Normalization Technique for Structure based Web Documents Classification using Rough Set Theory

Amit Rathore
Department of Computer Science and Engineering
SIRTE, Bhopal, India

Kamlesh Namdev
Department of Computer Science and Engineering
SIRTE, Bhopal, India

ABSTRACT

The rapid development of the internet and web publishing techniques create numerous information sources published as HTML document on World Wide Web. WWW is now a popular medium by which people all around the world can spread and gather the information of all kinds. But web document of various sites that are generated. Contain undesired information also. This information is called noisy or irrelevant content. The need for innovative and effective technologies to help find and use the useful information and knowledge from a large variety of data sources is continually increasing. Web information has become increasingly diverse. In order to utilize the Web information better, people pursue the latest technology, which can effectively organize and use online information. Classification is one of the vital and important data mining techniques that grouped various items in a collection to predefined classes or groups. The main goal of classification is to exactly predict the target class for each case in the data. Web Document Classification is technique of data mining to discover classification of Web Documents. The information providers on the web will be interested in techniques that could improve the effectiveness of the web search engine. In this paper, the relationships among the techniques used in data mining are studied. A study of web usage is also done on optimization of this web classification.

Keywords

Structured data, Unstrutured data,, Rough Set Theory, Web Document Classification.

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, to contextual advertising, and to analysis of the topical structure of the Web. Web Document 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 Document 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 Document 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.

At present, the numbers of Web-document on World Wide Web are increasing significantly. The task to find Web-document which present information satisfying our requirements by traversing hyperlinks is difficult. Therefore, we use search engines frequently on the portal site. There are two kinds of search engines. i.e., directory-style search engines such as Yahoo, and robot style ones such as goo, excite and AltaVista. The latter displays the lists of Web-documents which contain input keywords without checking themes characterizing respective Web-document. For this reason these search engines are likely to provide misdirected Web-Doc file. On the other hand, in directory-style search engines, Web-document stored in a database are classified with hierarchical categories compatible with their themes in order. This enables us to obtain Web-document including information that meets our purpose by not only following input keywords but also traversing hyperlinks classifying Web-document into categories in systematic order.

2. WEB DOCUMENT AND IT'S CLASSIFICATION

There are various kind of the classification based on the web documents some of them are as follows:-

Swarm intelligence studies the collective behavior of unsophisticated agents that interact locally through their environment. It is inspired by social insects, such as ants and termites, or other animal societies, such as fish schools and bird flocks. Although each individual has only limited capabilities, the complete swarm exhibits complex overall behavior. Therefore, the intelligent behavior can be seen as an emergent characteristic of the swarm. When focusing on ant colonies, it can be observed that ants communicate only in an indirect manner—through their environment—by depositing a substance called pheromone. Paths with higher pheromone levels will more likely be chosen and thus reinforced, while the pheromone intensity of paths that are not chosen is decreased by evaporation. This form of indirect communication is known as Stigmergy, and provides the ant colony shortest-path finding capabilities. Ant Colony Classification employs artificial ants that cooperate to find good solutions for discrete optimization problems. These software agents mimic the foraging behavior of their biological counterparts in finding the shortest-path to the food source. The first algorithm following the principles of the ACO met heuristic is the Ant System where ants iteratively construct solutions and add pheromone to the paths corresponding to these solutions. Documents 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. Documents classification can be applied as an information filtering tool and can also be used to improve the retrieval results from a query process.

3. STRUCTURED CLASSIFICATION

Web document classification has been widely studied in the past few years. Much research work has been done in this area. Chakrabarti et al. used predicted labels of neighboring documents to reinforce classification decisions for a given document. Qi and Davison summarize the various concepts used for automatic web page classification with respect to recent works. A dynamic and various leveled arrangement framework that is equipped for including new classifications as required, sorting out the website pages into a tree structure, and characterizing pages via seeking through just a single way of the tree structure is proposed in.

Structured systems are those where the activity of processing and output is predetermined and highly organized. Structured systems are designed, built and operated by the IT department. ATM transactions, airline reservations, manufacturing inventory control systems, point of sale systems are all forms of structured systems. By contrast, unstructured systems are those that have little or no predetermined form or structure. Unstructured systems include email, reports, contracts, and other communications. A person who performs a communications activity in an unstructured system has wide latitude to structure the message in whatever form is desired. The rules of unstructured systems are fewer and less complex

4. UNSTRUCTURED CLASSIFICATION

Web document classification is the process of classifying documents into predefined categories based on their content. The classifiers used for this purpose should be trained from the web documents that are already classified. The task is to assign a document to one or more classes or categories. Unstructured information documents frequently incorporate content and multimedia substance. Instances incorporate email messages, word handling records, videos, photographs, sound audio files, presentations, website pages and numerous different sorts of business archives. Take note of that while these sorts of records may have an inner structure, they are as yet thought to be "unstructured" in light of the fact that the information they contain doesn't fit conveniently in a database. Unstructured data represent around 80% of data. It frequently incorporates text and media content. Examples incorporate email messages, word processing reports, recordings, photographs, sound documents, presentations, website pages and numerous different sorts of business archives.

