ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 3, Issue 1, January 2017

A multiple database handling technique for enhancing the results quality using entity mining

Hemali Damania[1] ,Prof. Miss. Khushboo Sawant[2] ,Prof. Kuntal Barua[3]
PG Scholar, Dept.Of Computer Science & Engg., JDCT,Indore, M.P., India[1]
Professor, Dept.Of Computer Science & Engg., JDCT,Indore, M.P., India[2]
HOD, Dept.Of Computer Science & Engg., JDCT,Indore, M.P., India[3]

ABSTRACT— The use of technology is in current scenarios for reducing the problems into small sets of problem and finding efficient solutions are performed. In this context the algorithms are used to handle the complex issues to find their optimum solutions. In this presented work a new algorithm is introduced for finding the most optimum results during the database based query processing. In this case the data base is not a single source of data that can be published by the multiple product vendors and service providers. The key issue in such environment is handling the different attributes sets of data, handling of the duplicate results during the search or query processing. In addition of that the similarity issues of the source and target data source schemas. Therefore to deal with these issues a new solution is proposed. The proposed solution handles the data sources by its semantic and syntactic relationship basis. Therefore a common solution for the issues is reported. That is able to find the similar attributes among the source and target data source schema for generating a common data format by which the system efficiently satisfy the user queries and results much precise outcomes. The implementation of the proposed technique of multiple data source handling and entity mining concept is provided using JAVA technology more specifically using JSP (java server pages). After implementation of proposed concept the performance of the proposed algorithm is measured for finding their search relevancy in terms of precision, recall and f-measures. In addition of that for demonstrating their resource persevering ability the performance in terms of space and time complexity is also computed. The results show the proposed technique is effective and accurate technique for querying with the multiple data sources at the same time for comparative QoS study**.**

KEYWORDS— Multiple databases, entity mining, query processing, results optimization, entity mining.

I. INTRODUCTION

A number of techniques are developed recently to automate the manual task. In place of traditional shops the internet based applications are becomes popular for shopping and obtaining the different kinds of services. In these applications a single service can be offered by more than one vendor, but the quality and prices are different for both the services and/or products. Thus to select the more suitable services from the available services are a complicated task. The machine learning based technique can reduce human efforts by providing strict decisions for these real world issues. Not only can these methods offers the classification and categorization task these technique are also helpful to recognize the similar patterns. This property of data mining can help to evaluate the different attributes of data to find the similar attributes among different data sources. And also helps to refine the user required data according to the user query. In this presented work data mining techniques are utilized for finding the semantically and syntactically similar attributes among the rich set of different kinds of attributes in different data sources. In data mining techniques example based learning is performed to understand the data patterns based on the learning patterns the data is identified by the applications. Those are the learning process of supervised algorithms but in unsupervised learning the data consumes the available patterns and based on their internal similarity the outcomes are demonstrated. In the presented work a unsupervised manner of attributes identification is required by which the algorithm identify the similar semantic and syntactic attributes and grouped together for information retrieval purpose. In this

presented work the structured data for learning and knowledge extraction is utilized in unsupervised manner with some semantic constrains. Because attribute selection for comparing the services of different service providers over the internet based applications is required with different attribute names.

II. PROPOSED WORK

Information retrieval and their improvements is the classical domain of research and development. But there are fewer efforts are made to improve the results during organizing the data sources at the query time. The presented effort of named entity mining is dedicated to improve the information retrieval for data source aggregation and query time performance improvements. The section provides the detailed overview of the proposed work and included solution for efficient and accurate data retrieval.

A. System overview -Now in these days the internet becomes an important channel for communication, service and other daily routine work. A number of different applications are developed with the support of internet technology. Using the internet the service providers are directly connected with their clients. But sometimes for the same services a number of service providers are exist. These service providers offer the similar product or services with different quality of service parameters. Therefore a technique is required by which the different quality of products and services are compared with each other. Therefore in traditional techniques the query processing is performed for finding optimum results from the databases. In order to reduce the data redundancy and improve the quality of search results a new technique is need to develop.

