Proc. of Int. Conf. on Current Trends in Eng., Science and Technology, ICCTEST

Integration of Diverse Availability of Data for Extracting Semantic Properties

Mohan Kumar A V[1], Shankara Gowda S R[2], Dr. Nanda Kumar A N[3]
[1, 2] Dept. of CSE , Don Bosco Institute of Technology, Bengaluru
[1, 2] Email: {avmohan.sjce,shankargowdasr}@gmail.com
[3] Dept. of ISE,New Horizon College of Engineering, Bengaluru
[3]Email: nandarajnk@rediffmail.com
[1, 2, 3]Visvesvaraya Technological University, Belagavi

***Abstract*— We have many industries and companies today which are generating enormous amount of data in many diverse formats, either it may be structural or may it be in unstructured formats. There is need for extracting the semantic properties of these data are vital, the data available should be made use and repetitive work should be avoided as such, this can be achieved by analyzing the semantic properties hidden in the abundant data. For example, mobile operating companies usually have their clients who provide personal data in different form. However, making these diverse data into meaningful text requires lot of manual data processing which involves recognizing different formats of data, processing them and presenting in usable formats require further processing. Currently there is no automated software tool as such which can do this tedious work. In this paper, we are trying to propose a novel integrating tool which can extract the data from heterogeneous source and this tool can analyze the semantically properties involved in the data.**

***Index Terms*— Heterogeneous Data, Structured and Unstructured data, Data Integration**

I. Introduction

With the advanced development of the internet today, and due to its subsequent exponential growth, we are now in to information era of internet. We are drawing many source of information from the internet today. Taking up of internet to this level has seen that, the amounts of data with variety of information are accessible through the internet also growing exponentially. Starting from traditional method of writing data in papers to and now storing them in electronic formats has wide difference. There is an importance of knowing its semantics [1]. When data are done in electronic formats, editing is easier.

In fact user can even monitor and track their modifications done. Searching and tracing and storing of data are efficient when data is done in electronic formats. Making any documents in electronic formats provides a convenient way to share among different users across the world. Today almost all companies have abundant and voluminous data available which they usually present in unstructured formats. For example, accounting companies and banking divisions, who have many customers who will provide financial data to each sectors above mentioned. This financial data which they have submitted varies in different formats, either it may be

Grenze ID: 02.ICCTEST.2017.1.115

pay slips; bills etc. and they even contain n formats. More ever these data were may be written using different programming languages. Although the data are different and diverse in nature, they usually contain some common set of information. Like every pay slip contain the salary details of an employee, which intern contain employee name, employee number, his/ here EPF Number, income tax details and so on.

To make facilities to processes these data and to make it present in usable, readable and more importantly, to makes best use from the information available. Most companies even today insists on their employees enter the, data manually [2]. Even though many companies develop their own processing tool, which is used to indentify different formats available in the inputs and recognize them. However, still there is no existing automated soft ware's to process the data and recognize the formats.

In this paper, we are trying to propose a novel idea to generate automated engine, which can extract data, integrate data and analyze the data about as it is structured or unstructured data of diverse nature, consisting of data across the world. The technique can be widely applicable which can be having the ability to manage and handle languages that are used across the world. This engine can able to integrate the data which are written in various formats and also can differentiate data which contain context and meaning

II. RELATED WORK

Automated integration of data tool is usually divided into several types. First one contain, the tool which needs an extractor engine which extract the data from diverse internet sources, these data extracted from source then need to be cleaned and integrated into a single frame work. These frame work with needs to be defined with the data type and formats defined. Most automatic extraction tools which are available today do not make use of the semantic properties of data sets or records in their plan.

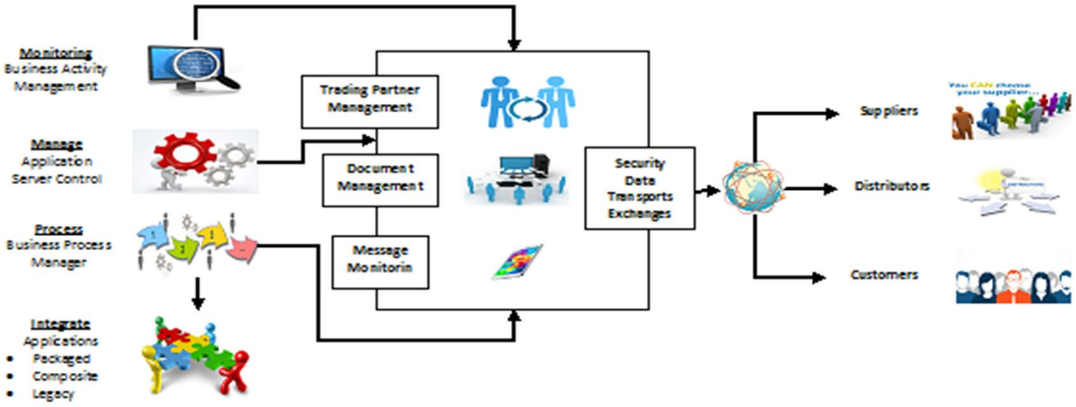

Figure 1. Data integration process

A. Data Extractor

These automatic extraction tools which are in use today, extract data records by inspecting and checking the patterns in the records [3]. For example Mining Data Region which usually called MDR, checks the recurring HTML tags and locates data records. Generally, this extraction engine is not accurate as it is unable to use links and page movements. Another example include, ViNT, which uses content lines like finding out rectangular box enclosing HTML tags and Text nodes.

