Proceedings of GLOGIFT 07
November 15-17, 2007
UP Technical University
Noida, pp. 611-619

A SURVEY ON WEB PAGE CLASSIFICATION TECHNIQUES

S. K. Chaturvedii*, D. K. Swami* and R. C. Jain*

ABSTRACT

With the huge amount of information available online, the World Wide Web is a fertile area for data mining research. The Web page classification research is at the cross road of research from several research communities, such as database, information retrieval, and within AI, especially the sub-areas of machine learning and natural language processing.. In this paper, we survey the research in the area of Web page classification. For the survey, we focus on representation issues, on the process, on the learning algorithm, and on the application of the recent works as the criteria.

Keywords: Web, Web mining, Information retrieval, classification, Information Extraction.

Introduction

The World Wide Web (Web) is a popular and interactive medium to disseminate information today. The Web is huge, diverse, and dynamic and thus raises the scalability, multimedia data, and temporal issues respectively. Due to those situations, we are currently drowning in information and facing information overload Information users could encounter, among others, the following problems when interacting with the Web

I. Finding relevant information: However today's a search tool have the following problems? People either browse or use the search service when they want to find specific information on the Web [28]. We are using different search engine like Google, yahoo, Khoj etc.When a user uses search service he or she usually inputs a simple keyword query and the query response is the list of pages ranked based on their similarity to the query. The first problem is low precision, which is due to the irrelevance of many of the search results. This results in a difficulty finding the relevant information. The second problem is low recall, which is due to the inability to index all the information available on the Web. This results in a difficulty finding the unindexed information that is relevant.

II. Creating new knowledge out of the information available on the Web: Actually this problem could be regarded as a Sub-problem of the problem above. While the problem above is usually a query-triggered process (retrieval oriented), this problem is a data-triggered process that presumes that we already have a collection of Web data and we want to extract potentially useful knowledge out of it (data mining oriented).

III. Personalization of the information: This problem is often associated with the type and presentation of information, since it is likely that people differ in the contents and presentations they prefer while interacting with the Web.

* Faculty, Department of Computer Applications, SATI Vidisha (M.P.)

IV. Learning about consumers or individual users: This is a problem that specifically deals with the problem c above, which about knows what the customers do and want. Inside this problem, there are sub-problems such as mass customizing the information to the intended consumers or even to personalize it to individual user, problems related to effective Web site design and management, problems related to marketing, etc.

Web mining techniques could be used to solve the information overload problems above directly or indirectly. However, we do not claim that Web mining techniques are the only tools to solve those problems. Other techniques and works from different research areas, such as database (DB), information retrieval (IR), natural language processing (NLP), and the Web document community, could also be used. By the direct approach we mean that the application of the Web mining techniques directly addresses the above problems. For example, a Newsgroup agent that classifies whether the news is relevant to the user. By the indirect approach we mean that the Web mining techniques are used as a part of a bigger application that addresses

	Web mining Tasks
(a)	Resource finding
(b)	Information selection & pre-processing
(c)	Generalization
(d)	Analysis

the above problems. For example, Web mining techniques could be used to create index terms for the Web search services.

Web mining performs following subtasks, namely

1. *Resource finding:* the task of retrieving intended Web documents.
2. *Information selection and pre-processing:* automatically selecting and pre-processing specific information from retrieved Web resources.
3. *Generalization:* automatically discovers general patterns at individual Web sites as well as across multiple sites.
4. *Analysis:* validation and/or interpretation of the mined patterns.

Web Mining and Information Retrieval

Actually IR is the automatic retrieval of all relevant documents while at the same time retrieving as few of the non-relevant as possible. [11]IR has the primary goals of indexing text and searching for useful documents in a collection and nowadays research in IR includes modeling, document classification and categorization, user interfaces, data visualization, filtering, etc.

