Web Crawl Detection and Analysis of Semantic Data

AbhishekYadav[#1], Piyush Singh[*2]

RKDF institute of science and technology bhopal

***Abstract*— Web mining can be defined as mining of the WWW to retrieve useful knowledge and data about user behavior, user query, content and structure of the web. In this paper, aim on processing of structured and unstructured data mining will take place. With a tremendous development growth in website, web portal to provide downloadable data to user, required a lead to demand of a specific strategy to provide knowledgeable data to user and also useful to predict otherwise uncertain user behavior on the server. Semantic web is about machine-understandable web pages to make the web more intelligent and able to provide useful services to the users. In this paper we propose agent based Semantic Web Mining System (SWMS). It will provide classification and clustering of the web contents according to user navigating links and time when navigating to other pages, thereby facilitating knowledge based response to the user and will highlight otherwise unnoticed patterns. It mainly comprises of Interface Agents, collection Agent supported with ontology database, content mining agent and clustering agent. Content mining agent works in collaboration with descriptive metadata agent and semantic metadata agent.**

***Keywords*—– Sementic web mining, Resource Description Framework**

I. INTRODUCTION

Mining of the web to retrieve useful knowledge and data about user behavior, user query, content and structure of the web. In this paper, focus on processing of structured and unstructured data mining will take place. With tremendous development growth in website, web portal to provide downloadable data to user, required a lead to demand of a specific strategy to provide knowledgeable data to user and also useful to predict otherwise uncertain user behavior on the server. Semantic web is about machine-understandable web pages to make the web more intelligent and able to provide useful services to the users. This means that information on web pages may have to be mined so that the machine can understand the contents.

Data Mining from Semantic Web Data

Edwards *et al.* [Edwards et al., 2002] present an empirical investigation of learning from the Semantic Web, where they apply different machine learning methods to a typical user-profiling problem. The goal of their experiments is to learn a model which could then be used to recommend products to a user according to his profile. The authors test different datasets and compare the performance of learning from plain text format with learning from semantic meta-data. For the first experiment, they use traditional statistical machine learning methods. The results are not very promising, showing that the learning from semantically annotated data is not able to outperform the learning from plain text for that particular experiment. For the second experiment they apply the Prego Inductive Logic Programming (ILP) system, which is able to learn from supplied example instances and supporting background information. The results indicate some improvements: the algorithm is able to find a couple of reasonable rules for the classification task. They conclude that the Semantic Web mark-up available at that time cannot be expected to outperform conventional machine learning applied to plain text, with regards to the accuracy of the learned model. Our work extends this evaluation by looking at new statistical approaches appropriate for Semantic Web data and ontological support.

Characteristics of Semantic Web Data

The Semantic Web enhances the traditional web by adding a semantic layer on top of the well-known web data formats to make the web machine readable. In this section we introduce the basic principles and characteristics of Semantic Web data, which will be necessary for the understanding of the remainder of this thesis.

Structure

As the corner stone for describing data in such a manner, the Resource Description Framework (RDF) has been generated.

The RDF Specification provides the following definition: 1"RDF is based on the idea of identifying things using Web identifiers (called Uniform Resource Identifiers, or URIs), and defining resources in terms of simple properties and property values". RDF can be defined by its graph data model which states that the underlying structure of an RDF expression is a collection of triples, each consisting of a subject, a predicate and an object.

Semantics

The OWL Web Ontology Language [Mcguinness and van Harmelen, 2004] allows an even greater machine interpretability of the web by providing additional vocabulary and formal semantics to make the data more expressive. It serves as a standard language to define the terms in vocabularies and the relationships between those terms. Opposed to databases, ontologies serve as conceptual structures to describe the entire application domain, instead of just describing one specific application.

Querying

A query mainly consists of the following parts: the prologue (line 1), which contains the definition of namespace prefix bindings. This allows a user to write the prefix inside a query instead of rewriting the whole URI again.