5. LITERATURE REVIEW

Web Document classification is significantly different from traditional text classification because of the presence of some additional information, provided by the HTML structure and by the presence of hyperlinks. In this paper we analyze these peculiarities and try to exploit them for representing Web Documents in order to improve categorization accuracy. We conduct various experiments on a corpus of 8000 documents belonging to 10 Yahoo! categories, using Kernel Perception and Naive Bays classifiers. Our experiments show the usefulness of dimensionality reduction and of a new, structure-oriented weighting technique. We also introduce a new method for representing linked document using local information that makes hypertext categorization feasible for real-time applications. Finally, we observe that the combination of the usual representation of Web Documents using local words with a hyper textual one can improve classification performance.

In this paper we use machine learning algorithms like SVM, KNN and GIS to perform a behavior comparison on the Web Documents classifications problem, from the experiment we see in the SVM with small number of negative documents to build the Centroids has the smallest storage requirement and the least on line test computation cost. But almost all GIS with different number of nearest neighbors have an even higher storage requirement and on line test computation cost than KNN. This suggests that some future work should be done to try to reduce the storage requirement and on list test cost of GIS.

The more broad issue of content characterization is past the extent of this article. Contrasted and standard content characterization, grouping of web substance is diverse in the accompanying perspectives. Initially, customary content characterization is commonly performed on "organized corpora with very much controlled writing styles", while web accumulations don't have such a property. Second, website pages are semi-organized records in HTML with the goal that they might be rendered outwardly for clients. Albeit other archive accumulations may have implanted data for rendering as well as a semi-organized arrangement, such markup is ordinarily stripped for order purposes.

6. WORK DONE

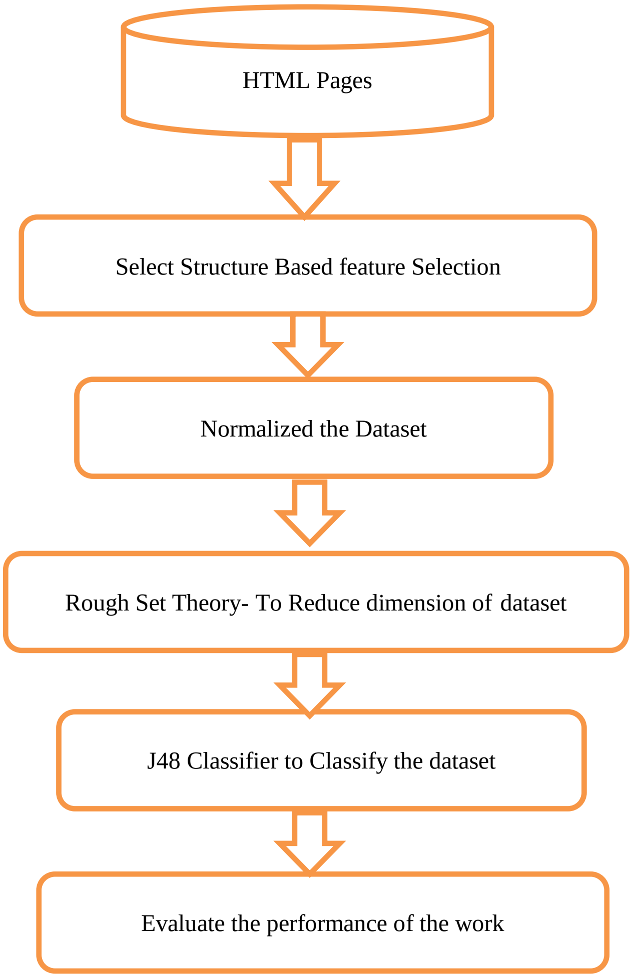
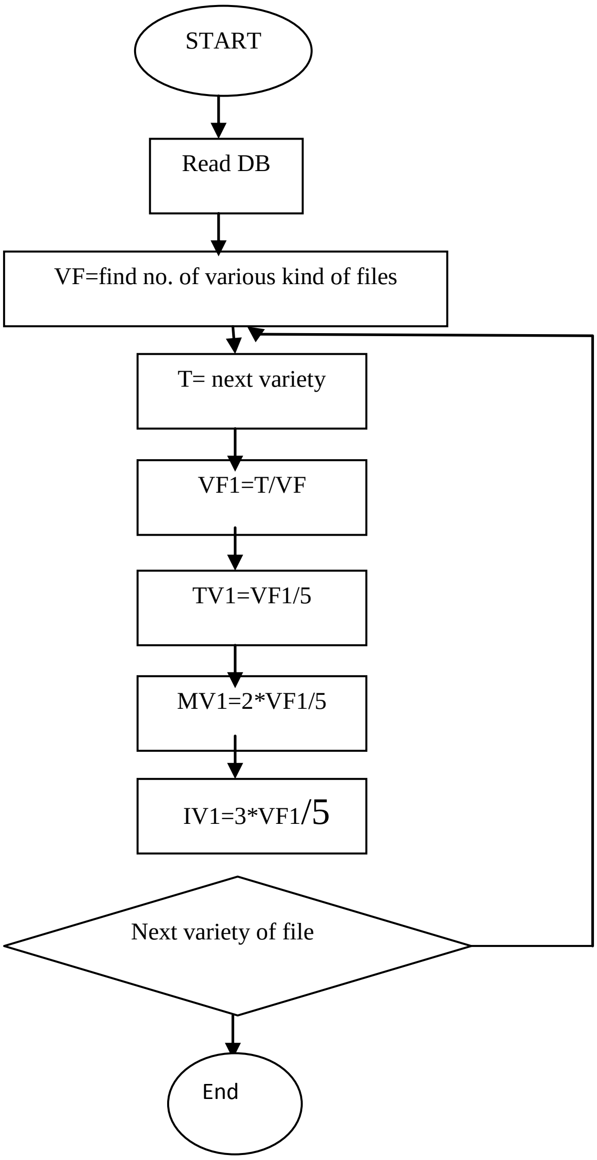