B.Data Integrator

ViPER, another tool which enhances algorithm of MDR by using primitive cycle to detects the repetitive series of HTML Tags using a matrix. WordNet is one which is used to analyze the semantic properties available in data, it also checks for the synonym and words containing disambiguate. OW is yet another tool which can wrap accurately the web deep data.

Today we have various tools available for integrating data. These tools which are designed especially for specific purpose and specific domain, and most of all are language dependents [5]. For an instance consider a tool which developed in 2009, used to analyze the human structural gene. So, data integration tools are usually a field which dependents and May usually contains different formats and structures

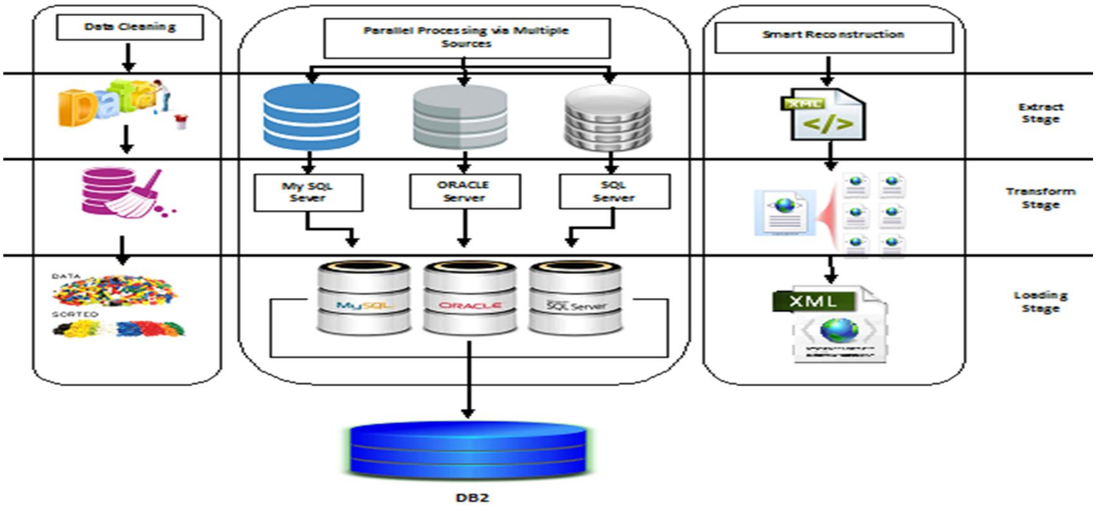

Figure 2. Various Stages in Data Integration

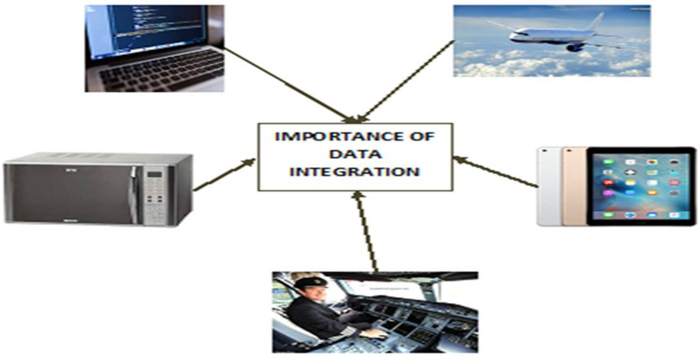

Figure 3. Importance of Data Integration

C. Data Analysis

Once the data is ready integrated, then it needs to be checked again for analysis purpose and further process. We have again various available tools which will do the analyze part [6]. Some of the worth mentioning work in data analysis part started in 70's where the subjective words are identified and classified. We have Jeff and Ales who developed a social media analysis tool which can be used to analyze user's idea of insight and online communications. Also in 2003, Jeonghee developed a sentimental mood analyzer tool which used natural language for processing the data.

III. Proposed Method

A. Overview

Here we try to propose a novel methodological tool which is used to extract the data from diverse heterogeneous and multiple data sources like World Wide Web. This tool will be able to analyze the semantic attributes i.e. properties of the data [7]. Using this semantic characteristic, we can analyze the data in to several classifications, where we can't extract meaningful approach for the additional processing. Once they are done with formatted things, we can develop a system or engine to analyze the data from incorporate the changes into business decision making things.

B. Data Extractor

The data extractor consists of components, which can be used to check complete contents of data in the document form. To able to achieve this, we need to tokenize the textual things into words. Then these tokens can be formed as list. Using this list, the word to word similarity is identified. Two words having similar meaning can be grouped to gather to form a new set of list [8]. This new list is then given a name or label. Which will show the list when another word with same meaning is encountered.