Web Mining

The purpose of Web mining is to develop methods and systems for discovering models of objects and processes on the World Wide Web and for web-based systems that show adaptive performance. Web Mining integrates three parent areas: Data Mining (we use this term here also for the closely related areas of Machine Learning and Knowledge Discovery), Internet technology and World Wide Web, and for the more recent Semantic Web. The World Wide Web has made an enormous amount of information electronically accessible. The use of email, news and mark-up languages like HTML allows users to publish and read documents at a world-wide scale and to communicate via chat connections, including information in the form of images and voice records. The HTTP protocol that enables access to documents over the network via Web browsers created an immense improvement in communication and access to information. For some years these possibilities were used mostly in the scientific world but recent years have seen an immense growth in popularity, supported by the wide availability of computers and broadband communication.

II. LITERATURE REVIEW

Sharma et. al [1], Kosala et. al [3] and Eirinaki et. al [4] provided detailed review on web mining focusing on different dimensions of this field. [1] Highlighted use of cloud computing in web mining, [3] focused on scope of agent technology in it whereas [4] provided details on web personalization through web mining. Bhatia et. al in [2] provided semantic web mining and suggested an ontology learning mechanism for the extraction of semantics through grammatical rule extraction technique. Meironget.al in [5] proposed an agent based web mining model for e-business. Zhan et. al in [8] provided a multi-agent module working as knowledge crawler. Ting H.I. in [6] employed web mining for on-line social network analysis, however strategy for selecting appropriate sample size to reflect exact real social networks and actual implementation is left as future research. Jichenget.al in [7] proposed an agent based web text mining system for mining HTML based documents on the web, however it still lacks efficient algorithm for very large document collections and use of XML specifications.

Critical review of literature highlights this fact that agent technology has widely been employed in semantic web applications at various fronts and researchers have agreed on its applicability for mining semantic web contents. Although some efforts had already been made to propose application specific agent based solution in diverse areas like e-business [5] or for social networking [6], but there is no standard framework for semantic web content mining. Thus, there is scope of research in this direction. Upcoming section elaborates our proposed framework. Singh et. Al[15] proposed the next agent based web mining but there is a scope to research in content file in contrast of unstructured data mining with concept of web. Multimedia mining already included in agent based web mining al[15] but the user timing log mining and file size mining can provide a better way to meet the requirements. Literature review highlighted the fact that agent based systems have already been employed in various area of semantic web due to their promising features. Dimouet. al. [9] developed an agent based framework called Biospider for developing and testing autonomous, intelligent& semantically focused web spiders. The framework takes the advantage of agent technology in distributing crawling load to a number of cooperating spiders.

III RELATED WORK

Clustering analysis is a widely used data mining algorithm for many data management applications. Clustering is a process of partitioning a set of data objects into a number of object clusters, where each data object shares the high similarity with the other objects within the same cluster but is quite dissimilar to objects in other clusters. Different from classification algorithm that assigns a set of data objects with various labels previously defined via a supervised learning process, clustering analysis is to partition data objects objectively based on measuring the mutual similarity between data objects, i.e. via a unsupervised learning process[21].

Due to the fact that the class labels are often not known before data analysis, for example, in case of being hard to assign class labels in large databases, clustering analysis is sometimes an efficient approach for analysing such kind of data. To perform clustering analysis, similarity measures are often utilized to assess the distance between a pair of data objects based on the feature vectors describing the objects, in turn, to help assigning them into different object classes/clusters. There are variety of distance functions used in different scenarios, which are really dependent on the application background [22].

IV PROBLEM DEFINITION

A major challenge was to find good datasets that can be used for data mining. To gain a good understanding of the data and to create models with reasonable support we are in need of complete and noise-free datasets. Most available datasets are not carefully selected nor up-to-date, hence, the task of predicting anything from this data will not yield good results. During this thesis, we came across a lot of datasets that were either incomplete or simply not expressive enough to allow an accurate prediction. Hence, we argue, that further experiments on data mining from Semantic Web data could be greatly facilitated with the creation of common datasets for the evaluation and comparison of different approaches. The benefit of our approach is based on the expressiveness of the underlying ontologies. While ontologies with a deep inheritance hierarchy can outperform data mining without ontology support,

