yang et al. (2002) considered various approaches to hypertext classification. Their results are mixed, finding that identification of hypertext regularities and appropriate representations are crucial to categorization performance. They note, however, that "algorithms focusing on automated discovery of the relevant parts of the hypertext neighborhood should have an edge over more naive approaches." Sun et al. (2002) showed that SVM classifiers using the text on the target page, page title (as separate features), and anchor text from parent pages can improve classification compared with a pure text classifier. Similarly, Glover et al. (2002) demonstrated that utilizing extended anchortext (the surrounding text of anchortext, including the anchor text itself) from parent pages can improve the accuracy compared with the classifier which uses on-page content only. Utard and Fürnkranz (2005) also proposed to use a portion of text as opposed to full text on parent pages. They studied individual features from parent pages such as anchor text, the neighborhood of anchor text, the paragraph containing the link, the headings before the anchor text, as well as the text on the target page, and showed significant improvement of pair-wise combination of such features over the individual features. Besides directly using a portion of the text on parent pages, implicitly utilizing information from parent pages could also be successful, such as applying wrapper learners to anchor text on parent pages proposed by Cohen (2002).

Sibling pages are even more useful than parents and children. This was empirically demonstrated by Chakrabarti et al. (1998) and again by Qi and Davison (Qi and Davison 2006). Such sibling relationships can also help in relational learning of functional categories (Slattery and Mitchell 2000).

Using multiple types of neighbor could provide additional benefit. Calado et al. (2003) studied the use of several link similarity measures in web page topical classification, in which the link similarities are derived from hyperlinked neighbors within two steps. Appropriate combinations of such link similarities and textual classifiers can make great improvement over textual classifiers. Qi and Davison (2006) proposed a method to enhance web page classification by utilizing the class and content information from neighboring pages in the link graph. The categories represented by four kinds of neighbors (parents, children, siblings and spouses) are combined to help with the page in question. That study of the contribution of the four types of neighbors revealed that while sibling pages are the most important type of neighbor to use, the other types are also of value.

Instead of individually considering each target page together with its neighbors, some algorithms may collectively consider the class labels of all the nodes within a graph. This type of approach is discussed in Section 4.2.

The idea of utilizing information from neighbors can also be applied to unsupervised learning. Drost et al. (2005) proposed to find communities in linked data by using similarity metrics based on co-citation and bibliographic coupling relationships, as well as content similarity. Angelova and Siersdorfer (2006) proposed an approach to linked document clustering by means of iterative relaxation of cluster assignments on a linked graph.

In summary, on one hand, although parent, child, sibling, and spouse pages are all useful in classification, siblings are found to be the best source; on the other hand, since using information from neighboring pages may introduce extra noise, they should be used carefully. The effect of grandparent and grandchild pages has not been well studied.

As mentioned earlier, little work has examined the effect of pages beyond two steps away. There are at least two reasons for this: first, due to the explosive number of neighbors, utilizing features of neighbors at a long distance is expensive; second, the farther away the neighbors are, the less likely they have topics in common with the page being classified (Chakrabarti, Joshi, Punera, and Pennock 2002), and thus they are less useful in classification.

3.2.4 Features of neighbors

The features that have been used from neighbors include labels, partial content (anchor text, the surrounding text of anchor text, titles, headers), and full content.

Some researchers take advantage of neighboring pages which have already been labeled by humans. For example, Chakrabarti et al. (1998), Slattery and Mitchell (2000), and Calado et al. (2003) used the labels of neighbors in their work. The advantage of directly using labels is that human labeling is more accurate than classifiers. The disadvantage is that these labels are not always available. (Human-labeled pages, of course, are available on only a very small portion of the Web.) When the labels are not available, these approaches would either suffer significantly in terms of coverage (leaving a number of pages undecidable) or reduce to the result of traditional content-based classifiers.

As discussed earlier in Section 3.2.3, using partial content of neighbors can reduce the effect of irrelevant topics on neighboring pages. As one kind of partial content of parents, anchor text is usually considered a tag or a concise description of the target page (Amitay 1998; Davison 2000). It is useful in web classification as well as in web search. However, given that anchor text is usually short, it may not contain enough information for classification. As shown by Glover et al. (2002), "anchortext alone is not significantly better (arguably worse) than using the full-text alone". Using surrounding text of anchor text (including the anchor text itself), as in (Glover, Tsioutsiouliklis, Lawrence, Pennock, and Flake 2002), or using anchor text indirectly, as in (Cohen 2002), can address this problem. For carefully created pages, information in titles and headers are generally more important than the prose. Therefore, it is reasonable to use titles and anchors instead of the full content of neighboring pages, or to emphasize them when using full content. Compared with using labels of neighbors, using partial content of neighboring

Approach	Assumption based upon	Types of neighbors used	On-page features utilized?	Types of features used	Combination method
Chakrabarti et al. (1998)	weak	sibling	no	label	N/A
Attardi et al. (1999)	weak	parent	no	text	N/A
Fürnkranz (1999)	weak	parent	no	anchor text, extended anchor text, & headings	multiple voting schemes
Slattery & Mitchell (2000)	weak	sibling	no	label	N/A
Fürnkranz (2001)	weak	parent	no	anchor text, extended anchor text, & headings	multiple voting schemes
Glover et al. (2002)	weak	parent	yes	extended anchor text	N/A
Sun et al. (2002)	weak	parent	yes	anchor text, plus text & title of target page	
Cohen (2002)	weak	parent	no	text & anchor	N/A
Calado et al. (2003)	strong	all six types types within two steps	yes	label	Bayesian Network
Angelova & Weikum (2006)	weak	"reliable" neighbors in a local graph	yes	text & label	N/A
Qi & Davison (2006)	strong	parent, child, sibling, spouse	yes	text & label	weighted average

Table 1: Comparison of approaches using features of neighbors.

pages does not rely on the presence of human labeled pages in the neighborhood. The benefit of using such partial content, however, partially relies on the quality of the linked pages.

Among the three types of features, using the full content of neighboring pages is the most expensive; however it may generate better accuracy. Oh et al. (2000) showed that using the class of neighbors provides a 12% gain in F-measure over the approach which only considers on-page content. They also showed that including content of neighbors can increase F-measure by another 1.5%. Qi and Davison (2006) demonstrated that additionally using the topical labeling of neighboring page content outperforms two other approaches which only use the labels of neighbors.

The approaches that utilize features of neighbors are compared in Table 1. From the table, we can see that class label is a frequently-used feature. Interestingly, anchortext is less frequently used than one would expect given its apparent descriptive power.

3.2.5 Utilizing artificial links

Although hyperlinks are the most straightforward type of connection between web pages, it is not the only choice. One might also ask which pages should be connected/linked (even if not linked presently). While simple textual similarity might be a reasonable start, a stronger measure is to consider pages that co-occur in top query results (Fitzpatrick and Dent 1997; Beeferman and Berger 2000; Glance 2000; Wen, Nie, and Zhang 2002; Zaiane and Strilets 2002; Davison 2004). In this model, two pages are judged to be similar by a search engine in a particular context, and would generally include pages that contain similar text and similar importance (so that they both rank high in a query). Based on the idea of utilizing information in queries and results, Shen et al. (2006) suggested an approach to creating connections between pages that appear in the results of the same query and are both clicked by users, which they term "implicit links". Thus, they utilize similarity as formed by the ranking algorithm, but also by human insight. Their comparison between implicit links and explicit links (hyperlinks) showed that implicit links can help web page classification. A similar approach which classifies web pages by utilizing the interrelationships between web pages and queries was proposed by Xue et al. (2006). The main idea is to iteratively propagate the category information of one type of object (pages or queries) to related objects. This approach showed an improvement of 26% in F-measure over content-based web page classification. In addition to web page classification, artificial connections built upon query results and query logs can be used for query classification.