Web Mining and Information Extraction

IE has the goal of transforming a collection of documents, usually with the help of an IR system, into information that is more readily digested and analyzed. IE aims to extract relevant facts from the documents while IR aims to select relevant documents. While IE is interested in the structure or representation of a document, IR views the text in a document just as a bag of unordered words. Thus, in general IE works at a finer granularity level than IR does on the documents. However, the differences between the two become blurred if the interest of IR is in extraction and when used in the context of vague forms of information in which a full text IR

system can provide some IE features.

Definition

The web is a huge repository of information and there is a need for categorizing web documents to facilitate the search and retrieval of pages. These algorithms rely solely on the text content of the web pages for classification. However, the web has a lot of information contained in structure, images, video etc present in the document. We propose a survey on classification of web pages into a few broad categories based on the structure of the web document. Earlier, domain experts did this classification manually. But very soon, the classification had to be done semi-automatically or automatically. Some of the approaches according to [25]; are web page classification based statistical and machine-learning algorithms like K-Nearest Neighbor approach, Bayesians probabilistic models, inductive rule learning, support vector mechanics, neural networks and decision trees. Very few learning methods exploit the hierarchical structure and an effort was made by [29] to classify web content based on hierarchical structure for classification.

Besides the text content of the web page, the images, video and other multimedia content and the structure of the document also provide a lot of information aiding in the classification of a page. A human mind categorizes pages into a few broad categories at first sight without knowing the exact content of the page. It uses other features like the structure of the page, the images, links contained in the page, their placement etc.

Different Categories for Web page Classification

Several attempts have been made to categorize the web pages with varying degree of success. The major classifications can be classified into the following broad categories

1. Manual classification by domain specific experts.
2. Clustering approaches.
3. META tags
4. A collection of document content (like images, multimedia data etc.)
5. Solely on web page content analysis and hyperlink Analysis.

Manual Classification

The traditional manual approach to classification would involve the analysis of the contents of the web page by a number of domain experts and classification based on the textual content as by Google. The real data on the web rules out this approach. Moreover, such a classification would be subject specific and hence open to question. This resulted in efforts to automate the entire classification process. However, this is based on a number of positive and negative training sets, for which again a number of domain experts are required.

Clustering Approaches

Clustering algorithms have been used widely as the clusters can be formed directly without any background information. However, most of the clustering algorithms like K-Means etc. require the number of clusters to be specified in advance.

META Tags

These classification techniques solely rely on content attributes of the <META name=”Keywords”> and <META name=”description”> tags. Though relying on these tags might give accurate results to a large extent, there is a possibility of the web page author to include keywords that don’t reflect the content of the page, just to increase the hit-rate of his page in

search engine results. So, some search engines that relied on this method failed to appropriately classify the web documents.

The fourth and fifth approaches use the text content of the web page for classification. In text-based approaches, first a database of keywords in a category is prepared as follows. The frequency of the occurrence of words, phrases etc in a category is computed from an existing corpus (a large amount of text). The commonly occurring words (called stop words) are removed from this list. The remaining words are the keywords for that particular category and can be used for classification. To classify a document, all the stop words are removed and the remaining keywords/phrases are represented in the form of a feature vector.

Structure Based Approach

Structure based approach relies on the fact that there are many other features apart from text content which form the whole gamut of information present in a web document. Structure based approach tends to exploit this fact. Web pages belonging to a particular category have some similarity in their structure. Based on these similarities, any web page can be categorized into at least three broad categories:

Web pages Categories		
(a) Information Pages	(b) Research Pages	(c) Personal Home Pages

A typical information page has a logo on the top followed by a navigation bar linking the page to other important pages. The amount of information available on the web is huge and growing each year. At present Google searches more than 4.2 billion pages. As the web has grown, the ability to mine for specific information has become almost important as the web itself. Data mining consists of a set of techniques used to find useful patterns within a set of data and to express these patterns in a way which can be used for intelligent decision making. In this project the knowledge is represented as classification rules. A rule consists of an antecedent (a set of attribute values) and a consequent (class): IF <attrib = value> AND ... AND <attrib = value> THEN <class>. The class part of the rule (consequent) is the class predicted by the rule for the records where the predictor attributes hold. An example rule might be IF <Salary = high> AND <Mort gate = No> THEN <Good Credit>. This kind of knowledge representation has the advantage of being intuitively comprehensible to the user. This is important, because the general goal of data mining is to discover knowledge that is not only accurate, but also comprehensible to the user. In the classification task, the goal is to discover rules from a set of training data and apply those rules to a set of test data (unseen during training), and hopefully predict the correct class in the test set. In this project, the goal is to discover a good set of classification rules to classify web pages based on their subject.