In this presented work a new technique for aggregating different database and for finding the optimal search results based on user query a new technique using attribute comparisons is proposed. The proposed technique involves the semantic and syntactic similarities for finding more effective outcomes from the databases. In addition of that the proposed technique is also helpful to deal with the huge amount of data in same place, therefore the data management cost and efforts are also effectively reduced. This section provides the overview of the proposed concept in next section the key aspects of the proposed model is provided. In next section the detailed problem domain is provides the issues involved in the system.

B. Detailed Problem discussion - This section provides the key issues and challenges that are required to resolve for finding the optimum solution for the multiple databases entity mining and comparative study.

- Now in these days a number of online databases are existing. These databases are open to extract the data for different products and services. In addition of that various intermediate applications are also offers the comparison of products. But the direct query execution in the multiple databases may affect the precise outcomes from the databases. In addition of that, for a single service or products a number of results can be found from the different data sources. That misguides the user for selection of accurately required product or service.
- Consider figure 1 where the similar products description is available in two different data sources and both are having their own definitions of product or service. In this context need to find similar attributes by which the database queries results less duplicate outcomes. Therefore need to provide a method by which the different attributes are mapped and aggregated in a common format.
- Form the above given figure 1 the semantically comparison based issues are also find. For example in a car database, database designer assume an attribute as COST and the other database where the designer assumed this attribute as PRICE.
- Synthetic issues are also a key challenge of the proposed work, suppose that a common field can be distributed by the multiple attributes such as name of a person can be denoted by “FULL NAME” or by the two attributes “FIRST NAME” and “LAST NAME” or “FIRST NAME”, “MIDDILE NAME” and “LAST NAME”.

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 3, Issue 1, January 2017

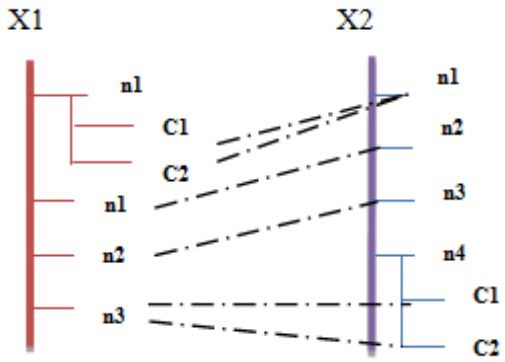

Fig. 1 Example Data Sources

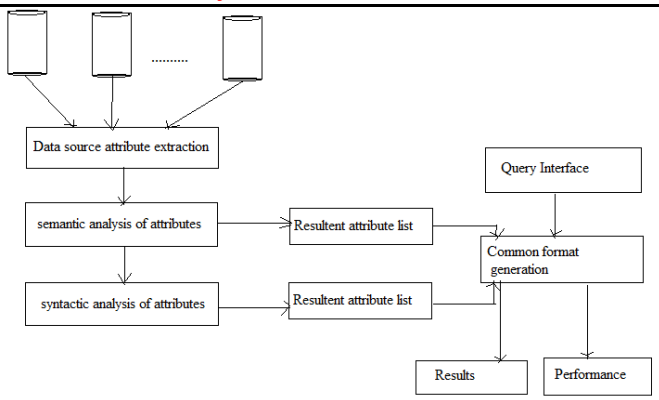

Fig.2 Proposed System

C. Solution Methodology

The proposed system for entity mining technique is understood using the figure 2. In this diagram the entire system components are demonstrated with their functional aspects. The proposed system is demonstrated using figure 2, in this diagram the input and the output of the system is represented by analysis of the input parameters. The individual steps and their key responsibility for the proposed system model are demonstrated in this section. Therefore the component basis description is provided.

Data sources: the proposed system needs more than one data sources as input. The input data sources can be belongs to any kind of products information or any class of services which is offered through the different kinds of vendors. Thus, in this phase first need to prepare or manage the multiple connections with the different services provider's databases.

Data source attribute extraction: now after establishing connection with the different service provider's databases. Need to find the target data tables or schemas which is required to compare and utilized for comparative study. Therefore in this phase a provision is need to made for selection of data tables or the attributes which is used for further processing.

