ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Booster in High Dimensional Data Classification

Shruti Hiremath[1], Sheba Pari N[2], Dr. S Mohan Kumar[3],

M. Tech. Student, Dept. of CSE., New Horizon College of Engineering Bengaluru, India

Assistant Professor, Dept of CSE, New Horizon College of Engineering, Bengaluru, India

Associate Professor, Dept of CSE, New Horizon College of Engineering, Bengaluru, India

ABTRACT: Data Mining is a technique used in various domains to give mean- ing to the available data. In classification tree modeling the data is classified to make predictions about new data. Using old data to predict new data has the danger of being too fitted on the old data. But that problem can be solved by pruning methods which degeneralizes the modelled tree. This paper describes the use of classification trees and shows two methods of pruning them. An experiment has been set up using different kinds of classification tree algorithms with different pruning methods to test the performance of the algorithms and pruning methods. This paper also analyzes data set properties to find relations between them and the classification algorithms and pruning methods. Classification problems in high dimensional data with small number of observations are becoming more common especially in microarray data. During the last two decades, lots of efficient classification models and feature selection (FS) algorithms have been proposed for higher prediction accuracies. However, the result of an FS algorithm based on the prediction accuracy will be unstable over the variations in the training set, especially in high dimensional data. This paper proposes a new evaluation measure Q-statistic that incorporates the stability of the selected feature subset in addition to the prediction accuracy.

KEYWORDS: Q-static, Data Mining, Feature Selection (FS).

I. INTRODUCTION

With the development of World Wide Web, web search engines have contributed a lot in searching infonl1ation from the web. They help in finding information on the web quick and easy. But there is still room for improvement. Current web search engines do not consider specific needs of user and serve each user equally. It is difficult to let the search engine know what we the user actually want. Generic search engines are following the "one size fits all" model which is not adaptable to individual users. When different users give same query, same result will be returned by a typical search engine, no matter which user submitted the query. This might not be appropriate for users which require different information. While searching for the information from the web, users need infonl1ation based on their interest. For the same keyword two users might require different piece of information. This fact can be explained as follows: a biologist and a programmer may need information on "virus" but their fields are is entirely different. Biologist is searching for the "virus" that is a microorganism and programmer is searching for the malicious software.

II. LITERATURE SURVEY

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy n company strength. Once these things r satisfied, ten next step is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration r taken into account for developing the proposed system.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

When different users give same query, same result will be returned by a typical search engine, no matter which user submitted the query. This might not be appropriate for users which require different information. While searching for the information from the web, users need infonl1ation based on their interest. For the same keyword two users might require different piece of information. This fact can be explained as follows: a biologist and a programmer may need information on "virus" but their fields are is entirely different. Biologist is searching for the "virus" that is a microorganism and programmer is searching for the malicious software. For this type of query, a number of documents on distinct topics are returned by generic search engines. Hence it becomes difficult for the user to get the relevant content. Moreover it is also time consuming. Personalized web search is considered as a promising solution to handle these problems, since different search results can be provided depending upon the choice and information needs of users. It exploits user information and search context to learning in which sense a query refer.

III. PROPOSED SYSTEM

We propose a framework for personalized web search which considers individual's interest into mind and enhances the traditional web search by suggesting the relevant pages of his/her interest. We have proposed a simple and efficient model which ensures good suggestions as well as promises for effective and relevant information retrieval. In addition to this, we have implemented the proposed framework for suggesting relevant web pages to the user. Framework for Personalized search engine consists of user modeling based on user past browsing history or application he/she is using etc. And then use this context to make the web search more personalized. This section presents different approaches and the related work done in the field of Personalized Web search.

Advantages

Personalized web search is considered as a promising solution to handle these problems, since different search results can be provided depending upon the choice and information needs of users. It exploits user information and search context to learning in which sense a query refer.

IV. MODULE DESCRIPTION

A. Customized Search Module:

Customized Search which considers individual's interest into mind and enhances the traditional web search by suggesting the relevant pages of his/her interest. We have proposed a simple and efficient model which ensures good suggestions as well as promises for effective and relevant information retrieval. In addition to this, we have implemented the proposed framework for suggesting relevant web pages to the user.

B. User Modeling Module:

Our system considers user's profile (based on user's weblog/navigation browsing history) and Domain Knowledge in order to perform Customized Search. Using a Domain Knowledge, the system stores information about different domain/categories. Information obtained from User Profile is classified into these specified categories. The learning agent learns user's choice automatically through the analysis of user navigation/browsing history, and creates/updates enhanced User Profile conditioning to the user's most recent choice. Once the user inputs query, the system provides good suggestions for Customized Search based on enhanced user profile. Further our model makes good use of the advantages of popular search engines, as it can re-rank the
results obtained by the search engine based on the enhanced user profile.