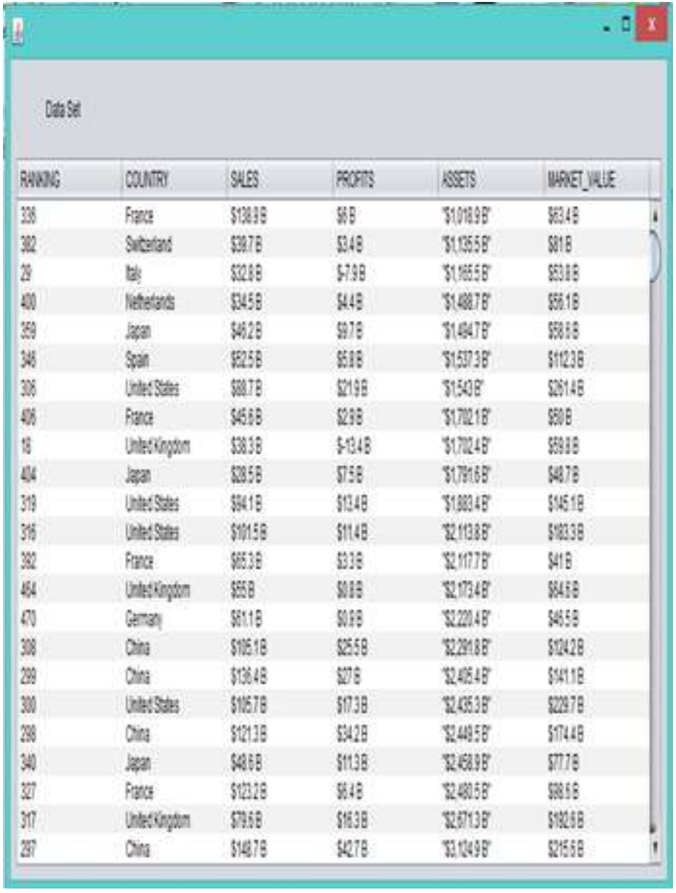
V PROPOSED WORK

This framework proposes agent based Semantic Web Mining System (SWMS). It will provide classification and clustering of the web contents according to user navigating links and time when navigating to other pages, thereby facilitating knowledge based response to the user and will highlight otherwise unnoticed patterns. It mainly comprises of Interface Agents, collection Agent supported with ontology database, content mining agent and clustering agent. Content mining agent works in collaboration with descriptive metadata agent and semantic metadata agent. Let us take an another example of web page searching like if user enters a query or phrase which contains the multiple meaning of that phrase. Ontology database will be searched for that query to mean and once meaning is derived it will hit to the DMA to descriptive metadata and then CMA, same process as above but here from user side, IA will send the information of place and top to bottom listing of the hitting link by user and ontology database will store it. And another time when users enters same query but hits on another link it will record the behavior of that site and next time it will proceed to provide correct result. As in example, A phrase entered Hotel in India‖ and this was queried in morning or before noon and user clicked to hotel to stay in that hotel but after noon or in evening user queried same phrase but clicked on hotel for dinner, lunch. This will recorded and will be stored in ontology database and next time when user enters same phrase after noon o in evening it will show the topmost hotels ready to dinner or lunch according to user behavior in past.

VI. EXPERIMENTAL SETUP AND RESULT ANALYSIS

Here in our research we are experimentally going to provide a simulation of semantic web mining system where query will work with different type of conditions related to past behaviour of query and ontology database and will help to provide accurate result.