Semantic analysis of attributes: after finding the data tables or the schemas of the target databases the attributes are analysed in Semantic manner. Thus a separate list of similar semantic data or synonyms is prepared which is used to map the given attribute into the possible semantic means. Suppose in the data base an attribute is defined using the COST thus there are the possible semantic attributes can be price, amount, or charges. In this phase by using the possible means of data attributes the target attributes are compared for finding the possible match on the target database.

Synthetic analysis of attributes: after analysing the data semantically the matched attributes are separated in previous phase. In this phase the remaining attributes are used for both the data sources to evaluate the attributes syntactically. Therefore the different combinations of the attributes are prepared and that is compared to the target attributes list. The most similar attributes are separated in this phase.

Resultant attributes: the system first compares the databases and their attributes to find the matched attributes at a time a single data source and their attributes is selected and compared with the other available attributes list. The matched attributes are separated and preserved at both the stages of evaluation.

Common format generation: by the outcomes of the semantic and syntactic analysis the matched attributes are identified. Using the obtained information from the both kinds of analysis the matched attributes and their data is combined in this phase for finding the common structure of database.

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 3, Issue 1, January 2017

Query interface: that is basically a user interface designed for supporting the user oriented query firing and obtaining the search outcomes. The search outcomes are generated on the basis of multiple data sources organized in a common format.

Results: after performing the search operation the search results are collected through the data structure and according to the query most relevant search results are produced.

Performance: during the generation of outcomes for queried data the performance of the system is also computed in terms of their precision, recall and f-measures.

This section introduces the proposed system model used for performing the search using the multiple data sources. In the next section the proposed algorithm steps are reported.

D. Proposed algorithm- In order to transform the described data model in terms of the solution algorithm steps the following step of solution is proposed. Suppose the system involve the N number of service provider's databases. These data bases can be denoted using the $D = \{D_1, D_2, \dots, D_N\}$ where the D is database and N is number of parties. To compare the similarity among the different databases attributes first a list of attributes are required to find, thus a data table can be defined as $D_i = \{A_1, A_2, \dots, A_m\}$ where the A is the participating attributes and M is the number of attributes in list. For simplicity here need to consider two different data sources at a time. Therefore the first data source is termed as source data table S and the second data source which is compared with the first one is known as target data table T. thus if $D_i = S$ then the target data attributes can also be defined by $T = \{A_1, A_2, \dots, A_o\}$ where the target data source contains O number of attributes list. Now need to find the best matched attributes from source and target attributes list.

Input : Source attribute list S, target attribute list T
Output: common format of data CD

Process:
SCount = S.length()
Tcount = T.length()
$for(i = 1; i \leq Scount; i++)$
$Selected = S[i]$
$Smeta = selected.getMetaData()$
$for(j = 1; j \leq Tcount; j++)$
$Targeted = T[j]$
$Tmeta = Targeted.getMetaData$
$if\ Smeta == Tmeta$
Result [i]=T [j];
End if
End for
Update S and T as Snew and Tnew
End for
NewScount = Snew.length()
$for(n = 1; n \leq NewScount; n++)$
$newSelect = Snew[n]$
$LST = list.synonyms(newSelected)$

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 3, Issue 1, January 2017

$for(m = 1; m \leq NewTcount; m++)$

$Bool = LST.findSim(Tnew[m])$

$if\ Bool = true$

Result.Add($Tnew[m]$)

End if

$end\ for$

$end\ for$

CD = copy(S, T, Results)

Return CD

Table 1 similarity index

After finding the attributes those are synthetically or semantically similar, the attributes are separated. After combining the attributes the new data source CD is prepared. That CD is used making the search by the user, thus the user fires their queries using the query interface for finding the relevant records from the CD or common data format, for comparative study of product and/or services.

III. RESULTS ANALYSIS

The chapter introduces performance study of the proposed entity mining technique. In order to evaluate the developed system different performance factors are evaluated and their outcomes with the different amount of datasets are provided.