WorldNet Similarity then measure the similarity measure by identifying meaning with the hierarchy of words. For example, if we take up reptiles like mamba and taipan, as both are legless lizards which lack their eye lids and external ears. Although several algorithms are present to measure the similarity of the words or phrases in both, we can consider these features as data sets. One thing also to be note is that, the algorithms should find the similarity evaluation based on the return value.

The return value decides the similarity, where the similarity can be considered if it found more than 75% with the matching words. Some instances even may occur where the same words may contain single entity or thing, consider for example, the word king cobra, is usually a name of most poisonous snake, which contain two words, the method checks for the words which usually have more than two words [9].

To check these types of multiple words, which contain single meaning and convey single thing or entity, we have to first group these two adjacent words in to single word and every pair of words are grouped together with their respective entities. We can match this for the entity using similarity check engine or tool.

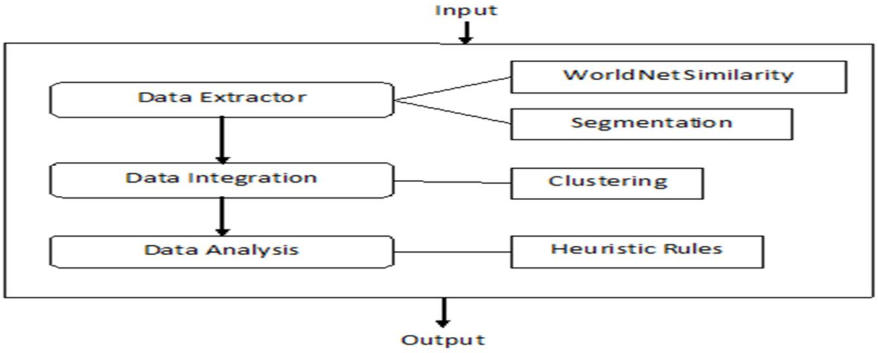

Figure 4. Framework for Data Integration Tool

Words are considered similarity only if they found 75 % similarity and once the words are matched, we can construct entities table with more than three words and can procedure for matching repeatedly. This procedure can even be repeated for the remaining entities also [10]. At last, to the end of the extracting stage, we can group a list of similar words. After this we can document the results into several regions.

C. Data Integrator

Once the segments are constructed and done, we can further group the individual data segments. To group them, we can use Word Net similarity check to find out the similarity of the words. Once the data are grouped and categorized in a table, they can be further classified upon their common features. We can use Self Organizing Map Clustering technique is used to group these data into large aspects of groups; it reduces the threshold value of similarity check to 0.7 Percentage to group further.

D. Decision Support System

Finally, the developed engine can be automated to use for the data in a significant way. To understand this, we have to know about the business rules and organizational rules. Data from comparable clusters can be displayed on the monitors according to the requirements. We can also develop a hierarchical association of data and display the content. For example, we can inform users about that both mamba and taipan are of reptile category and they have most poisonous venom in them. The link with the detailed document can also be shown to the user with API interface, so that users can integrate with their systems effectively.

IV. Conclusion

In this paper, we are proposing a novel idea, which is yet to be implanted for heterogeneous data available in the World Wide Web. As integrating the real world data, unlike other integration engines or tools, using this integrating engine can be able to recognize data existing in various formats, and also identifies the similarity words, sets and other forms of semantics involved in them. We can use Word Net to authenticate and validate the contents of data extracted. This tool which will be developed can merge and integrate data with similarity meaning and syntax. This tool once developed can sure be helpful to the organization to make better decisions for their marketing future trends.

References

[1] J. D. Ullman. "Information Integration Using Logical Views". International Conference on Data Technology, pp. 19–40. 1997.

[2] J. Zabin, A. Jefferies. Social media monitoring and analysis: Generating consumer insights from online conversation. Aberdeen Group Benchmark Report, January 2008.

[3] C. Koch . Data Integration against Multiple Evolving Autonomous Schemata, Thesis, 2001.

[4] "Rapid Architectural Consolidation Engine – The enterprise solution for disparate data model", Innovative Routines International, Inc., 2011.

[5] M. M. Kwakye. "A Practical Approach To Merging Multidimensional Data Models", Thesis, 2011.

[6] L. Li, Y. Liu, A. Obregon, M. A. Weatherston "Visual Segmentation-Based Data Record Extraction from Web Documents," IEEE International Conference on Information Reuse and Integration, pp. 502--507, 2007.

[7] J. L. Hong, "Data Extraction for Deep Web using WordNet," IEEE Transaction on Systems, Man, and Cybernetics Part C: Application and Reviews, pp. 854-868, 2011.

[8] J. Zabin, A. Jefferies. Social media monitoring and analysis: Generating consumer insights from online conversation. Aberdeen Group Benchmark Report, January 2008.

[9] B. Liu, R. Grossman, Y. Zhai, "Mining data records in Web pages," Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 601-606, 2003.

[10] V. Stoyanov, C. Cardie, J. Wiebe. Multi-perspective question answering using the OpQA corpus. Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pp. 923–930, 2005.