Fig1 to show: Obtained result from the multi attribute


Data Set

RANKING	COUNTRY	SALES	PROFITS	ASSETS	MARKET_VALUE
338	France	$138.9 B	$6 B	"$1,018.9 B"	$63.4 B
382	Switzerland	$39.7 B	$3.4 B	"$1,135.5 B"	$81 B
29	Italy	$32.8 B	$-7.9 B	"$1,165.5 B"	$53.8 B
400	Netherlands	$34.5 B	$4.4 B	"$1,488.7 B"	$56.1 B
359	Japan	$46.2 B	$9.7 B	"$1,494.7 B"	$58.8 B
346	Spain	$52.5 B	$5.8 B	"$1,537.3 B"	$112.3 B
306	United States	$88.7 B	$21.9 B	"$1,543 B"	$261.4 B
406	France	$45.6 B	$2.9 B	"$1,702.1 B"	$50 B
16	United Kingdom	$38.3 B	$-13.4 B	"$1,702.4 B"	$59.8 B
404	Japan	$28.5 B	$7.5 B	"$1,791.6 B"	$48.7 B
319	United States	$84.1 B	$13.4 B	"$1,883.4 B"	$145.1 B
316	United States	$101.5 B	$11.4 B	"$2,113.8 B"	$183.3 B
392	France	$65.3 B	$3.3 B	"$2,117.7 B"	$41 B
464	United Kingdom	$55 B	$0.8 B	"$2,173.4 B"	$64.6 B
470	Germany	$61.1 B	$0.9 B	"$2,220.4 B"	$46.5 B
308	China	$105.1 B	$25.5 B	"$2,291.8 B"	$124.2 B
299	China	$136.4 B	$27 B	"$2,405.4 B"	$141.1 B
300	United States	$106.7 B	$17.3 B	"$2,435.3 B"	$229.7 B
298	China	$121.3 B	$34.2 B	"$2,449.5 B"	$174.4 B
340	Japan	$48.6 B	$11.3 B	"$2,458.9 B"	$77.7 B
327	France	$123.2 B	$6.4 B	"$2,480.5 B"	$98.5 B
317	United Kingdom	$79.6 B	$16.3 B	"$2,671.3 B"	$192.6 B
297	China	$146.7 B	$42.7 B	"$3,124.9 B"	$215.6 B

Performance Measures

1 Training Time

A training time of a dataset in Javais computed with the help of start and end time class variables defined in the tool and here as we load the dataset and verifies the eligibility and taking their features for consideration or not is the time taking process to identify and to load the dataset comes under training time of a dataset, extracting the properties and making them in process format is training time.

2 Testing Time

A testing time is the time of process we calculate and obtain the various threshold or classification related activity, we perform testing over the dataset where the dataset need to process after this step, Processing the extracted dataset information with our algorithm we called as testing time.

3 Accuracy

Accuracy is the value of exactness where we calculate a value and technique with which we match the expected output and the exact output we receive, we compare them and calculate the accuracy or analysis of result in this form and thus we get the quality of our work in this basis and justify the approach as best among the other algorithms which we have taken for consideration.

Performance evaluation based on the computation time based upon the results or the dataset we actually take for the data or ranking optimization:

Number of Outliers we have detected after performing outlier technique on the complete dataset and performing attribute based ranking on individual technique, After applying such technique we have received few dataset to further work on and optimization.

	Existing Approach	Outlier Approach
SALES	500	121
PROFIT	500	119
ASSET	500	111
MARKET VALUE	500	60

Figure 2: Number of Attribute received after applying outlier technique

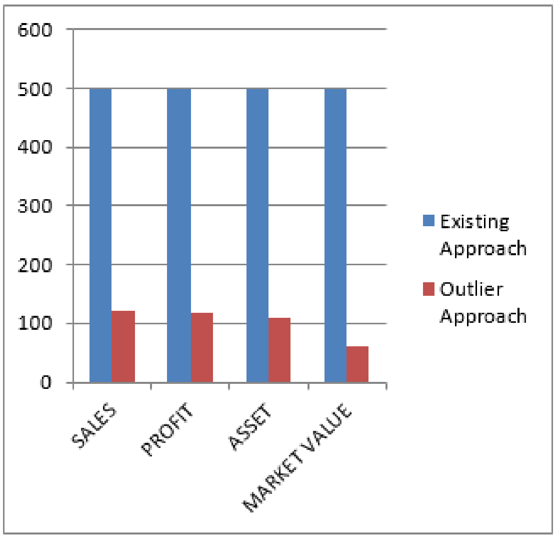


Figure 3: Graph – Dataset retrieved on applying different dataset attribute

VII CONCLUSION AND FUTURE WORK:

In this paper I have proposed a powerful enhanced line-up in finding very high quality solution for the ranking optimization. The proposed has found optimal or best known solution for most benchmark instances with up to max. Number of categories. One of the strengths of my technique is the use of line-up, the local version of ELU and global version of LU significantly reduced the computational cost, with the help of efficient implementation techniques. This resolves the common problem that line-up for ranking optimization are usually much more time consuming than efficiently implemented local search based algorithms. Another important contribution is the development of ELU in generating even better solution from very high quality parent solution at the phase of the line-up. An interesting feature is that I design a simple local search procedure into ELU to determine good combination of edges of attributes. I am going to demonstrate that the enhancements significantly improve the performance of the ranking optimization and other related tools which provide us efficient way to get a ranking on changing ranking attributes. I believe that the proposed ELU provides a good example of a sophisticated product comparison application for a reprehensive combinational optimization problem and that some of the ideas can be successfully applied to the design of LU for other combinational optimization problems. In this paper we conclude to get a best optimization technique on enhancing the current algorithm which is line-up available today where we are performing ranking optimization solution on the particular provide algorithms.