Web page Classification Algorithms

Web-page Classification through Summarization

Web-page classification can borrow directly from the machine learning literature for text classification. Web pages have their own underlying embedded structure in the HTML language. Authors in [32] quote that web pages typically contain noisy content such as advertisement banner and navigation bar. If a pure-text classification method is directly applied to these pages, it will incur much bias for the classification algorithm, making it possible to lose focus on the main topics and important content. Thus, a critical issue is to design an intelligent preprocessing technique to extract the main topic of a Web page. Web-page summarization

techniques for preprocessing in Web-page classification is a viable and effective technique. Authors in [32] show that instead of using an off-the-shelf summarization technique that is designed for pure-text summarization, it is possible to design specialized summarization methods catering to Web-page structures. In order to collect the empirical evidence that summarization techniques can benefit Web classification, they first conduct an ideal case experiment, in which each Web page is substituted by its summary generated by human editors. Compared to using the full-text of the Web pages, we gain an impressive 14.8% improvement in Classification's measurement. They also propose a new automatic Web summarization algorithm, which extracts the main topic of a Web page by a page-layout analysis to enhance the accuracy of classification. [32]. they evaluate the classification performance with this algorithm and compare to some traditional state-of-the-art automatic text summarization algorithms which including supervised methods and unsupervised learning methods. Experiment results on Look Smart Web directory show that all summarization methods can improve the other measurement results. Finally, they show that an approach of summarization methods can achieve about 12.9% improvement relatively on other measurement results, which is very close to the upper bound achieved in our ideal case experiment. They will consider four different methods for conducting the Web page summarization. The first method uses to an adaptation of Luhn's summarization technique. The second method uses to using Latent Semantic Analysis on Web pages for summarization. The third method uses to finding the important content body as a basic summarization component. Finally, the fourth method looks at summarization as a supervised learning task. They gather the results of all four summarization methods into an ensemble of summarizers, and use it for Web page summarization.

Web Page Classification Based on Fuzzy Association

Web mining is discovering and collecting of useful information from WWW. Some of Web mining techniques include analysis of user access patterns, Web document clustering, and classification. Document classification or text categorization (as used in information retrieval context) is the process of assigning a document to a predefined set of categories based on the document content. Document classification can be applied as an information filtering tool and can also be used to improve the retrieval results from a query process. To help the users search and browse for specific information on the Web, many of the well-known Web portals such as *Google!* have organized the information, in form of Web documents, into some predefined categories such as *Arts & Humanities*, *Computers & Internet*, and *Entertainment*. A method of automatically classifying Web documents into a set of categories using the *fuzzy association* concept is proposed.[12] The fuzzy association uses the concept of the *Fuzzy Set* theory to model the vagueness in the information retrieval process. Examples of the research works involving the use of the fuzzy association technique include. The basic concept of fuzzy association involves the construction of a *pseudo thesaurus* of keywords or index terms from a set of documents. [12]By constructing a *pseudo thesaurus*, the relationship among different index terms or keywords in the documents is captured, i.e., each pair of words has an associated value to distinguish itself from other pairs of words. Therefore, the ambiguity in word usage is minimized.