Links derived from textual similarities can also be useful. Based on a set of feature vectors generated from web directories, Gabrilovich and Markovitch (2005) proposed to use feature vectors that are similar enough to the content of the target document to help classification, although such links are not explicitly generated. Given a web directory such as dmoz or Yahoo!, a feature vector is generated for each node (category) by selecting highly representative terms in the category description, URL, and the web sites within that category. After that, the feature generator compares the input text with the feature vectors of all the directory nodes, and vectors that are similar enough are chosen to enrich the bag-of-words representation of the input page's text. A similar approach that utilizes encyclopedic knowledge for automatic feature generation is described in (Gabrilovich and Markovitch 2006).

There are other methods to generate such artificial links but have not been tested with respect to web classification, such as the "generation links" proposed by Kurland and Lee (2005, 2006), in which links are created between documents if the language model induced from one document assigns high probability to another. Another example is the links proposed by Luxenburger and Weikum (2004), which are generated through high order links within query logs and content similarity.

There are opportunities for future research in this direction. For example, co-click-through data is generally quite sparse. An interesting question is how the sparsity of such data affects its usefulness in web classification. In addition, it would be useful to see how well the combination of artificial links and hyperlinks works.

3.3 Discussion: features

On-page features directly reside on the page to be classified. Methods to extract on-page features are fairly well-developed and relatively inexpensive to extract. While obtaining features of neighbors is computationally more expensive (particularly for researchers not within search engine companies), these features provide additional information that cannot be obtained otherwise. When designing a classifier, a decision needs to be made regarding the the trade-off between accuracy and efficiency. Yang et al. (2003) compared the computational complexity of several popular text categorization algorithms by means of formal mathematical analysis and experiments. However, most work on web specific classification lacks an analysis of computational complexity, which makes an implementer's decision more difficult.

A comparison of many of the approaches reviewed in this section across several characteristics is shown in Table 2. In order to provide a rough idea of performance for each approach, the baseline classifier with which the approach is compared and its reported improvement over the baseline is listed in the table. However, since the results of different approaches are based on different implementations and different datasets, the performance comparison provided here should only serve as a start of a comprehensive evaluation. From Table 2, we can see that

- web classification techniques achieve promising performance, although there appears to be room for improvement;
- bag-of-words and set-of-words are popular document representations; and
- existing approaches are evaluated on a wide variety of metrics and datasets, making it difficult to compare their performance.

Although the benefit of utilizing features of neighbors has been shown in many papers, little work has been done to analyze the underlying reason why such features are useful. In general, features of neighbors provide an alternative view of a web page, which supplements the view from on-page features. Therefore, collectively considering both can help reduce classification error. Jensen et al. (2004) investigated the underlying mechanism of collective inference, and argued that the benefit does not only come from a larger feature space, but from modeling dependencies among neighbors and utilizing known class labels. Such explanations may also apply to why web page classification benefits from utilizing features of neighbors.

Sibling pages are even more useful than parents and children. We speculate that the reason may lie in the process of hyperlink creation. When linking to other pages, authors of web pages often tend to link to pages with related (but not the same) topics of the current page. As a result, this page, as a whole, may not be an accurate description of its outgoing links. The outgoing links, however, are usually on the same topic, especially those links adjacent to each other. In other words, a page often acts as a bridge to connect its outgoing links, which are likely to have common topics. Therefore, sibling pages are more useful in classification than parent and child pages.

4 Algorithms

The types of features used in web page classification have been reviewed in the previous section. In this section, we focus on the algorithmic approaches.

4.1 Dimension reduction

Besides deciding which types of features to use, the weighting of features also plays an important role in classification. Emphasizing features that have better discriminative power will usually boost classification. Feature selection can be seen as a special case of feature weighting, in which features that are eliminated are assigned zero weight. Feature selection reduces the dimensionality of the feature space, which leads to a reduction in computational complexity. Furthermore, in some cases, classification can be more accurate in the reduced space. A review of traditional feature selection techniques used in text classification can be found in (Yang and Pedersen 1997).

There are a variety of measures to select features. Some simple approaches have been proven effective. For example, Shanks and Williams (2001) showed that only using the first fragment of each document offers fast and accurate classification of news articles. This approach is based on an assumption that a summary is present at the beginning of each document, which is usually true for news articles, but does not always hold for other kinds of documents. However, this approach was later applied to hierarchical classification of web pages by Wibowo and Williams (2002a), and was shown to be useful for web documents.

Approach	Task	Baseline classifier	Reported improv.	Document represent.	Feature selection criteria	Evaluation dataset
Chakrabarti et al. (1998)	Topical	A term-based classifier built on TAPER	From 32% to 75% (accuracy)	Bag-of-words	A score-based function derived from Duda & Hart (1973)	Yahoo directory
Mladenic (1999)	Topical	N/A	N/A	N-gram	Gram frequency	Yahoo directory
Fürnkranz (1999)	Functional	A text classifier	From 70.7% to 86.6% (accuracy)	Set-of-words	Entropy (for baseline classifier)	WebKB
Slattery & Mitchell (2000)	Functional	N/A	N/A	Relations	No feature selection	WebKB
Kwon & Lee (2000)	Topical	A kNN classifier with traditional cosine similarity measure	From 18.2% to 19.2% (micro-averaging breakeven point)	Bag-of-words	Expected mutual information and mutual information	Hanmir
Fürnkranz (2001)	Functional	A text classifier	From 70.7% to 86.9% (accuracy)	Set-of-words	Entropy (for baseline classifier)	WebKB
Sun et al. (2002)	Functional	A text classifier	From 0.488 to 0.757 (F-measure)	Set-of-words	No feature selection	WebKB
Cohen (2002)	Topical	A simple bag-of-words classifier	From 91.6% to 96.4% (accuracy)	Bag-of-words	No feature selection	A custom crawl on nine company web sites
Calado et al. (2003)	Topical	A kNN textual classifier	From 39.5% to 81.6% (accuracy)	Bag-of-words	Information gain	Cade directory
Golub & Ardi (2005)	Topical	N/A	N/A	Word/phrase	No feature selection	Engineering Electric Library
Qi & Davison (2006)	Topical	An SVM textual classifier	From 73.1% to 91.4% (accuracy)	Bag-of-words	No feature selection	ODP directory

Table 2: Comparison of web page classification approaches.

Besides these simple measures, there have been a number of feature selection approaches developed in text categorization, such as information gain and mutual information. These approaches can also be useful for web classification. Kwon and Lee (2000) proposed an approach based on a variation of the k-Nearest Neighbor algorithm, in which features are selected using two

well-known metrics: expected mutual information and mutual information. They also weighted terms according to the HTML tags that the term appears in, i.e., terms within different tags bear different importance. Calado et al. (2003) used information gain, another well-known metric, to select the features to be used. However, they did not show to what extent feature selection and feature weighting contributed to the improvement. Furthermore, based on existing work, it is not clear which feature selection algorithms are superior for web classification.

In text categorization, there is a class of problem where the categories can be distinguished by a few number of features while a large number of other features only add little additional differentiation power. Gabrilovich and Markovitch (2004) studied such types of classification problems and showed that the performance of SVM classifiers can be improved in such problems by aggressive feature selection. They also developed a measure that is able to predict the effectiveness of feature selection without training and testing classifiers.