A. Precision- In most of the cases any information retrieval systems is evaluated on the basis of their outcomes and their accuracy. The precision is the similar to the accuracy for the search systems. Actually precision is the part of search result that is relevant to input user query. This can be computed using the search results produced by the information retrieval systems and by following formula.

$$precision = \frac{releventdocument \cap retrieveddocumets}{retrieveddocuments}$$

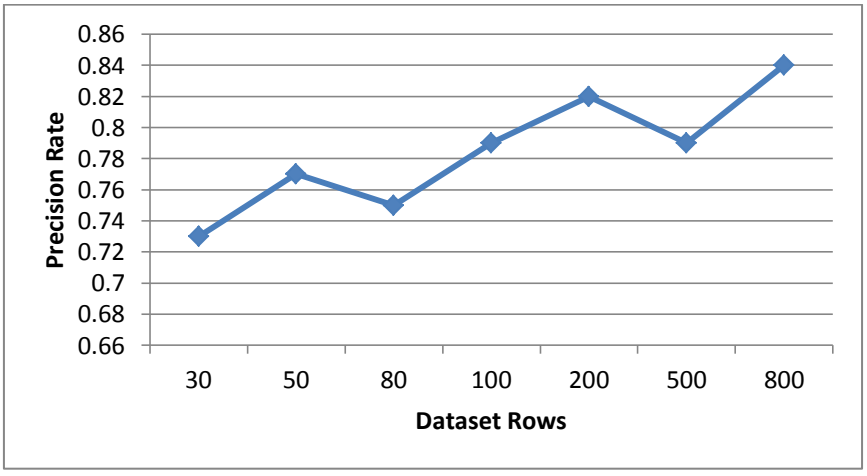

Fig. 3 Precision rate

Dataset size	Precision
30	0.73
50	0.77
80	0.75

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 3, Issue 1, January 2017

100	0.79
200	0.82
500	0.79
800	0.84

Table 3.1 precision

The precision rate of the proposed technique is given using figure 3 and table 3.1. In this diagram X axis includes the different size of dataset rows used during experiments. In the similar figure Y axis shows corresponding precision rate. According to the obtained results the performance of the proposed technique is more precise in addition of that that becomes more accurate as the amount of data is increases. But that is depends on the consistence of the attributes participating in the mining and the values of objects.

B. Recall- In data retrieval application or the search applications the recall values are measured for finding the accuracy in terms of relevant document retrieved among the available set of documents or relevant data obtained according to input user query over the given set of documents. This can be evaluated using the following formula.

$$recall = \frac{relevantdoucment \cap retrieveddocuments}{relevant\ documents\ in\ database}$$

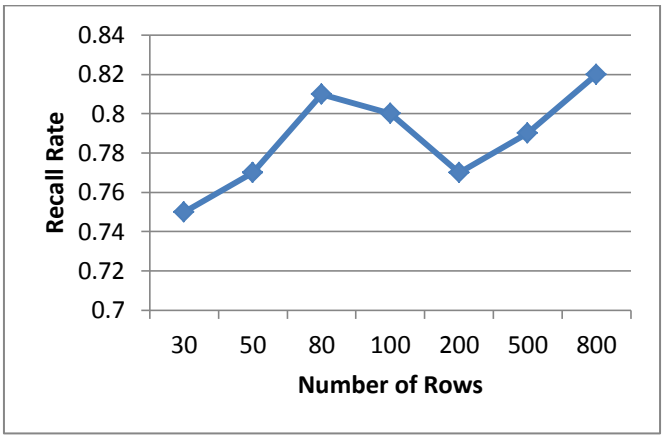

Fig.4 recall rate

Dataset size	Recall
30	0.75
50	0.77
80	0.81
100	0.80
200	0.77
500	0.79
800	0.82

Table 3.2 recall values

The computed recall values of the proposed entity mining technique are given using figure 4 and table 3.2. In this diagram X axis shows different experiments performed on variable length of data and the Y axis shows the recall values according to experiments. According to the obtained performance the performance of the proposed technique is accurate.