REFERENCES

[1] Sharma K., Shrivastava G. & Kumar V., ‗Web Mining: Today and Tommorrow'. In Proceedings of the IEEE 3rd International Conference on Electronics Computer Technology, 2011.

[2] Bhatia C.S. & Jain S., ‗Semantic Web Mining: Using Ontology Learning and Grammatical Rule Interface Technique'. In IEEE 2011.

[3]Kosala R. &Blockeel H., ‗Web Mining Research: A Survey'. Published in ACM SIGKDD, Vol. 2, Issue 1,July 2000.

[4] Eirinaki M. &Vazirgiannis M., ‗Web Mining for Web Personalization'. Published in ACM Transactions on Internet Technology, Vol.3 , No. 1, February 2003, pp. 1-27 [05] Z. Yang, B. Zhang, J. Dai, A. C. Champion, D. Xuan, and D. Li, "E-smalltalker: A distributed mobile system for social networking in physical proximity," in ICDCS, 2010, pp. 468–477.

[5] Meirong T. & Xuedong C. , ‗Application of Agent Based Web Mining in E-business'. Published in 2010 IEEE Second International Conference on Intelligent Human-Machine Systems and Cybernetics, pp. 192-195.

[6] Ting I.H., ‗Web Mining Techniques for On-line Social Networks Analysis'. In Proceedings of the 5th International Conference on Service Systems and Service Management, Melbourne, Australia, 30 June-2 July 2008, pp. 696-700.

[7] Jicheng W., Yuan H., Gangshan W. &Fuyan Z., ‗Web Mining: Knowledge Discovery on the Web'. In Proceedings of IEEE International Conference on System, Man and Cybernetics 1999 (IEEE SMC'99), Vol. 2 , pp. 137-141.

[08] Zhan L. &Zhijing L., ‗Web Mining based on Multi-Agents'. Published in proceedings of Fifth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'03), 2003.

[9] C.Dimou, A.Batzios, A.L.Symeonidis and P.A.Mitkas, ‗A Multi-agent framework for Spiders Traversing the Semantic Web'. In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence.

[10] F. Buccafurri, G. Lax, D. Rosaci and D. Ursino, ‗Dealing with Semantic Heterogeneity for Improving Web Usage'. Data Knowledge Eng. Vol. 58, Issue 3, pp. 436–465,2006.

[11] Singh A., Juneja D. and Sharma A.K., ‗Design of Ontology-Driven Agent based Focused Crawlers'. In proceedings of 3rd International Conference on Intelligent Systems & Networks (IISN-2009),Organized by Institute of Science and Technology, Klawad, 14 -16 Feb 2009, pp. 178-181. Available online in ECONOMICS OF NETWORKS ABSTRACTS, Volume 2, No. 8: Jan 25, 2010.

[12] Singh A., Juneja D., Sharma A.K., 'Design of An Intelligent And Adaptive Mapping Mechanism For MultiagentInterface'.In Proceedings of International Conference on High Performance Architecture and Grid Computing Communications in Computer and Information Science (HPAGC'11), 2011, Volume 169, Part 2, 373-384, DOI: 10.1007/978-3-642-22577-2_51.

[13] Singh A., Juneja D., Sharma A.K., ‗General Design Structure of Ontological Databases in Semantic Web'. Published in International Journal of Engineering, Science & Technology, Vol. 2, Issue 5, pp. 1227-1232, 2010.

[14] Karayannidis N. &Sellis T., ‗Hierarchical Clustering for OLAP: The CUBE File Approach'. Published in The VLDB Journal — The International Journal on Very Large Data Bases, Vol. 17, Issue 4, July 2008.

[15] Aarti Singh, ‗Agent Based Framework for Semantic Web Content Mining'. Published in International Journal of Advancements in Technology,Vol. 3 No.2 (April 2012), ISSN 0976-4860.