Latent Semantic Indexing (LSI) (Deerwester, Dumais, Landauer, Furnas, and Harshman 1990) is a popular dimension reduction approach, which reinterprets text documents in a smaller transformed, but less intuitive space. However, its high computational complexity makes it inefficient to scale. Therefore, research experiments utilizing LSI in web classification (e.g., (Zelikovitz and Hirsh 2001; Riboni 2002)) are based on small datasets. Some work has improved upon LSI (e.g., probabilistic (Hofmann 1999b; Hofmann 1999a)) and make it more applicable to large datasets. Research has demonstrated the effectiveness of such improvements (Cohn and Hofmann 2001; Fisher and Everson 2003). Their efficiency on large datasets, however, needs further study.

4.2 Relational learning

Since web pages can be considered as instances which are connected by hyperlink relations, web page classification can be solved as a relational learning problem, which is a popular research topic in machine learning. Therefore, it makes sense to apply relational learning algorithms to web page classification. Relaxation labeling is one of the algorithms that work well in web classification.

Relaxation labeling was originally proposed as a procedure in image analysis (Rosenfeld, Hummel, and Zucker 1976). Later, it became widely used in image and vision analysis, artificial intelligence, pattern recognition, and web mining. "In the context of hypertext classification, the relaxation labeling algorithm first uses a text classifier to assign class probabilities to each node (page). Then it considers each page in turn and reevaluates its class probabilities in light of the latest estimates of the class probabilities of its neighbors" (Chakrabarti 2003).

Relaxation labeling is effective in web page classification (Chakrabarti, Dom, and Indyk 1998; Lu and Getoor 2003; Angelova and Weikum 2006). Based on a new framework for modeling link distribution through link statistics, Lu and Getoor (2003) proposed a variation of relaxation labeling, in which a combined logistic classifier is used based on content and link information. This approach not only showed improvement over a textual classifier, but also outperformed a single flat classifier based on both content and link features. In another variation proposed by Angelova and Weikum (2006), not all neighbors are considered. Instead, only neighbors that are similar enough in content are used.

Besides relaxation labeling, other relational learning algorithms can also be applied to web classification. Sen and Getoor (2007) compared and analyzed relaxation labeling along with two other popular link-based classification algorithms: loopy belief propagation and iterative classification. Their performance on a web collection is better than textual classifiers. Macskassy and Provost (2007) implemented a toolkit for classifying networked data, which utilized a relational classifier and a collective inference procedure (Jensen, Neville, and Gallagher 2004), and demonstrated its powerful performance on several datasets including web collections.

4.3 Modifications to traditional algorithms

Besides feature selection and feature weighting, efforts have also been made to tweak traditional algorithms, such as k-Nearest Neighbor and Support Vector Machine (SVM), in the context of web classification.

k-Nearest Neighbor classifiers require a document dissimilarity measure to quantify the distance between a test document and each training document. Most existing kNN classifiers use cosine similarity or inner product. Based on the observation that such measures cannot take advantage of the association between terms, Kwon and Lee (2000, 2003) developed an improved similarity measure that takes into account the term co-occurrence in documents. The intuition is that frequently co-occurring terms constrain the semantic concept of each other. The more co-occurred terms two documents have in common, the stronger the relationship between the two documents. Their experiments showed performance improvements of the new similarity measure over cosine similarity and inner product measures. Gövert et al. (1999) reinterpreted k-Nearest Neighbor algorithm with probability computation. In this probabilistic kNN, the probability of a document d being in class c is determined by its distance between its neighbors and itself and its neighbors' probability of being in class c.

Most supervised learning approaches only learn from training examples. Co-training, introduced by Blum and Mitchell (1998), is an approach that makes use of both labeled and unlabeled data to achieve better accuracy. In a binary classification scenario, two classifiers that are trained on different sets of features are used to classify the unlabeled instances. The prediction of each classifier is used to train the other. Compared with the approach which only uses the labeled data, this co-training approach is able to cut the error rate by half. Ghani (2001, 2002) generalized this approach to multi-class problems. The results showed that co-training does not improve accuracy when there are a large number of categories. On the other hand, their proposed method which combines error-correcting output coding (a technique to improve multi-class classification performance by using more than enough classifiers, see (Dietterich and Bakiri 1995) for details) with co-training is able to boost performance. Park and Zhang (2003) also applied co-training in web page classification which considers both content and syntactic information.

Classification usually requires manually labeled positive and negative examples. Yu et al. (2004) devised an SVM-based approach to eliminate the need for manual collection of negative examples while still retaining similar classification accuracy. Given positive data and unlabeled data, their algorithm is able to identify the most important positive features. Using these positive features, it filters out possible positive examples from the unlabeled data, which leaves only negative examples. An SVM classifier could then be trained on the labeled positive examples and the filtered negative examples.

4.4 Hierarchical classification

Most existing web classification approaches focus on classifying instances into a set of categories on the same level. Research specifically on hierarchical web classification is comparatively scarce.

Based on classical "divide and conquer", Dumais and Chen (2000) suggested the use of hierarchical structure for web page classification. It is demonstrated in the paper that splitting the classification problem into a number of sub-problems at each level of the hierarchy is more efficient and accurate than classifying in the non-hierarchical way. Wibowo and Williams (2002b) also studied the problem of hierarchical web classification and suggested methods to minimize errors by shifting the assignment into higher level categories when lower level assignment is uncertain. Peng and Choi (2002) proposed an efficient method to classify a web page into a topical hierarchy and update category information as the hierarchy expands.

Liu et al. (2005) studied the scalability and effectiveness of using SVMs in classifying documents into very large-scale taxonomies. They found that, although hierarchical SVMs are more

efficient than flat SVMs, neither of them can provide satisfying result in terms of effectiveness in large taxonomies. Although hierarchical settings are beneficial to SVM classifiers compared with flat classification, it is also shown (Liu, Yang, Wan, Zhou, Gao, Zeng, Chen, and Ma 2005) that hierarchical settings do more harm than good to k-Nearest Neighbor and naive Bayes classifiers.

With regard to the evaluation of hierarchical classification, Sun and Lim (2001) proposed to measure the performance of hierarchical classification by the *degree* of misclassification, as opposed to measuring the correctness, considering distance measures between the classifier-assigned class and the true class. Kiritchenko (2005) provides a detailed review of hierarchical text categorization.

4.5 Combining information from multiple sources

Methods utilizing different sources of information can be combined to achieve further improvement, especially when the information considered is orthogonal. In web classification, combining link and content information is quite popular (Calado, Cristo, Moura, Ziviani, Ribeiro-Neto, and Goncalves 2003; Qi and Davison 2006).

A common way to combine multiple information is to treat information from different sources as different (usually disjoint) feature sets, on which multiple classifiers are trained. After that, these classifiers are combined together to generate the final decision. There are various methods to combine such classifiers (Kuncheva 2004), including well-developed methods in machine learning such as voting and stacking (Wolpert 1992). Besides these direct combining methods, co-training, as we discussed previously in Section 4.3, is also effective in combining multiple sources since different classifiers are usually trained on disjoint feature sets.

Based on the assumption that each source of information provides a different viewpoint, a combination has the potential to have better knowledge than any single method. However, it often has the disadvantage of additional resource requirements. Moreover, the combination of two does not always perform better than each separately. For example, Calado et al. (2003) show that the combination of a bibliographic coupling similarity measure with a kNN content-based classifier is worse than kNN alone.

5 Other issues

In Sections 3 and 4, we discussed the two important factors in classification: features and algorithms. Other related issues, such as web page preprocessing and dataset selection, also have an effect on classification. We cover such issues here.