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 3, Issue 1, January 2017

C. F-measures-The f-measures of information retrieval system demonstrate the fluctuation in the computed performance in terms of precision and recall rates. The f-measures of the system can be approximated using the following formula.

$$F - measures = 2.\frac{precision\ X\ recall}{Precision + recall}$$

The comparative f-measures of the proposed technique are given using figure 5 and table 3.3. In this diagram the X axis shows the different experiments performed with different number of rows in databases and Y axis shows the corresponding f-measures of the system. According to the obtained f-measures the proposed technique has more stable and accurate results. Thus the proposed technique is more adoptable.

Dataset size	f-measures
30	0.73
50	0.77
80	0.77
100	0.79
200	0.79
500	0.79
800	0.82

Table 3.3 f-measures

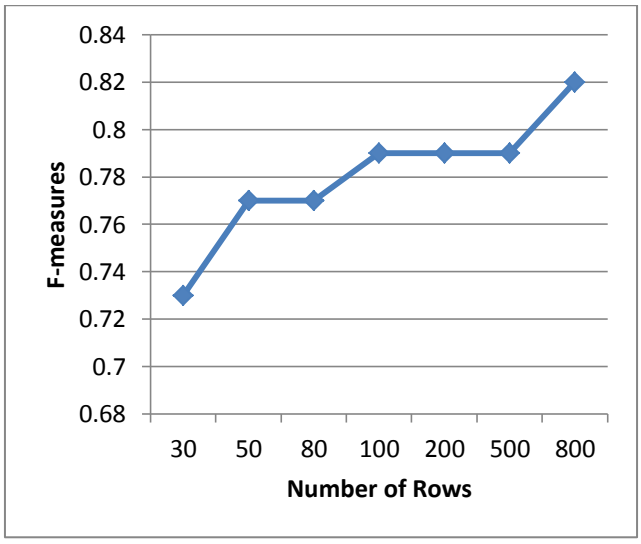

Fig.5 f-measures

D. Memory usages- The amount of main memory required to aggregate the different data source and extract the similar attributes is known as the memory consumption or space complexity of algorithm. The memory consumption of the implemented system is given using figure 6. In this diagram the X axis shows the different experimental scenarios with different size of datasets and the Y axis shows the respective memory consumption in terms of kilobytes (KB). According to the obtained performance of the system the proposed technique consumes higher memory space.

Dataset size	Memory consumption
30	26589
50	27485
80	28649

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 3, Issue 1, January 2017

100	29965
200	30014
500	31584
800	33225

Table 3.4 memory consumption

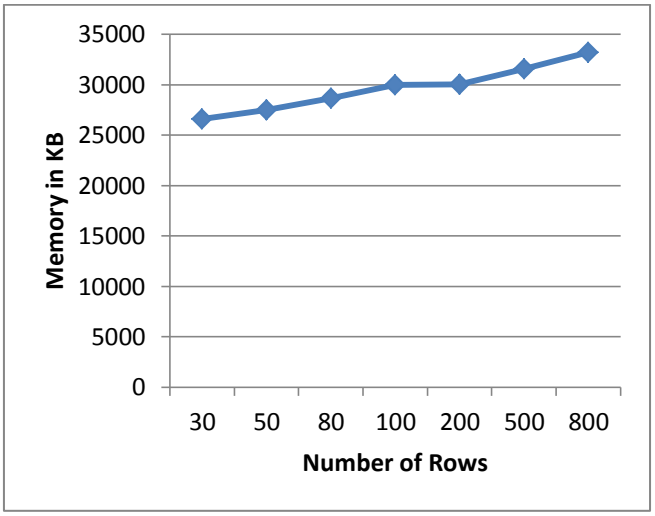

Fig.6 memory consumption

E. Time consumption- The amount of time required to process the number of data sources during the knowledge extraction is termed as the time consumption or time complexity of algorithm. The time consumption of the proposed algorithm is given using figure 7 and table 3.5. In this diagram the X axis contains the different experiments performed with different amount of data rows and the Y axis shows the amount of time consumed during experiments. The computed time consumption is given here in terms of milliseconds. According to the obtained performance the proposed algorithm requires less amount of time during the evaluation of entity and produces fewer amounts of results.