Web Page Classification with an Ant Colony Algorithm

In Ant Colony Optimization (ACO) algorithm for discovering classification rules. Investigating the use of Ant-Miner in web mining is an important research direction, as follows. [1]First, an empirical comparison between Ant- Miner and two very popular rule induction algorithms (C4.5 and CN2), across six data sets, has shown that Ant-Miner is not only competitive with respect to predictive accuracy, but also tends to discover much simpler rules However, that comparison

involved only "conventional" data mining – i.e., mining structured data sets. Web mining is more challenging, because it involves unstructured or semi-structured text found in web pages. there are a potentially very large number of attributes (words) associated with web pages, and a theoretical analysis of Ant-Miner (under very pessimistic assumptions) shows that its computational time is quite sensitive to the number of attributes Hence, it is important to understand how scalable Ant- Miner is to data sets with a large number of attributes in practice, in a challenge real world domain such as web mining. In nature ants are seen creating "highways" to and from their food, often using the shortest route. Each ant lays down an amount of pheromone and the other ants are attracted to the strongest scent. As a result, ants tend to converge to the shortest path.

This is because a shorter path is faster to transverse, so if an equal amount of ants follow the long path and the short path, the ants that follow the short path will make more trips to the food and back to the colony. If the ants make more trips when following the shorter path, then they will deposit more pheromone over a given distance when compared to the longer path. This is a type of positive feedback and the ants following the longer path will be more likely to change to follow the shorter path, where scent from the pheromone is stronger.

Webpage Classification Using URL Features

In Current webpage classification techniques use a variety of information to classify a target page: the text of the page itself, its hyperlink structure, the link structure and anchor text from pages pointing to the target page and its location (given by its URL)[11]. Of this information, a web page's uniform resource locator (URL) is the least expensive to obtain and one of the more informative sources with respect to classification. Past systems have incorporated URL features into machine learning frameworks before. URLs are often meant to be easily recalled by humans, and websites that follow good design techniques will encode useful words that describe their resource in the website's domain name as advocated by best practice guidelines. Websites that present a large amount of expository information often break their contents into a hierarchy of pages on subtopics. This information structuring for the web often is mirrored in their URLs as well. As the URL is short, ubiquitous (all web pages, whether or not they are accessible or even exist, have URLs) and is largely content-bearing, it seems logical to expend more effort in making full use of this resource. They approach this problem by considering a classifier that is restricted to using the URL as the sole source of input. Such a classifier is of interest as it would be magnitudes faster than traditional approaches as it does not require pages to be fetched or the full text or links to be analyzed. We have implemented such a classifier which uses a two-step machine learning approach. A URL is first segmented into meaningful tokens using information-theoretic measures. This is necessary as some components of a URL are not delimited by spaces (especially domain names). These tokens are then fed into an analysis module that derives useful composite features for classification. These features model sequential dependencies between tokens, their orthographic patterns, length, and originating URI component. In the second step, machine learning is used to induce a multiclass or regression model from labeled training URLs that have been processed by the above pipeline. New, unseen test URLs can then be classified by processing them first to extract features, and then applying the derived model to obtain a final classification. A key result is that the combination of quality URL segmentation and feature extraction results in a significant improvement in classification accuracy over baseline approaches.

Hierarchical Classification of Web Content

Authors in [32] make use of automatic classification methods to supplement human effort in creating structured knowledge hierarchies. Although many real world classification systems

have complex hierarchical structure (e.g., MeSH, U.S. Patents, Yahoo!, Look Smart), few learning methods capitalize on this structure. Most of the approaches mentioned above ignore hierarchical structure and treat each category or class separately, thus in effect "flattening" the class structure. [12]A separate binary classifier is learned to distinguish each class from all other classes. The binary classifiers can be considered independently, so an item may fall into none, one, or more than one category. Or they can be considered as an m-ary problem, where the best matching category is chosen. The hierarchical structure can also be used to set the negative set for discriminative training and at classification time to combine information from different levels. First, they test the approach on a large collection of very heterogeneous web content, which we believe is increasingly characteristic of information organization problems. Second, they use a learning model, support vector machine (SVM) that has not previously been explored in the context of hierarchical classification. SVMs have been found to be more accurate for text classification than popular approaches like naïve Bayes, neural nets, and Rocchio We use a reduced-dimension binaryfeature version of the SVM model that is very efficient for both initial learning and real-time classification, thus making it applicable to large dynamic collections.