5.1 Web page content preprocessing

In most experimental work reviewed in this survey, preprocessing is performed before the content of web pages are fed into a classifier. HTML tags are usually eliminated. However, the content of meta keyword, meta description and "ALT" fields of image tags are usually preserved. Although stemming is often used in web search, it is rarely utilized in classification. The intuition is that stemming is used in indexing mainly in order to improve recall, while in the scenario of web classification, given enough training instances, different forms of a particular term will appear if the term is important.

5.2 Dataset selection and generation

Since web page classification is usually posed as a supervised learning problem, it requires the presence of labeled training instances. In addition, test instances also need to be labeled for the purpose of evaluation.

Since manual labeling can require an excessive amount of human effort, many researchers use subsets of the existing web directories that are publicly available. The two most frequently used web directories are the Yahoo! directory and the dmoz ODP. Others include Cade directory[1] (a Brazilian web directory, now merged with Yahoo!), HanMir[2] (a Korean web directory), and Engineering Electric Library[3] (EEL, a directory providing engineering information).

The use of different datasets makes it difficult to compare performance across multiple algorithms. Therefore, the web classification research community would benefit from a standard dataset for web page classification, such as the TREC datasets (NIST 2007) for information retrieval. Although the "WebKB" dataset (Craven, DiPasquo, Freitag, McCallum, Mitchell, Nigam, and Slattery 1998) is one such dataset, it is small and provides only limited functional classes.

In addition to using web directories as evaluation dataset, one might consider artificially generated datasets. Davidov et al. (2004) proposed a method that can automatically generate labeled datasets for text categorization. Based on existing web directories such as ODP, this algorithm generates datasets with desired properties, e.g., classification difficulty. Although the generated dataset may potentially inherit the bias coming from the web directory on which it is based (such as toward older or high-quality pages), this approach practically eliminates the need for human effort in the generation of some kinds of datasets, while providing flexibility for users to control the characteristics of the generated dataset. Unfortunately, this approach does not provide datasets containing an appropriate web structure that more recent neighborhood-based classification approaches would require.

5.3 Web site classification

Web *sites* (rather than pages) can also be classified. One branch of research only uses a web site's content. Pierre (2001) proposed an approach to the classification of web sites into industry categories utilizing HTML tags.

Another branch focuses on utilizing the structural properties of web sites. It has been shown that there is close correlation between a web site's link structure and its functionality. Amitay et al. (2003) used structural information of a web site to determine its functionality (such as search engines, web directories, corporate sites). Motivated by the same intuition, Lindemann and Littig (2006) further analyzed the relation between structure and functionality of web sites.

There is also research that utilizes both structural and content information. Ester et al. (2002) investigated three different approaches to determining the topical category of a web site based on different web site representations. In their algorithms, a web site can be represented by a single virtual page consisting of all pages in the site, by a vector of topic frequencies, or by a tree of its pages with topics. Experiments showed that the tree classification approach offers the best accuracy. Tian et al. (2003) proposed to represent a web site by a two-layer tree model in which each page is modeled by a DOM (Document Object Model) tree and a site is represented by a hierarchical tree constructed according to the links among pages. Then a Hidden-Markov-Tree-based classifier is used for classification.

Classification of web pages is helpful to classifying a web site. For example, in (Ester, Kriegel, and Schubert 2002), knowing the topic of the pages in a site can help determine the web site's topic. Presumably, site categorization could also benefit web page classification but the results of such an approach have not been reported.

[1]Cade: `http://br.cade.yahoo.com/`
[2]HanMir: `http://hanmir.com/`
[3]EEL: `http://eels.lub.lu.se/`

5.4 Blog classification

The word "blog" was originally a short form of "web log", which, as defined by the Merriam-Webster Dictionary[4], is a web site that contains online personal journal with reflections, comments, and often hyperlinks provided by the writer. As blogging has gained in popularity in recent years, an increasing amount of research about blogs has also been conducted. Research in blog classification can be broken into three types: blog identification (to determine whether a web document is a blog), mood classification, and genre classification.

Research in the first category aims at identifying blog pages from a collection of web pages, which is essentially a binary classification of blog and non-blog. Nanno et al. (2004) presented a system that automatically collects and monitors blog collections, identifying blog pages based on a number of simple heuristics. Elgersma and Rijke (2006) examined the effectiveness of common classification algorithms on blog identification tasks. Using a number of human-selected features (some of which are blog specific, e.g., whether characteristic terms are present, such as "Comments" and "Archives"), they found that many off-the-shelf machine learning algorithms can yield satisfactory classification accuracy (around 90%).

The second category of research includes identification of the mood or sentiment of blogs. Most existing approaches in this category recognize mood at the level of individual post; some target recognition of mood reflected by a collection of posts. Mihalcea and Liu (2006) showed that blog entries expressing the two polarities of moods, happiness and sadness, are separable by their linguistic content. A naive Bayes classifier trained on unigram features achieved 79% accuracy over 10,000 mood-annotated blogposts. Similarly, Chesley et al. (2006) demonstrated encouraging performance in categorizing blog posts into three sentiment classes (Objective, Positive, and Negative). However, real-world blog posts indicate moods much more complicated than merely happiness and sadness (or positive and negative). Mishne (2005) showed that classifying blog posts into a more comprehensive set of moods is a challenging task (for both machine and human). When doing binary classification of blog posts on more than a hundred predefined moods, SVM classifiers trained on a variety of features (content and non-content) made small (while consistent) improvement (8%) over random guess. Using a similar approach, Leshed and Kaye (2006) achieved 76% overall accuracy when classifying into 50 most frequent moods. While recognizing mood for individual blog posts can be difficult, later work done by Mishne and Rijke (2006) showed that determining aggregate mood across a large collection of blog posts can achieve a high accuracy.

The third category focuses on the genre of blogs. Research in this category is usually done at blog level. Nowson (2006) discussed the distinction of three types of blogs: news, commentary, and journal. Qu et al. (2006) proposed an approach to automatic classification of blogs into four genres: personal diary, news, political, and sports. Using unigram tfidf document representation and naive Bayes classification, Qu et al.'s approach can achieve an accuracy of 84%.

So far, it seems research from both the second and the third category suffers from the lack of a well-defined taxonomy. For mood/sentiment classification, one of the main reasons is the blurry boundary of different moods. In genre classification, the intended purpose of a blog can evolve over time, making genre distinctions difficult.

6 Conclusion

After reviewing web classification research with respect to its features and algorithms, we conclude this article by summarizing the lessons we have learned from existing research and pointing out future opportunities in web classification.

Web page classification is a type of supervised learning problem that aims to categorize web pages into a set of predefined categories based on labeled training data. Classification

[4] `http://mw1.merriam-webster.com/dictionary/blog`

tasks include assigning documents on the basis of subject, function, sentiment, genre, and more. Unlike more general text classification, web page classification methods can take advantage of the semi-structured content and connections to other pages within the Web.

We have surveyed the space of published approaches to web page classification from various viewpoints, and summarized their findings and contributions, with a special emphasis on the utilization and benefits of web-specific features and methods.

We found that while the appropriate use of textual and visual features that reside directly on the page can improve classification performance, features from neighboring pages provide significant supplementary information to the page being classified. Feature selection and the combination of multiple techniques can bring further improvement.

We expect that future web classification efforts will certainly combine content and link information in some form. In the context of the research surveyed here, future work would be well-advised to:

- Emphasize text and labels from siblings (co-cited pages) over other types of neighbors;
- Incorporate anchortext from parents; and,
- Utilize other sources of (implicit or explicit) human knowledge, such as query logs and click-through behavior, in addition to existing labels to guide classifier creation.