C. Domain Knowledge ModelingModule:

Domain knowledge is the background knowledge that we used to enhance the user profile. The source which we have used for preparing Domain Knowledge is DMOZ directory. For preparing Domain Knowledge, first we have crawled the Web pages from DMOZ directory for some specified categories, where each category is represented by collection of URL's present in that category.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

D. Enhanced User Profile Module:

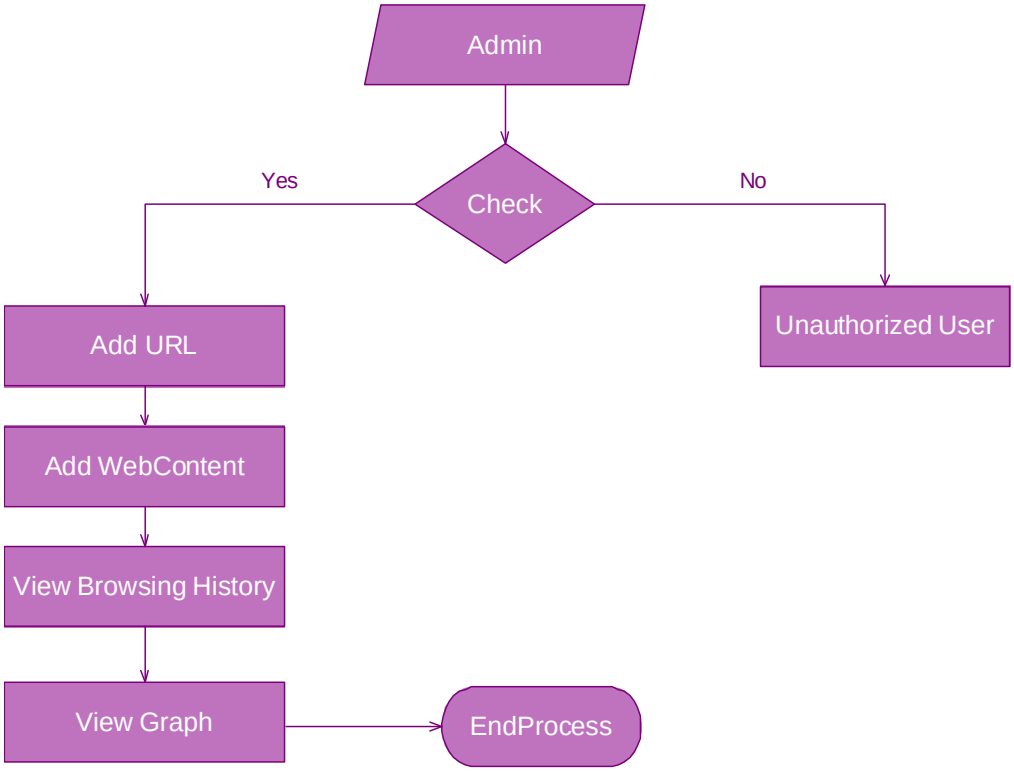
Using the information of user browsing history and domain knowledge, we create an Enhanced User Profile. Once the Enhanced User Profile is created, we take the user query and suggest the relevant web pages with respect the query. In our Experiment, we have used User Profile as a base case for suggesting the relevant pages and compared the results with the pages suggested from Enhanced User Profile. For each query, we suggest top 20 relevant documents from User Profile and for the same query we also suggest top 20 relevant documents from Enhanced User Profile. In order to compare the efficiency of the result, we compared the similarity of suggested documents with the user query.

V. SYSTEM DESIGN

Data Flow Diagram / Use Case Diagram / Flow Diagram

The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of the input data to the system, various processing carried out on these data, and the output data is generated by the system.