Heterogeneous Learner for Web Page Classification

Classification of an interesting class of Web pages (e.g., personal homepages, resume pages) has been an interesting problem. Typical machine learning algorithms for this problem require two classes of data for training: positive and negative training examples. However, in application to Web page classification, gathering an unbiased sample of negative examples appears to be difficult. They propose a heterogeneous learning framework for classifying Web pages, which (1) eliminates the need for negative training data, and (2) increases classification accuracy by using two heterogeneous learners. This framework uses two heterogeneous learners – a decision list and a linear separator which complement each other – to eliminate the need for negative training data in the training phase and to increase the accuracy in the testing phase. They present here a new machine learning framework that exactly matches these problems of Web page classification. There have been many attempts to use multiple homogeneous learners to increase classification accuracy. However, combination of homogeneous learners generally does not overcome the genuine weakness of each learner. The purpose of the decision list in training phase is to eliminate the need for negative training data in constructing a linear separator. The decision list in testing phase enhances the accuracy of the linear separator especially for low-margin data. As a result, our heterogeneous framework (1) makes easier to create a classifier for a new concept by reducing the work to collect training documents, and also (2) increases the final classification accuracy by complementing the weakness of linear separators for low-margin data. The contributions of our framework are the following. Our heterogeneous framework enables *pre-filtering* stage in training phase to induce negative training data from universe and positive training data. Previous machine learning schemes need to classify large number of pages manually to prepare *unbiased* positive and negative training documents. The pre-filtering stage makes possible to construct a classifier without requiring negative training data, which speeds up the process of creating a classifier for a new class, but also opens a possible way to support type-specific queries on the Internet from sample pages. We propose a new *early-inclusion* stage for correctly classifying low-margin data. Linear separators such as Winnow, Perceptron, and SVMs have been studied extensively and have proved their outstanding performances when the environment has high dimensions, the number of active features is small, and the instance spaces are sparse. Consequently, the linear separators are the most widely used algorithms for Web page classification problems since Web page classification has the same properties as environment these linear separators work well.

Conclusion

In this paper we have surveyed state of art techniques of Web page classification. We reviewed some standard approaches to classify web pages such as summarization based classification, fuzzy association based, URL based approach and hierarchical classification and Heterogeneous Learner for Web Page Classification. The classification of a page detecting very useful approaches and it would increase the classification accuracy. Classification by URL features has advantages like real-time efficiency. An alternative approach of automatically classifying the Web documents into some predefined categories using the fuzzy association .Several approaches exist in the literature for web page classification. Still challenges like web images classification, applets categorization, and multimedia data classification etc exist and beckon further research in the area of we page classification.

