International Journal of Science and Engineering Invention (IJSEI)
Volume 01 Issue 01 November 2015
Available Online at www.isij.in, page no. 1to 8

WEB SITE CLASSIFICATION FEATURES AND ALGORITHMS -A SURVEY

C.Gunasundari[1], N.Eswari[2] G.Vignesh[3]
Assistant Professor, Department of Computer science and Engineering
Roever College of Engineering and Technology, Perambalur, Tamilnadu-621 220
Email id- gunasundari.cs@gmail.com

Abstract: In recent years the Internet has massive growth of data stored in various forms. There is a need for innovative and effective technologies to help find and use the valuable information and knowledge from a multiple disciplinary crew. Always the data is not to be static it is dynamically increasing and varying In order to utilize the Web information better, people pursue the latest technology, which can effectively organize and use online information. Classification of Web page content is important 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. In this paper in Web page classification, to indicate the importance of these Web-specific features and algorithms.

Keywords: Web mining, Classification.

1. Introduction

Internet is a shared world wide computing network. All computing devices are connected for sharing data and getting services. This platform provides web services and Internet services for people. Web contains lot of resources which is freely available for Internet users. The web does not contain single form of data, It has several types of multimedia data like text, image, audio, video and etc. The different forms of data have to be well organized in proper way, so that only it will be efficiently used. Data mining defines extraction of data in terms of patterns or rules from huge amount of data [1]. The term web mining was coined by Etzioni in 1996, to denote the use of data mining techniques to automatically discover web documents, extract information from web resources and uncover general patterns on the web.

The research in the field of web is classified on two ways: information retrieval and knowledge discovery. The information retrieval focuses on retrieving relevant information from large storage area whereas mining research focuses on extracting new information already existing data[3]. In past, techniques like information extraction, information retrieval and machine learning were used to discover new knowledge from huge amount of data available on web. Information extraction focuses on extracting relevant facts whereas information retrieval focus selects relevant document. Now, Web mining is a part of both information extraction and information retrieval. Web mining supports machine learning because it improves the classification of text [4]. The main aim of web mining is to extract information. Web mining is integration of information that is gathered by traditional data mining techniques with information gathered over World Wide Web.

(i) To discover resources: It helps in retrieving services and unknown documents on web.

(ii) To select Information and preprocessing: It automatically selects and preprocesses specific information from the web sources.

(iii) Generalization: It uncovers general pattern at individual web sites as well as across multiple sites.

(iv) Analysis: It validates and interprets the mined pattern.

(v) Visualization: It presents the result in visual and easy to understand way. Web mining is divided into three main categories depending on the type of data as web content mining, web structure mining and web usage mining [2].

Web mining deals with three main areas: web content mining, web usage mining and web structure mining. Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to web content has been the most widely researched. Issues addressed in text mining include topic discovery and tracking, extracting association patterns, clustering of web documents and classification of web pages. Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP).

Web structure mining aims at developing techniques to identify quality of web page which can be find out with the help of hyperlinks. For example, from the links, we can discover important Web pages, which, incidentally, is a key technology used in search engines. We can also discover communities of users who share common interests. Traditional data mining does not perform such tasks because there is usually no link structure in a relational table.

Web usage mining focuses on techniques to study the user behavior when navigating the web. It is also known as Web log mining. It refers to the discovery of user access patterns from Web usage logs, which record every click made by each user.

2. Classification

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, to contextual advertising, and to analysis of the topical structure of the Web. Web page classification can also help improve the quality of Web search.

The problem of data classification has various applications in a wide variety of mining applications. This is because the problem attempts to learn the relationship between a set of attribute variables and a target variable of interest. Since many practical problems can be expressed as associations between feature and target variables, this provides a broad range of applicability of this model.

The problem of classification may be stated as follows:

Given a set of training data points along with associated training labels, determine the class label for an unlabeled test instance. Some common application domains, in which the classification problem arises, are as follows:

- **Customer Target Marketing:** Since the classification problem relates feature variables to target classes, this method is extremely popular for the problem of customer target marketing. In such cases,

feature variables describing the customer may be used to predict their buying interests on the basis of previous training examples. The target variable may encode the buying interest of the customer.

- **Medical Disease Diagnosis:** In recent years, the use of data mining methods in medical technology has gained increasing traction. The features may be extracted from the medical records, and the class labels correspond to whether or not a patient may pick up a disease in the future. In these cases, it is desirable to make disease predictions with the use of such information.

- **Supervised Event Detection:** In many temporal scenarios, class labels may be associated with time stamps corresponding to unusual events. For example, an intrusion activity may be represented as a class label. In such cases, time-series classification methods can be very useful.

- **Multimedia Data Analysis:** It is often desirable to perform classification of large volumes of multimedia data such as photos, videos, audio or other more complex multimedia data. Multimedia data analysis can often be challenging, because of the complexity of the underlying feature space and the semantic gap between the feature values and corresponding inferences.

- **Biological Data Analysis:** Biological data is often represented as discrete sequences, in which it is desirable to predict the properties of particular sequences. In some cases, the biological data is also expressed in the form of networks. Therefore, classification methods can be applied in a variety of different ways in this scenario.