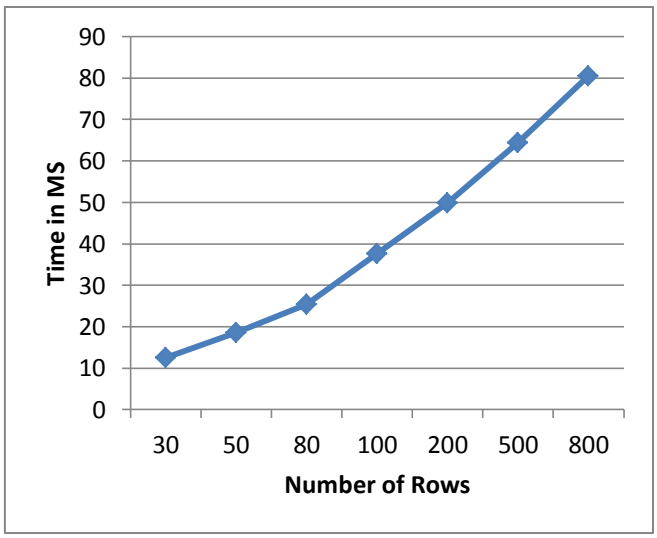

Fig.7 time consumption

Dataset size	Time complexity
30	12.56

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 3, Issue 1, January 2017

50	18.64
80	25.39
100	37.68
200	49.92
500	64.35
800	80.52

Table 3.5 time consumption

IV. CONCLUSIONS

The main of the proposed work to enhance the multiple data sources query processing and obtaining the refined and precise data is implemented successfully. In addition of that the performance of the presented technique is also evaluated. This chapter provides the summary of the proposed research work as conclusion and the future extension of the work is also suggested.

A. Conclusion- The entity mining is domain where the similar data sources are mined for combining the similar features of different data sources. The key area of application for this approach is to combine the data bases in same place and find the comparative attributes during the search processes. Using the current concept different comparisons web applications are now in existence. In order to improve the performance of the existing technique of comparative entity mining a new method is required. By which the database instances can be handled in less efforts. In addition of that the issues of duplicate results are also reduced. Finally the more relevant information can be queried by number of data sources parallel manner.

Therefore the proposed work is intended to design a new solution that optimizes the search outcomes from the multiple data sources, without reforming the user query. The proposed technique considers a number of data sources with the different location and different attributes. First both the database attributes are compared semantically and syntactically for finding the similar attributes. Finally the data is transformed into a common format for effective query satisfaction. Therefore the proposed technique involves a combine information processing infrastructure that processes data and generate the combine semantic outcomes in less redundant manner.

The implementation of the proposed concept of the entity mining technique is provided using the JAVA base web application. Thus it is designed with the help of JSP (java server pages) based technology. Finally for justifying the technique the proposed technique is evaluated in different performance parameters such as precision, recall and f-measures. In addition of that to compute the resources consumption the method is tested for space and time complexity estimation. The observed performance summary is provided using table 4.1.

S. No.	Parameters	Remark
1	Precision	The proposed technique provides the high precise results and low
2	Recall	The technique produces less duplicate results in query processing thus
3	F-measures	The less fluctuating the performance is found thus the f-measures of
4	Memory consumption	The technique consumes less memory space for processing the multiple
5.	Time complexity	Less time consuming technique for processing large number of

Table 4.1 performance summary

According to the obtained results the proposed technique is efficient and produces more precise results for comparing the attributes form different data sources of the similar categories.

B. Future work- The main aim of the proposed work to find the aggregated knowledge from the multiple data sources are implemented successfully. Additionally the technique is found the efficient and accurate. In near future the proposed technique is extended for the following area of applications.

1. The current system only able to distinguish the similar attributes from the similar categories of products in near future the technique is modified for finding the similar schema databases also.
2. The system is demonstrated with the two party data sources in near future the technique is need to extend for multiple data sources.