References

- Abraham and V. Ramos. Web Usage Mining Using Artificial Ant Colony Clustering and Genetic Programming. *Proc. Congress on Evolut. Comp. (CEC-2003)*. IEEE Press, 2003.
- C. R. Anderson and E. Horvitz. Web montage: a dynamic personalized start page. In *Proceedings of the Eleventh ntl.Conference on World Wide Web*, pages 704–712. ACM Press, 2002.
- C.-N. Hsu and M.-T. Dung. Generating finite-state transducers for semi-structured data extraction from the web. Information Systems, 23(8):521–538, 1998.
- D. Hawking and N. Craswell. Overview of the TREC-2001web track. In *The Tenth Text REtrieval Conference (TREC 2001*). NIST, 2001. Special Publication 500-250.
- H. Yu, J. Han, and K. C.-C. Chang. Pebl: Positive-example based learning for web page classification using svm. In *KDD*, Edmonton, Alberta, Canada, 2002.J. Kivinen, M. K. Warmuth, and P. Auer. The Perceptron algorithm vs. Winnow: linear vs. logarithmic mistake bound when few input variables are relevant. *Artificial Intelligence*, 1-2:325–343, 1997.
- I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools with Java Implementations*, Morgan Kaufmann Publications, 2000.
- J. Nielsen and M. Tahir. *Homepage usability: 50 websites deconstructe*d. New Riders Publishing, USA, 2001.
- J. Smith and S. Chang. Multi-stage classification of images from features and related text. In *EDLOS Worksho*p, San Miniato, Italy, 1997.
- J. Yi and N. Sundaresan. A classifier for semi-structured documents. In *KDD 200*0, Boston, MA USA, 2000.
- John.M.Pierre, *Practical Issues for Automated Categorization of Web Pages*, September 2000.
- K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filterin*g, 1999.
- Kosorukoff. Genetic synthesis of cascade structures for [19] R. A. Servedio. On pac learning using winnow, perceptron, and a perceptron-like algorithm. cite- seer.nj.nec.com/329971.html.
- L. K. Shih and D. Karger. Using URLs and table layout for web classification tasks. In *Proc. of WWW '0*4, 2004.
- Lewis, D.D. and Ringuette, M. *A Classification of two learning algorithms for text categorization,* Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94), pp.81-93, 1994.
- M.-Y. Kan. Web page classification without the web page. In *Proc. of WWW '0*4, 2004. Poster paper.
- Oh-Woog Kwon, Jong-Hyoek Lee, Web page classification based on k-Nearest Neighbor approach.
- R. Agrawal and R. Srikant. On integrating catalogs. In *Proceedings of 10th Intl. Conference on the World Wide Web*, pages 603–612, Hong Kong, CN, 2001. ACM Press, New York, US.
- R. Barzilay, N. Elhadad, and K. R. McKeown. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35–55, 2002.
- R. Klivans and R. A. Servedio. Learning dnf in time citeseer.nj.nec.com/329971.html.

- R. Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *Proc. of 6th Conf. on Natural Language Learnin*g, pages 49-55, 2002.
- R.S. Parpinelli, H.S. Lopes and A.A. Freitas. An Ant Colony Algorithm for Classification Rule Discovery. In: H.A. Abbass, R.A. Sarker, and C.S. Newton. (Eds.) *Data Mining: a Heuristic Approach*, pp. 191-208. London: Idea Group Publishing, 2002.
- R.S. Parpinelli, H.S. Lopes and A.A. Freitas. Data Mining with an Ant Colony Optimization Algorithm. *IEEE Trans. on Evolutionary Computation, special issue on Ant Colony algorithms*, 6(4), pp. 321-332, Aug. 2002.
- S. Chakrabarti *Mining the web: discovering knowledge from hypertext data*. Morgan Kaufmann, 2003.
- S. Slattery and M. Craven. Combining statistical and relational methods for learning in hypertext domains. In *8th Int'l Conf. on Inductive Logic Programmin*g, 1998.
- Sun, E.-P. Lim, and W.-K. Ng. Web classification using support vector machine. In *4th Int'l Workshop on Web Information and Data Management (WIDM 2002*), Virginia, USA, November 2002.
- T. Joachims, D. Freitag, and T. Mitchell.Webwatcher: A tour guide for the world wide web. In prceedings of the International Joint Conference on Artificial In- telligence IJCAI-97, pages 770–777, 1997.
- Using URLs and Table Layout for Web Classification Tasks Lawrence Kai Shih and David R.
- Web Mining Research: A Survey Raymond Kosala Department of Computer Science Katholieke Universiteit Leuven 1-15 July 2000.
- Weigend, A.S., Weiner, E.D. and Peterson, J.O., *Exploiting Hierarchy on Text Categorization,* Information Retrieval, I(3), pp.193-216, 1999.
- Y. Yang, S.Slattery, and R. Ghani. A study of approaches to hypertext categorization. *J. of Intelligent Information System*s, 18(2-3):219-241, 2002.
- Yiming Yang, Xin Lui *A Reexamination of Text Categorization methods*, In proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99), pp. 42-49, University of California, Berkeley, USA 1999.
- Web-page Classification through Summarization Dou Shen Zheng Chen Qiang Yang Hua-Jun Zeng Benyu Zhang Yuchang Lu Wei-Ying Ma USA 2004.Addesaa