HTML Pages
Select Structure Based feature Selection
Normalized the Dataset
Rough Set Theory- To Reduce dimension of dataset
J48 Classifier to Classify the dataset
Evaluate the performance of the work

Figure 2: Flow Chart of Proposed Work

1. Start
2. Load Web Document dataset
3. Find the Entities(titles)
4. If the search successful then initialize the title with TV1
5. If no proceed to next search
6. Find Meta
7. If successful initialize the meta with MV1
8. If no search found proceed to next
9. Find Images
10. If the search successful then initialize the images with IV1
11. End

NORMALAIZATION



7. RESULT ANALYSIS

This section is dealing with the implementation portion of the proposed work. For the same, this section in divided into two parts:

a. System Configuration

The system on which these experiments are performed and verify the results:

Table I: System Configuration

Model:	**Sony Vaio**
Processor:	Intel® Core™ i5-2450M 2.5GHz
RAM:	4GB
System Type:	64 Bit Operating System
Windows Edition:	Windows 10 Home
MATLAB	R2014a

Table II shows the detail about the dataset used. According to the information, dataset is divided into four main categories. These main categories are further divided into sub-categories

Dataset Character	Dataset Category	Associated Theme	No. of associated files
A	Account	Banking	200
B	Loans	Banking	200
C	Retail Finance	Banking	200
D	JavaLessonns	Programming languages	200
E	Interpreter	Programming languages	200
F	Visual Basic	Programming languages	200
G	Galaxies	Science	200
H	Complexity	Science	200
I	Training programme	Sport	200
J	Formula One	Motor Sport	200
K	Basketball	Sport	200
		Total	**2200**

In the case of distinct categories (category A: Commercial Banks and K: Sport), utilizing only full text from documents, the average accuracy is 0.94% whereas the proposed work based on the distinct categories in case of the full text of the document the average accuracy is 0.98%