REFERENCES

[1] Pavlos Fafalios, Claudio Baldassarre, Ioannis Kitsos, Michail Salampasis, Yannis Marketakis, and Yannis Tzitzikas, "Web Searching with Entity Mining at Query Time", 5th Information Retrieval Facility Conference, IRF 2012, Vienna, July 2012

[2] Data Mining: What is Data Mining?,http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.html. Data Mining - Applications & Trends, http://www.tutorialspoint.com/data_mining/dm_applications_trends.htm

[3] Mahak Chowdhary, Shrutika Suri and Mansi Bhutani, "Comparative Study of Intrusion Detection System", 2014, IJCSE All Rights Reserved, Volume-2, Issue-4

[4] Mrs. Pradnya Muley, Dr. Anniruddha Joshi, "Application of Data Mining Techniques for Customer Segmentation in Real Time Business Intelligence", International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163, Issue 4, Volume 2 (April 2015)

[5] "ER Model : Basic Concepts", http://www.idc-online.com/technical_references/pdfs/information_technology/Er_Model.pdf

[6] Kalaivani. K, Amutha Prabakar. M, "ANALYSIS OF BIG DATA WITH DATA MINING USINGHACE THEOREM", Journal of Recent Research in Engineering and Technology ISSN, Volume 2 Issue 4 Apr 2015

[7] VASANT DHAR, "Data Science and Prediction", COMMUNICATIONS OF THE ACM | DECEMBER 2013 | VOL. 56 | NO 12

[8] Distributed Databases, Learning Objectives, chapter 12 and 13 http://wps.prenhall.com/wps/media/objects/3310/3390076/hoffer_ch13.pdf

[9] Tusal Patel, Preeti Gupta, Nishant Khatri, "Distributed SAP HANA Database for Efficient processing", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 6, June 2013

[10] Shrutika Narayane, Sudipta Giri, "A Review on Comparable Entity Mining", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 12, December 2014

[11] Gu Xu, Shuang-Hong Yang, Hang Li, "Named Entity Mining from Click-Through Data Using Weakly Supervised Latent Dirichlet Allocation", KDD'09, June 28–July 1, 2009, Paris, France. Copyright 2009 ACM

[12] Pavlos Fafalios and Yannis Tzitzikas, "Post-Analysis of Keyword-based Search Results using Entity Mining, Linked Data and Link Analysis at Query Time", 2014 IEEE International Conference on Semantic Computing (ICSC)

[13] Pavlos Fafalios, Manolis Baritakis and Yannis Tzitzikas, "Configuring Named Entity Extraction through Real-Time Exploitation of Linked Data", WIMS'14, June 02-04 2014, Thessaloniki, Greece Copyright 2014 ACM

[14] Jiawei Han,Chi Wang, "Mining Latent Entity Structures from Massive Unstructured and Interconnected Data", SIGMOD'14, June 22–27, 2014, Snowbird, UT, USA Copyright 2014 ACM 978-1-4503-2376-5/14/06

[15] Kyoji Kawagoe, Carson Kai-Sang Leung, "Similarities of Frequent Following Patterns and Social Entities", 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, Procedia Computer Science 60 (2015) 642 – 651, 2015 by Elsevier

[16] Hongchang Lai, ShuoXu, Lijun Zhu, "Chemical and Biological Entity Recognition System from Patent Documents", Proceedings of the Second International Workshop on Patent Mining and its Applications (IPAMIN) May 27–28, 2015, Beijing, China

[17] ZhengXu, XiangfengLuo, Shunxiang Zhang, Xiao Wei, Lin Mei, Chuanping Hu, "Mining temporal explicit and implicit semantic relations between entities using web search engines", Future Generation Computer Systems, © 2013 Elsevier B.V. All rights reserved.

[18] Panikala Madhavi, S. Vijayalaxmi, "COMPARABLE ENTITY MINING FROMCOMPARATIVE QUERIES", Aurora's International Journal of Computing | 2015 | Vol. 2. Issue 1 | January-June 2015.

[19] Maksim Tkachenko, Hady W. Lauw, "Generative Modeling of Entity Comparisons in Text", CIKM'14, November 3–7, 2014, Shanghai, China, ACM 978-1-4503-2598-1/14/11.