[16] Agarwal, S., Pandey, G. N., & Tiwari, M. D.:Data Mining in Education: Data Classification and Decision Tree Approach.(2012).

[17] Nicolas Garcia-Pedrajas, Javier Perez-Rodriguez, Aida de Haro-Garcia , ―OligoIS: Scalable Instance Selection for Class-Imbalanced Data Sets‖ , IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics.

[18] Dr.D.Ramyachitra, P.Manikandan,‖ Imbalanced DataSet Classification and solution:A Reviw‖ International Journal of Computing and Business Research (IJCBR) ISSN (Online) : 2229-6166 Volume 5 Issue 4 July 2014.

[19] Anyanwu, M. N., & Shiva, S. G.: Comparative analysis of serial decision tree classification algorithms. Vol.3, 230-240 International Journal of Computer Science and Security(2009).

[20] Asuncion, A., & Newman, D. J. UCI Machine Learning Repository. Irvine, CA: University of California. School of Information and Computer Science. 2007.

[21] Bakar, A. A., Othman, Z. A., & Shuib, N. L. M. : Building a new taxonomy for data discretization techniques. In Data Mining and Optimization, 2nd Conference. 132-140. IEEE (2009).

[22] Balagatabi, Z. N., & Balagatabi, H. N.: Comparison of Decision Tree and SVM Methods in Classification of Researcher's Cognitive Styles in Academic Environment. vol.1, 31-43. Indian Journal of Automation and Artificial Intelligence, (2013).

[23] Bramer, M.: Decision Tree Induction: Using Entropy for Attribute Selection. In Principles of Data Mining. 49-62. Springer London. (2013).

[24] Bunkar, K., Singh, U. K., Pandya, B., & Bunkar, R.. : Data mining: Prediction for performance improvement of graduate students using classification. In Wireless and Optical Communications Networks (WOCN), Ninth International Conference, 1-5. IEEE. (2012).

[25] Burrows, W. R., Benjamin, M., Beauchamp, S., Lord, E. R., McCollor, D., & Thomson, B.: CART decision-tree statistical analysis and prediction of summer season maximum surface ozone for the Vancouver,

Montreal, and Atlantic regions of Canada. Vol. 34 1848-1862, Journal of applied meteorology(1995).

[26] Cortes, C., & Vapnik, V.: Support vector machine. Vol.203, 273-297. Machine learning (1995).

[27] Cover, T., & Hart, P.: Nearest neighbor pattern classification. Information Theory. Vol. 13, 21- IEEE Transactions (1967).

[28] Dasarathy, B.V.: Nosing Around the Neighborhood: A New System Structure and Classification Rule for Recognition in Partially Expos Environments. Vol. PAMI-2, No. 1, 67-71. Pattern Analysis and Machine Intelligence. IEEE Transactions (1980).

[29] Stonebraker, M., Çetintemel, U., Zdonik, S.: The 8 requirements of real-time stream processing. ACM SIGMOD Record. 34, 42-47 (2005).

[30] Cugola, G., Margara, A.: Processing Flows of Information : From Data Stream to Complex Event Processing. ACM Computing Surveys.

[31] Niblett, P.: Event Processing In Action. (2010).

[32] Wang, Q., Meegan, J., Freund, T., Li, F.T., Cosgrove, M.: Smarter City: The Event Driven Realization of City-Wide Collaboration. 2010 International Conference on Management of e-Commerce and eGovernment. 195-199 (2010).

[33] Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., Pichler-Milanovic, N., Meijers, E.: Smart cities Ranking of European medium-sized cities. , Vienna, Austria (2007).

[34] Transport for London, Live Traffic Disruptions – Data Dictionary, (last accessed 1st September 2012), http://www.tfl.gov.uk/assets/downloads/businessandpar tners/data-dictionary-live-traffic-disruptions.pdf.

[35] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten. The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1, 2009.

[36] C. Romero, S. Ventura, E. Garcia, "Data mining in course management systems: Moodle case study and tutorial", Computers & Education, Vol. 51, No. 1, pp. 368-384, 2008.

[37] C. Romero, S. Ventura "Educational data Mining: A Survey from 1995 to 2005", Expert Systems with Applications (33), pp. 135-146, 2007.

[38] Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M. Inayat Khan, “Data Mining Model for Higher Education