In Section 2.1 we described the a variety of classification problems that can be applied to web pages. However, most web (and text) classification work focuses on hard classification with a single label per document. Some applications, however, require multi-label classification and even soft classification—such as in the form of probability distributions over the possible classes. Multi-label and soft classification better represent the real world—documents are rarely well-represented by a single predefined topic. Unfortunately, complexity of evaluation and a lack of appropriate datasets has prevented straightforward progress in these areas. Although somewhat more difficult, we suggest that embedding such classification systems within a more easily evaluated task (perhaps just measuring 'user satisfaction' of the resulting system) would be a productive path forward.

Although our survey has included many efforts in web classification, additional opportunities for improvement in classification remain. Future work might address one or more of the following:

- How much can intermediate classification help? This question applies both to assisting in web page classification, and the converse, to use web page classification to aid in another task. This might apply to web page and site classification, or benefit other tasks such as query classification. It might also be applicable across different types of classification, such as using topical classes to enhance functional classification.
- How much do text and link similarity measures reflect the semantic similarity between documents? Although the work by Menczer and colleagues (Menczer 2005; Maguitman, Menczer, Roinestad, and Vespignani 2005) casts some light on this question, a more definitive answer requires further study.
- The use of information from neighboring nodes is valuable. But such information is also noisy, and so all neighbors (even of the same type) are unlikely to be equally valuable. How might neighbors (or portions of neighbors) be weighted or selected to best match the likely value of the evidence provided?
- Hyperlink information often encodes semantic relationships along with voting for representative or important pages. Would the complete integration of content information into link form be beneficial? This could be performed with various types of artificial links as we have surveyed, or in a combined model of underlying factors, as in the combination of PHITS and PLSA (Cohn and Hofmann 2001) for web information retrieval.
- Search engine spam (Gyöngyi and Garcia-Molina 2005a) is a significant concern in web information retrieval. What effect does web spam have on topical or functional classification?

- The lack of a standardized dataset, especially one with the spatial locality representative of the web, is a significant disadvantage in web classification research. How can a truly representative dataset with these properties that is multiple orders of magnitudes smaller than the actual Web be selected?

It is expected that solutions or even a better understanding of these problems may lead to the emergence of more effective web classification systems, as well as improvements in other areas of information retrieval and web mining.

Finally, we wish to explicitly note the connection between machine learning and information retrieval (especially ranking). This idea is not new, and has underpinned many of the ideas in the work presented here. A learning to rank community (Joachims 2002; Burges, Shaked, Renshaw, Lazier, Deeds, Hamilton, and Hullender 2005; Radlinski and Joachims 2005; Roussinov and Fan 2005; Agarwal 2006; Cao, Xu, Liu, Li, Huang, and Hon 2006; Richardson, Prakash, and Brill 2006) is making advances for both static and query-specific ranking. There may be unexplored opportunities for using retrieval techniques for classification as well. In general, many of the features and approaches for web page classification have counterparts in analysis for web page retrieval. Future advances in web page classification should be able to inform retrieval and vice versa.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0328825.

References

Aas, K. and L. Eikvil (1999, June). Text categorisation: A survey. Technical report, Norwegian Computing Center, P.B. 114 Blindern, N-0314, Oslo, Norway. Technical Report 941.

Agarwal, S. (2006). Ranking on graph data. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, New York, NY, pp. 25–32. ACM Press.

Amitay, E. (1998). Using common hypertext links to identify the best phrasal description of target web documents. In *Proceedings of the SIGIR'98 Post-Conference Workshop on Hypertext Information Retrieval for the Web*, Melbourne, Australia.

Amitay, E., D. Carmel, A. Darlow, R. Lempel, and A. Soffer (2003). The connectivity sonar: Detecting site functionality by structural patterns. In *HYPERTEXT '03: Proceedings of the 14th ACM Conference on Hypertext and Hypermedia*, New York, NY, pp. 38–47. ACM Press.

Angelova, R. and S. Siersdorfer (2006). A neighborhood-based approach for clustering of linked document collections. In *CIKM '06: Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, New York, NY, pp. 778–779. ACM Press.

Angelova, R. and G. Weikum (2006). Graph-based text classification: Learn from your neighbors. In *SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, pp. 485–492. ACM Press.

Armstrong, R., D. Freitag, T. Joachims, and T. Mitchell (1995, March). WebWatcher: A learning apprentice for the World Wide Web. In *Proceedings of the AAAI Spring Symposium on Information Gathering from Distributed, Heterogeneous Environments*, Menlo Park, CA, pp. 6–12. AAAI Press.

Asirvatham, A. P. and K. K. Ravi (2001). Web page classification based on document structure. Awarded second prize in National Level Student Paper Contest conducted by IEEE India Council.

Attardi, G., A. Gulli, and F. Sebastiani (1999). Automatic web page categorization by link and context analysis. In C. Hutchison and G. Lanzarone (Eds.), *Proceedings of THAI'99, First European Symposium on Telematics, Hypermedia and Artificial Intelligence*, Varese, IT, pp. 105–119.

Beeferman, D. and A. Berger (2000). Agglomerative clustering of a search engine query log. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, pp. 407–415. ACM Press.

Bennett, P. N., S. T. Dumais, and E. Horvitz (2005). The combination of text classifiers using reliability indicators. *Information Retrieval 8*(1), 67–100.

Blum, A. and T. Mitchell (1998). Combining labeled and unlabeled data with co-training. In *COLT' 98: Proceedings of the 11th Annual Conference on Computational Learning Theory*, New York, NY, pp. 92–100. ACM Press.

Burges, C., T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender (2005). Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pp. 89–96.

Calado, P., M. Cristo, E. Moura, N. Ziviani, B. Ribeiro-Neto, and M. A. Goncalves (2003). Combining link-based and content-based methods for web document classification. In *CIKM '03: Proceedings of the 12th International Conference on Information and Knowledge Management*, New York, NY, pp. 394–401. ACM Press.

Cao, Y., J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon (2006). Adapting ranking svm to document retrieval. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, pp. 186–193. ACM Press.

Cardoso-Cachopo, A. and A. L. Oliveira (2003, October). An empirical comparison of text categorization methods. In *Proceedings of the 10th International Symposium on String Processing and Information Retrieval (SPIRE)*, Volume 2857 of *LNCS*, Berlin, pp. 183–196. Springer.

Castillo, C., D. Donato, A. Gionis, V. Murdock, and F. Silvestri (2007). Know your neighbors: Web spam detection using the web topology. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY. ACM Press. In press.

Chakrabarti, S. (2000, January). Data mining for hypertext: a tutorial survey. *SIGKDD Explorations Newsletter 1*(2), 1–11.

Chakrabarti, S. (2003). *Mining the Web: Discovering Knowledge from Hypertext Data*. San Francisco, CA: Morgan Kaufmann.

Chakrabarti, S., B. E. Dom, and P. Indyk (1998). Enhanced hypertext categorization using hyperlinks. In *SIGMOD '98: Proceedings of the ACM SIGMOD International Conference on Management of Data*, New York, NY, pp. 307–318. ACM Press.

Chakrabarti, S., M. M. Joshi, K. Punera, and D. M. Pennock (2002). The structure of broad topics on the web. In *WWW '02: Proceedings of the 11th International Conference on World Wide Web*, New York, NY, pp. 251–262. ACM Press.

Chakrabarti, S., M. van den Berg, and B. Dom (1999, May). Focused crawling: A new approach to topic-specific Web resource discovery. In *WWW '99: Proceeding of the 8th International Conference on World Wide Web*, New York, NY, pp. 1623–1640. Elsevier.

Chekuri, C., M. Goldwasser, P. Raghavan, and E. Upfal (1997, April). Web search using automated classification. In *Proceedings of the Sixth International World Wide Web Conference*, Santa Clara, CA. Poster POS725.