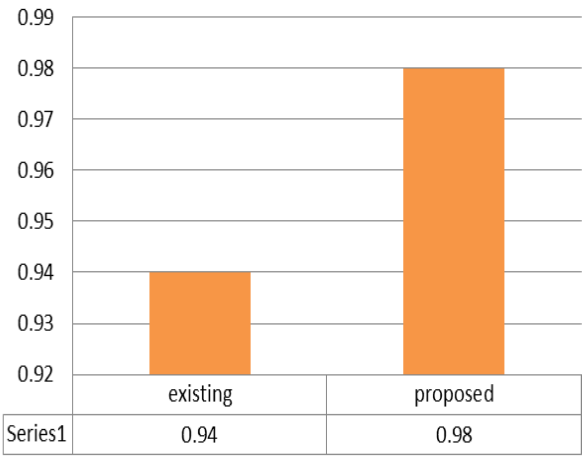
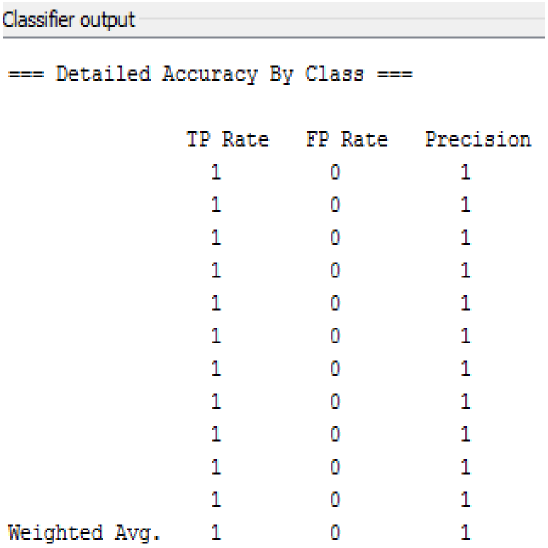


Table III gives detail about the dataset used for training and testing phase. According to this table every category of the dataset is further categories

TABLE IV: Comparison of Accuracy

Existing Work	Proposed Work
0.94	0.98



```
Classifier output

=== Detailed Accuracy By Class ===

                 TP Rate    FP Rate    Precision
                   1          0          1
                   1          0          1
                   1          0          1
                   1          0          1
                   1          0          1
                   1          0          1
                   1          0          1
                   1          0          1
                   1          0          1
                   1          0          1
                   1          0          1
Weighted Avg.      1          0          1
```

As resulted the True Positive Rate, False Positive Rate along with the Precision.

8. CONCLUSION

This paper has discussed about three areas i) Web Document classification, ii) Optimization method (which is used by the web document classification method to increase the accuracy) iii) Structure and Unstructured classification has been discussed . This article is helpful for evaders to have a deep insight in to Web Document classification. This study motivates us to do further work in the area of optimized Web Document classification with the help of other theory.

9. REFERENCES

[1] "A Review on Optimization in Web Document Classification", International Journal of Advance Foundation and Research in Computer (IJAFRC) Volume 1, Issue 9, September 2014. ISSN 2348 – 4853.

[2] XIAOGUANG QI and BRIAN D. DAVISON, "Web Document Classification: Features and Algorithms".

[3] Makoto Tsukada, Takashi Washio, Hiroshi Motoda, Mihogaoka, Ibaraki, Osaka 567-0047, JAPAN, "Automatic Web-Page Classification by Using Machine Learning Methods".

[4] E. Bonabeau, M. Dorigo, and G. Theraulaz, Swarm Intelligence: From Natural to Artificial Systems. New York: Oxford Univ.Press, 1999.

[5] M. Dorigo and T. Stützle, Ant Colony Optimization. Cambridge, MA:MIT Press, 2004.

[6] M. Dorigo, V. Maniezzo, and A. Colorni, Positive feedback as a search strategy Dipartimento di Elettronica e Informatica, Politecnico di Milano, Milano, Italy, Tech. Rep. 91016, 1991.

[7] "Ant system: Optimization by a colony of cooperating agents," IEEE Trans. Syst., Man, Cybern. Part B, vol. 26, no. 1, pp. 29–41, Feb. 1996.

[8] Yang, X. S. Nature-Inspired Metaheuristic Algorithms. Frome: Luniver, Press. (2008). ISBN 1-905986-10-6.

[9] Breiman,L.: Random Forest. Machine Learning. vol. 45, No. 1, pp.5-32 (2001).

[10] Daniele Riboni, D.S.I., University' degli Studi di Milano, Italy,"Feature Selection for Web Document Classification".

[11] International Journal of Computer Science & Information Technology (IJCSIT) Vol 4, No 5, October 2012, "Machine Learning Algorithms In Web Document Classification".

[13] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In SIGMOD '98: proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, pages 307–318, New York, NY, USA, 1998.

[14] Qi, X. and B. D. Davison (2009). "Web page classification: Features and algorithms." ACM Computing Surveys (CSUR) 41(2): Article No.: 12.

[15] 3.Xiaogang Peng, Ben Choi (2002), "Automatic Web Page Classification in a Dynamic and Hierarchical Way", In proceedings of Second IEEE International Conference on Data Mining, Washington DC, IEEE Computer Society, pp: 386-393.