Data Flow Diagram



VI. RESULTS

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

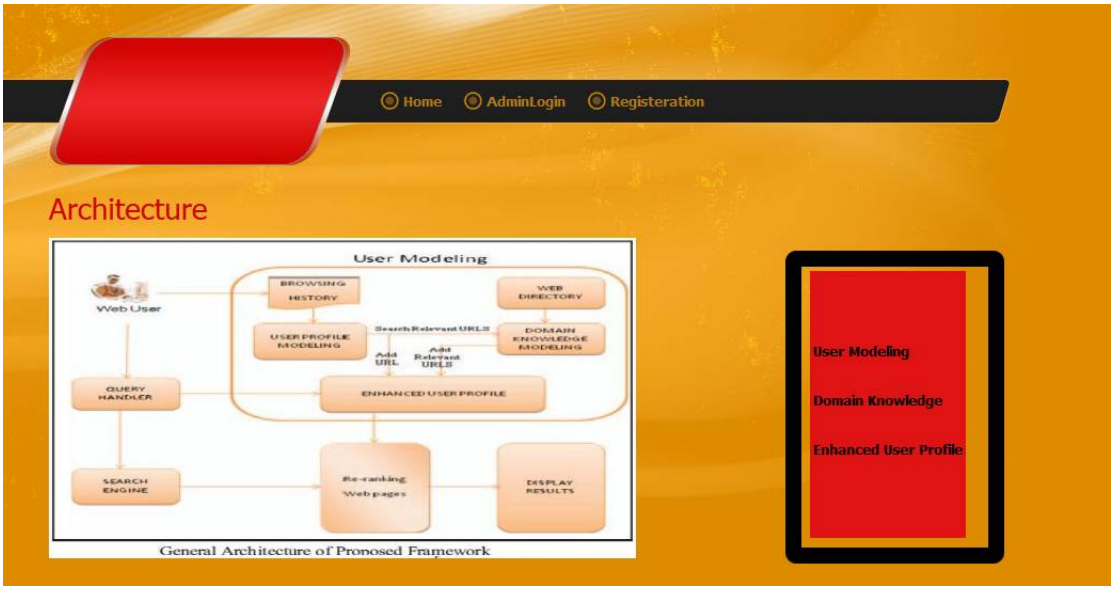
(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy.

Admin login



The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

User Search

This is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

Domain Knowledge matrix

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

VII. OBJECTIVES

A. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

B. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

C. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user
will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow.

VIII. CONCLUSION

In this paper, we have proposed a framework for personalized web search using User Profile and Domain Knowledge. Based on the User Profile and the Domain Knowledge, the system keeps on updating the user profile and thus builds an enhanced user profile. This Enhanced user profile is then used for suggesting relevant web pages to the user. The

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

proposed framework has been implemented by performing some experiments. These experiments shows that the performance of the system using enhanced user profile is better than those which are obtained through the simple user profile. Our work is significant as it improves the overall search efficiency, catering to the personal interest of the user's. Thus, it may be a small step in the field of personalized web search. In future this framework may be applied for re-ranking the web pages retrieved by search engines on the basis of user priorities. We may also apply collaborative filtering for personalized web search in our framework.

REFFERENCES

[1] T. Abeel, T. Helleputte, Y. V. de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," Bioinformatics, vol. 26, no. 3, pp. 392-398, 2010.
[2] D. Aha, and D. Kibler, "Instance-based learning algorithms," Machine Learning, vol. 6, no. 1, pp. 37-66, 1991.
[3] S. Alelyan, "On Feature Selection Stability: A Data Perspective," PhD dissertation, Arizona State University, 2013.
[4] A.A. Alizadeh, et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," Nature, vol. 403, no. 6769, pp. 503-511, 2000.
[5] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," Proc. of the National Academy of Sciences, vol. 96, no. 12, pp. 6745-6750, 1999.
[6] F. Alonso-Atienza, and J.L. Rojo-Alvare, et al., "Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection," Expert Systems with Applications, vol. 39, no.2, pp. 1956-1967, 2012.
[7] P.J. Bickel, and E. Levina, "Some theory for Fisher's linear discriminant function, naive Bayes, and some alternatives when there are many more variables than observations," Bernoulli, vol.
10, no. 6, pp. 989-1010, 2004.
[8] Z.I. Botev, J.F. Grotowski, and D.P. Kroese, "Kernel density estimation via diffusion," The Annals of Statistics, vol. 38, no. 5, pp. 2916-2957, 2010.

BIOGRAPHY

Ms Shruti Hiremath. Mtech Student, Computer Science and Engineering in New Horizon College of Engineering, which is located in Outer Ring Road, Panathur Post, Kadubisanahalli, Bangalore – 560087.

Ms Sheba Pari N. Assistant professor, Dept. of Computer Science and Engineering in New Horizon College of Engineering, which is located in Outer Ring Road, Panathur Post, Kadubisanahalli,Bangalore – 560087.

Dr. S Mohan Kumar. Associate professor, Dept. of Computer Science and Engineering in New Horizon College of Engineering, which is located in Outer Ring Road, Panathur Post, Kadubisanahalli,Bangalore – 560087.