Chen, H. and S. Dumais (2000). Bringing order to the Web: Automatically categorizing search results. In *CHI '00: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, pp. 145–152. ACM Press.

Chen, Z., O. Wu, M. Zhu, and W. Hu (2006). A novel web page filtering system by combining texts and images. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, Washington, DC, pp. 732–735. IEEE Computer Society.

Chesley, P., B. Vincent, L. Xu, and R. K. Srihari (2006, March). Using verbs and adjectives to automatically classify blog sentiment. In N. Nicolov, F. Salvetti, M. Liberman, and J. H. Martin (Eds.), *Computational Approaches to Analyzing Weblogs: Papers from the 2006 Spring Symposium*, Menlo Park, CA, pp. 27–29. AAAI Press. Technical Report SS-06-03.

Choi, B. and Z. Yao (2005). Web page classification. In W. Chu and T. Y. Lin (Eds.), *Foundations and Advances in Data Mining*, Volume 180 of *Studies in Fuzziness and Soft Computing*, pp. 221–274. Berlin: Springer-Verlag.

Cohen, W. W. (2002). Improving a page classifier with anchor extraction and link analysis. In S. Becker, S. Thrun, and K. Obermayer (Eds.), *Advances in Neural Information Processing Systems*, Volume 15, pp. 1481–1488. Cambridge, MA: MIT Press.

Cohn, D. and T. Hofmann (2001). The missing link — a probabilistic model of document content and hypertext connectivity. In *Advances in Neural Information Processing Systems (NIPS) 13*.

Corporation, N. C. (2007). The dmoz open Directory Project (ODP). `http://www.dmoz.com/`.

Craven, M., D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery (1998, July). Learning to extract symbolic knowledge from the World Wide Web. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, Menlo Park, CA, USA, pp. 509–516. AAAI Press.

Davidov, D., E. Gabrilovich, and S. Markovitch (2004). Parameterized generation of labeled datasets for text categorization based on a hierarchical directory. In *SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, pp. 250–257. ACM Press.

Davison, B. D. (2000, July). Topical locality in the Web. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, pp. 272–279. ACM Press.

Davison, B. D. (2004, November). The potential of the metasearch engine. In *Proceedings of the Annual Meeting of the American Society for Information Science and Technology*, Volume 41, Providence, RI, pp. 393–402. American Society for Information Science & Technology.

Deerwester, S. C., S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science 41*(6), 391–407.

Dietterich, T. G. and G. Bakiri (1995). Solving multiclass learning problems via error-correcting output codes. *Journal of Articial Intelligence Research 2*, 263–286.

Drost, I., S. Bickel, and T. Scheffer (2005, March). Discovering communities in linked data by multi-view clustering. In *From Data and Information Analysis to Knowledge Engineering: Proceedings of 29th Annual Conference of the German Classification Society*, Studies in Classification, Data Analysis, and Knowledge Organization, Berlin, pp. 342–349. Springer.

Duda, R. O. and P. E. Hart (1973). *Pattern Classification and Scene Analysis*. New York, NY: Wiley.

Dumais, S. and H. Chen (2000). Hierarchical classification of web content. In *SIGIR '00: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, pp. 256–263. ACM Press.

Elgersma, E. and M. de Rijke (2006, March). Learning to recognize blogs: A preliminary exploration. In *EACL 2006 Workshop: New Text - Wikis and blogs and other dynamic text sources*.

Ester, M., H.-P. Kriegel, and M. Schubert (2002). Web site mining: A new way to spot competitors, customers and suppliers in the World Wide Web. In *KDD '02: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, pp. 249–258. ACM Press.

Fisher, M. J. and R. M. Everson (2003, April). When are links useful? Experiments in text classification. In *Advances in Information Retrieval. 25th European Conference on IR Research*, pp. 41–56. Springer.

Fitzpatrick, L. and M. Dent (1997, July). Automatic feedback using past queries: Social searching? In *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, pp. 306–313. ACM Press.

Fürnkranz, J. (1999). Exploiting structural information for text classification on the WWW. In D. J. Hand, J. N. Kok, and M. R. Berthold (Eds.), *Proceedings of the 3rd Symposium on Intelligent Data Analysis (IDA-99)*, Volume 1642 of *LNCS*, Amsterdam, Netherlands, pp. 487–497. Springer-Verlag.

Fürnkranz, J. (2001). Hyperlink ensembles: A case study in hypertext classification. *Journal of Information Fusion 1*, 299–312.

Fürnkranz, J. (2005). Web mining. In O. Maimon and L. Rokach (Eds.), *The Data Mining and Knowledge Discovery Handbook*, pp. 899–920. Berlin: Springer.

Gabrilovich, E. and S. Markovitch (2004). Text categorization with many redundant features: Using aggressive feature selection to make SVMs competitive with C4.5. In *Proceedings of the 21st International Conference on Machine learning*, New York, NY, pp. 41. ACM Press.

Gabrilovich, E. and S. Markovitch (2005, July). Feature generation for text categorization using world knowledge. In *Proceedings of the 19th International Joint Conference for Artificial Intelligence (IJCAI)*, pp. 1048–1053.

Gabrilovich, E. and S. Markovitch (2006, July). Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the 21st National Conference on Artificial Intelligence*, Menlo Park, CA, pp. 1301–1306. AAAI Press.

Getoor, L. and C. Diehl (2005, December). Link mining: A survey. *SIGKDD Explorations Newsletter Special Issue on Link Mining 7*(2).

Ghani, R. (2001). Combining labeled and unlabeled data for text classification with a large number of categories. In *First IEEE International Conference on Data Mining (ICDM)*, Los Alamitos, CA, pp. 597. IEEE Computer Society.

Ghani, R. (2002). Combining labeled and unlabeled data for multiclass text categorization. In *ICML '02: Proceedings of the 19th International Conference on Machine Learning*, San Francisco, CA, pp. 187–194. Morgan Kaufmann.

Ghani, R., S. Slattery, and Y. Yang (2001). Hypertext categorization using hyperlink patterns and meta data. In *ICML '01: Proceedings of the 18th International Conference on Machine Learning*, San Francisco, CA, pp. 178–185. Morgan Kaufmann.

Glance, N. S. (2000, July). Community search assistant. In *Artificial Intelligence for Web Search*, pp. 29–34. AAAI Press. Presented at the AAAI-2000 workshop on Artificial Intelligence for Web Search, Technical Report WS-00-01.

Glover, E. J., K. Tsioutsiouliklis, S. Lawrence, D. M. Pennock, and G. W. Flake (2002). Using web structure for classifying and describing web pages. In *Proceedings of the 11th International Conference on World Wide Web*, New York, NY, pp. 562–569. ACM Press.

Golub, K. and A. Ardo (2005, September). Importance of HTML structural elements and metadata in automated subject classification. In *Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, Volume 3652 of *LNCS*, Berlin, pp. 368–378. Springer.

Gövert, N., M. Lalmas, and N. Fuhr (1999). A probabilistic description-oriented approach for categorizing web documents. In *CIKM '99: Proceedings of the 8th International Conference on Information and Knowledge Management*, New York, NY, pp. 475–482. ACM Press.

Gyöngyi, Z. and H. Garcia-Molina (2005a). Link spam alliances. In *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB)*, Trondheim, Norway.

Gyöngyi, Z. and H. Garcia-Molina (2005b, May). Web spam taxonomy. In B. D. Davison (Ed.), *Proceedings of the First International Workshop on Adversarial Information Retrieval (AIRWeb)*, Bethlehem, PA, pp. 39–47. Lehigh University, Department of Computer Science. Technical Report LU-CSE-05-030.