- **Document Categorization and Filtering:** Many applications, such as newswire services, require the classification of large numbers of documents in real time. This application is referred to as document categorization, and is an important area of research in its own right.

- **Social Network Analysis:** Many forms of social network analysis, such as collective classification, associate labels with the underlying nodes. These are then used in order to predict the labels of other nodes. Such applications are very useful for predicting useful properties of actors in a social network.

3. Features

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 to 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 use. However, due to the variety of uncontrolled noises 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,

N-gram representation is another method that has been 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 five 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 between them, while a 2-gram representation can. However, the 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 were used: title, headings, metadata, and main text. They showed that the best result was 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.

3.1.2. Visual Analysis.

Each Web page has two representations, if not more. One is the text representation written in HTML. The other 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 results of visual analysis, heuristic rules are applied to recognize multiple logical areas, which correspond to different meaningful parts of the page. The authors 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, they used the histogram of the images on the page 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 et al. 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.

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. In such cases, it is difficult for classifiers to make reasonable judgments based on the 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.

3.2.3. Neighbor Selection.

Another question when using features from neighbors is that of which neighbors to examine. Existing research has mainly focused 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 3. 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.

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. 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.

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 et al. 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 Shen et al. termed implicit links. Thus, they utilized 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.

4. ALGORITHMS

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].Besides these simple measures, there have been a number of feature selection approaches developed in text categorization, such as information gain, mutual information, document frequency, and the $\chi 2$ test.

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 in which the term appears, that is, 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, based on existing research, it is not clear to what extent feature selection and feature weighting contributed to the improvement. Yan et al. [2005] proposed a novel feature selection approach which is more efficient and effective than information gain and $\chi 2$ test on large-scale datasets. In text categorization, there is a class of problems in which categories can be distinguished by a small number of features while a large number of other features only add little additional differentiation power.

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 works well in Web classification. Relaxation labeling was originally proposed as a procedure in image analysis [Rosen-feld et al. 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], pages 190–191).Relaxation labeling is effective in Web page classification [Chakrabarti et al. 1998; Luand 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 was better than textual classifiers. Macskassy and Provost [2007] implemented a toolkit for classifying networked data, which utilized a collective inference procedure [Jensen et al. 2004], and demonstrated its powerful performance on several datasets including Web collections. Unlike others, Zhang et al. [2006] proposed a novel approach to relational learning based on both local text and link graph, and showed improved accuracy

4.3. Hierarchical Classification

Most existing Web classification approaches focus on classifying instances into a set of categories on a single 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. Their work demonstrated that splitting the classification problem into a number of subproblems at each level of the hierarchy is more efficient and accurate than classifying in a nonhierarchical fashion. 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.

5. 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 tasks include assigning documents to categories on the basis of subject, function, sentiment, genre, and more. Unlike more general text classification, Web page classification methods can take advantage of the semistructured content and connections to other pages within the Web. 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 (cocited pages) over other types of neighbors;

—incorporate anchor text 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.

6. References

[1]Singh, Brijendra, and Hemant Kumar Singh."Web data mining research: A survey."Computational Intelligence and

Computing Research (ICCIC), 2010 IEEE International Conference on. IEEE, 2010.

[2]R. Malarvizhi and K. Saraswathi.“Web Content Mining Techniques Tools & Algorithms -A Comprehensive Study."International Journal of Computer Trends and Technology (IJCTT), Volume 4, 2013.

[3]Deepti Sharda and Sonal Chawla."Web Content Mining Techniques: A Study."International Journal of Innovative Research in Technology & Science.

[4]Johnson, Faustina, and Santosh Kumar Gupta."Web Content Mining Techniques: A Survey."International Journal of Computer Applications (0975–888) Volume (2012).

[5]Sharma, Arvind Kumar, and P. C. Gupta."Study & Analysis of Web Content Mining Tools to Improve Techniques of WebData Mining."International Journal of Advanced Research in ComputerEngineering & Technology (IJARCET) Volume 1 (2012)

[6] Asirvatham,A.P.Andrav I, K. K. 2001. Web page classification based on document structure. Awarded second prize in National Level Student Paper Contest conducted by IEEE India Council.

[7] Chakrabarti, S. 2000. Data mining for hypertext: A tutorial survey.SIGKDD Explorat. Newsl. 1,2 (Jan.),1–11.

[8]Chakrabarti, S. 2003.Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann,San Francisco, CA.

[9] Huang, C.-C., Chuang, S.-L.,Andchien, L.-F. 2004b. Using a Web-based categorization approach to generate thematic metadata from texts.ACM Trans. Asian Lang. Inform. Process. 3,3, 190–212.

[10]J̈Aschke, R., Marinho, L. B., Hotho, A., Schmidt-Thieme, L.,Andstumme, G. 2007. Tag recommendationsin folksonomies. InProceedings of Knowledge Discovery in Databases: 11th European Conference onPrinciples and Practice of Knowledge Discovery in Databases (PKDD)

[11]J. N. Kok, J. Koronacki, R. L.de Mntaras, S. Matwin, D. Mladenic, and A. Skowron, Eds. Lecture Notes in Computer Science, vol.4702. Springer, Berlin, Germany, 506–514.