Hammami, M., Y. Chahir, and L. Chen (2003). Webguard: Web based adult content detection and filtering system. In *WI '03: Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence*, Washington, DC, pp. 574. IEEE Computer Society.

Harabagiu, S. M., M. A. Pasca, and S. J. Maiorano (2000). Experiments with open-domain textual question answering. In *Proceedings of the 18th Conference on Computational Linguistics*, Morristown, NJ, USA, pp. 292–298. Association for Computational Linguistics.

Haveliwala, T. H. (2002, May). Topic-sensitive PageRank. In *Proceedings of the Eleventh International World Wide Web Conference*, New York, NY, pp. 517–526. ACM Press.

Hermjakob, U. (2001, July). Parsing and question classification for question answering. In *Proceedings of the ACL Workshop on Open-Domain Question Answering*, pp. 1–6.

Hofmann, T. (1999a). Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm.

Hofmann, T. (1999b). Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, pp. 50–57. ACM Press.

Huang, C.-C., S.-L. Chuang, and L.-F. Chien (2004a). Liveclassifier: Creating hierarchical text classifiers through web corpora. In *WWW '04: Proceedings of the 13th International Conference on World Wide Web*, New York, NY, pp. 184–192. ACM Press.

Huang, C.-C., S.-L. Chuang, and L.-F. Chien (2004b). Using a web-based categorization approach to generate thematic metadata from texts. *ACM Transactions on Asian Language Information Processing (TALIP) 3*(3), 190–212.

Jensen, D., J. Neville, and B. Gallagher (2004). Why collective inference improves relational classification. In *KDD '04: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, pp. 593–598. ACM Press.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 133–142. ACM Press.

Joachims, T., D. Freitag, and T. Mitchell (1997, August). WebWatcher: A tour guide for the World Wide Web. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pp. 770–775. Morgan Kaufmann.

Käki, M. (2005). Findex: Search result categories help users when document ranking fails. In *CHI '05: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, pp. 131–140. ACM Press.

Kan, M.-Y. (2004). Web page classification without the web page. In *WWW Alt. '04: Proceedings of the 13th International World Wide Web Conference Alternate Track Papers & Posters*, New York, NY, pp. 262–263. ACM Press.

Kan, M.-Y. and H. O. N. Thi (2005). Fast webpage classification using URL features. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM '05)*, New York, NY, pp. 325–326. ACM Press.

Kiritchenko, S. (2005). *Hierarchical Text Categorization and Its Application to Bioinformatics*. Ph. D. thesis, University of Ottawa.

Kohlschutter, C., P.-A. Chirita, and W. Nejdl (2007). Utility analysis for topically biased PageRank. In *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, New York, NY, pp. 1211–1212. ACM Press.

Kosala, R. and H. Blockeel (2000, June). Web mining research: A survey. *SIGKDD Explorations Newsletter 2*(1), 1–15.

Kovacevic, M., M. Diligenti, M. Gori, and V. Milutinovic (2004, September). Visual adjacency multigraphs - a novel approach for a web page classification. In *Proceedings of the Workshop on Statistical Approaches to Web Mining (SAWM)*, pp. 38–49.

Kuncheva, L. I. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience.

Kurland, O. and L. Lee (2005, July). PageRank without hyperlinks: Structural re-ranking using links induced by language models. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, pp. 306–313. ACM Press.

Kurland, O. and L. Lee (2006, July). Respect my authority!: HITS without hyperlinks, utilizing cluster-based language models. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, pp. 83–90. ACM Press.

Kwok, C. C. T., O. Etzioni, and D. S. Weld (2001). Scaling question answering to the web. In *WWW '01: Proceedings of the 10th International Conference on World Wide Web*, New York, NY, pp. 150–161. ACM Press.

Kwon, O.-W. and J.-H. Lee (2000). Web page classification based on k-nearest neighbor approach. In *IRAL '00: Proceedings of the 5th International Workshop on Information Retrieval with Asian languages*, New York, NY, pp. 9–15. ACM Press.

Kwon, O.-W. and J.-H. Lee (2003, January). Text categorization based on k-nearest neighbor approach for web site classification. *Information Processing and Management 29*(1), 25–44.

Leshed, G. and J. J. Kaye (2006). Understanding how bloggers feel: Recognizing affect in blog posts. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, pp. 1019–1024. ACM Press.

Lindemann, C. and L. Littig (2006). Coarse-grained classification of web sites by their structural properties. In *WIDM '06: Proceedings of the 8th ACM International Workshop on Web Information and Data Management*, New York, NY, pp. 35–42. ACM Press.

Liu, T.-Y., Y. Yang, H. Wan, H.-J. Zeng, Z. Chen, and W.-Y. Ma (2005). Support vector machines classification with a very large-scale taxonomy. *SIGKDD Explorations Newsletter 7*(1), 36–43.

Liu, T.-Y., Y. Yang, H. Wan, Q. Zhou, B. Gao, H.-J. Zeng, Z. Chen, and W.-Y. Ma (2005, May). An experimental study on large-scale web categorization. In *WWW '05: Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, New York, NY, pp. 1106–1107. ACM Press.

Lu, Q. and L. Getoor (2003, August). Link-based classification. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, Menlo Park, CA. AAAI Press.

Luxenburger, J. and G. Weikum (2004, November). Query-log based authority analysis for web information search. In *Proceedings of 5th International Conference on Web Information Systems Engineering (WISE)*, Volume 3306 of *LNCS*, Berlin, pp. 90–101. Springer.

Macskassy, S. A. and F. Provost (2007, May). Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research 8*, 935–983.

Maguitman, A. G., F. Menczer, H. Roinestad, and A. Vespignani (2005). Algorithmic detection of semantic similarity. In *WWW '05: Proceedings of the 14th International Conference on World Wide Web*, New York, NY, pp. 107–116. ACM Press.

Menczer, F. (2005, May/June). Mapping the semantics of web text and links. *IEEE Internet Computing 9*(3), 27–36.

Mihalcea, R. and H. Liu (2006, March). A corpus-based approach to finding happiness. In N. Nicolov, F. Salvetti, M. Liberman, and J. H. Martin (Eds.), *Computational Approaches to Analyzing Weblogs: Papers from the 2006 Spring Symposium*, Menlo Park, CA, pp. 139–144. AAAI Press. Technical Report SS-06-03.

Mishne, G. (2005, August). Experiments with mood classification in blog posts. In *Workshop on Stylistic Analysis of Text for Information Access*.

Mishne, G. and M. de Rijke (2006, March). Capturing global mood levels using blog posts. In N. Nicolov, F. Salvetti, M. Liberman, and J. H. Martin (Eds.), *Computational Approaches to Analyzing Weblogs: Papers from the 2006 Spring Symposium*, Menlo Park, CA, pp. 145–152. AAAI Press. Technical Report SS-06-03.

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

Mladenic, D. (1998). Turning Yahoo into an automatic web-page classifier. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, pp. 473–474.

Mladenic, D. (1999, Jul/Aug). Text-learning and related intelligent agents: A survey. *IEEE Intelligent Systems and Their Applications 14*(4), 44–54.

Nanno, T., T. Fujiki, Y. Suzuki, and M. Okumura (2004). Automatically collecting, monitoring, and mining Japanese weblogs. In *WWW Alt. '04: Proceedings of the 13th international World Wide Web Conference on Alternate Track Papers & Posters*, New York, NY, pp. 320–321. ACM Press.

Nie, L., B. D. Davison, and X. Qi (2006, August). Topical link analysis for web search. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval*, New York, NY, pp. 91–98. ACM Press.

NIST (2007). Text REtrieval Conference (TREC) home page. http://trec.nist.gov/.

Nowson, S. (2006, June). *The Language of Weblogs: A study of genre and individual differences*. Ph. D. thesis, University of Edinburgh, College of Science and Engineering.

Oh, H.-J., S. H. Myaeng, and M.-H. Lee (2000). A practical hypertext catergorization method using links and incrementally available class information. In *SIGIR '00: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, pp. 264–271. ACM Press.

Page, L., S. Brin, R. Motwani, and T. Winograd (1998). The PageRank citation ranking: Bringing order to the Web. Unpublished draft, Stanford University.

Park, S.-B. and B.-T. Zhang (2003, April). Large scale unstructured document classification using unlabeled data and syntactic information. In *Advances in Knowledge Discovery and Data Mining: 7th Pacific-Asia Conference (PAKDD)*, Volume 2637 of *LNCS*, Berlin, pp. 88–99. Springer.

Pazzani, M., J. Muramatsu, and D. Billsus (1996). Syskill & webert: Identifying interesting Web sites. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, Menlo Park, CA, pp. 54–61. AAAI Press.

Peng, X. and B. Choi (2002). Automatic web page classification in a dynamic and hierarchical way. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'02)*, Washington, DC, pp. 386–393. IEEE Computer Society.

Pierre, J. M. (2001, February). On the automated classification of web sites. *Linköping Electronic Articles in Computer and Information Science 6*. http://www.ep.liu.se/ea/cis/2001/001/.

Qi, X. and B. D. Davison (2006). Knowing a web page by the company it keeps. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM)*, New York, NY, pp. 228–237. ACM Press.

Qu, H., A. L. Pietra, and S. Poon (2006, March). Automated blog classification: Challenges and pitfalls. In N. Nicolov, F. Salvetti, M. Liberman, and J. H. Martin (Eds.), *Computational Approaches to Analyzing Weblogs: Papers from the 2006 Spring Symposium*, Menlo Park, CA, pp. 184–186. AAAI Press. Technical Report SS-06-03.

Radlinski, F. and T. Joachims (2005). Query chains: learning to rank from implicit feedback. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, New York, NY, USA, pp. 239–248. ACM Press.

Riboni, D. (2002). Feature selection for web page classification. In *Proceedings of the Workshop on Web Content Mapping: A Challenge to ICT (EURASIA-ICT)*.

Richardson, M., A. Prakash, and E. Brill (2006). Beyond pagerank: machine learning for static ranking. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, New York, NY, pp. 707–715. ACM Press.

Rosenfeld, A., R. Hummel, and S. Zucker (1976). Scene labeling by relaxation operations. *IEEE Transactions on Systems, Man and Cybernetics 6*, 420–433.

Roussinov, D. and W. Fan (2005). Discretization based learning approach to information retrieval. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Morristown, NJ, pp. 153–160. Association for Computational Linguistics.

Sebastiani, F. (1999). A tutorial on automated text categorisation. In *Proceedings of 1st Argentinean Symposium on Artificial Intelligence (ASAI-99)*, pp. 7–35.

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

Sen, P. and L. Getoor (2007). Link-based classification. Technical Report CS-TR-4858, University of Maryland.

Shanks, V. and H. E. Williams (2001, November). Fast categorisation of large document collections. In *Proceedings of Eighth International Symposium on String Processing and Information Retrieval (SPIRE)*, pp. 194–204.

Shen, D., Z. Chen, Q. Yang, H.-J. Zeng, B. Zhang, Y. Lu, and W.-Y. Ma (2004). Web-page classification through summarization. In *SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, pp. 242–249. ACM Press.

Shen, D., J.-T. Sun, Q. Yang, and Z. Chen (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, pp. 643–650. ACM Press.

Slattery, S. and T. M. Mitchell (2000, June). Discovering test set regularities in relational domains. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*, pp. 895–902. Morgan Kaufmann.

Sun, A. and E.-P. Lim (2001, November). Hierarchical text classification and evaluation. In *Proceedings of IEEE International Conference on Data Mining (ICDM)*, Washington, DC, pp. 521–528. IEEE Computer Society.

Sun, A., E.-P. Lim, and W.-K. Ng (2002). Web classification using support vector machine. In *WIDM '02: Proceedings of the 4th International Workshop on Web Information and Data Management*, New York, NY, pp. 96–99. ACM Press.

Tan, A.-H. (1999). Text mining: The state of the art and the challenges. In *Proceedings of PAKDD Workshop on Knowledge Discoverery from Advanced Databases*, pp. 65–70.

Tian, Y., T. Huang, W. Gao, J. Cheng, and P. Kang (2003). Two-phase web site classification based on hidden markov tree models. In *WI '03: Proceedings of the IEEE/WIC International Conference on Web Intelligence*, Washington, DC, USA, pp. 227. IEEE Computer Society.

Tong, S. and D. Koller (2001, November). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research 2*, 45–66.

Utard, H. and J. Fürnkranz (2005, October). Link-local features for hypertext classification. In *Semantics, Web and Mining: Joint International Workshops, EWMF/KDO*, Volume 4289 of *LNCS*, Berlin, pp. 51–64. Springer.

Wen, J.-R., J.-Y. Nie, and H.-J. Zhang (2002). Query clustering using user logs. *ACM Transactions on Information Systems (TOIS) 20*(1), 59–81.

Wibowo, W. and H. E. Williams (2002a). Simple and accurate feature selection for hierarchical categorisation. In *DocEng '02: Proceedings of the 2002 ACM Symposium on Document Engineering*, New York, NY, pp. 111–118. ACM Press.

Wibowo, W. and H. E. Williams (2002b). Strategies for minimising errors in hierarchical web categorisation. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM '02)*, New York, NY, pp. 525–531. ACM Press.

Wolpert, D. (1992). Stacked generalization. *Neural Networks 5*, 241–259.

Xue, G.-R., Y. Yu, D. Shen, Q. Yang, H.-J. Zeng, and Z. Chen (2006). Reinforcing web-object categorization through interrelationships. *Data Mining and Knowledge Discovery 12*(2-3), 229–248.

Yahoo!, Inc. (2007). Yahoo! `http://www.yahoo.com/`.

Yang, H. and T.-S. Chua (2004a). Effectiveness of web page classification on finding list answers. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, pp. 522–523. ACM Press.

Yang, H. and T.-S. Chua (2004b). Web-based list question answering. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, Morristown, NJ, pp. 1277. Association for Computational Linguistics.

Yang, Y. and J. O. Pedersen (1997). A comparative study on feature selection in text categorization. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, San Francisco, CA, pp. 412–420. Morgan Kaufmann.

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

Yang, Y., J. Zhang, and B. Kisiel (2003). A scalability analysis of classifiers in text categorization. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, New York, NY, pp. 96–103. ACM Press.

Yu, H., J. Han, and K. C.-C. Chang (2004). PEBL: Web page classification without negative examples. *IEEE Transactions on Knowledge and Data Engineering 16*(1), 70–81.

Zaiane, O. R. and A. Strilets (2002, September). Finding similar queries to satisfy searches based on query traces. In *Proceedings of the International Workshop on Efficient Web-Based Information Systems (EWIS)*, Volume 2426 of *LNCS*, Berlin, pp. 207–216. Springer.

Zelikovitz, S. and H. Hirsh (2001). Using LSI for text classification in the presence of background text. In *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM)*, New York, NY, pp. 113–118. ACM Press.

Zhang, D. and W. S. Lee (2003). Question classification using support vector machines. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, New York, NY, pp. 26–32. ACM Press.

zu Eissen, S. M. and B. Stein (2004). Genre classification of web pages. In *Proceedings of the 27th German Conference on Artificial Intelligence*, Volume 3238 of *LNCS*, Berlin, pp. 256–269. Springer.