Novel Heterogeneous Data Integration Using KNN Algorithm

Anto Germin Sweeta J
M.Phil Scholar, S.T. Hindu College, Nagercoil – 629002. Manonmaniam Sundaranar University, Tirunelveli 627012, Tamil Nadu, India.

Dr. B. Ramakrishnan
Associate Professor, Department of Computer Science and Research Centre, S.T. Hindu College, Nagercoil– 629002, Tamil Nadu, India.

Abstract – The huge amount of data is produced daily owing to the tremendous growth of information and communication technology. As these data are individual they have no use. In this paper, we propose a heterogeneous data integration framework for multi-source data. Already many authors have done heterogeneous data integration successfully. However, everyone has its own disadvantage. In all the methods developed already has been only given importance to data integration. However, no one checks the genuineness of the data. Consequently, the existing data integration architecture cannot provide an accurate business model to the customers. In this paper, we develop a four-tier multisource data integration framework for Business Intelligence System (BIS). In this, we first solve the data genuineness problem by using pattern mining method. Moreover, the KNN algorithm is used for solve the multisource text classification problem. The experimental result of our proposed system proves that the novel multisource data integration framework yields better results than the already existing methods.

Index Terms – Text Mining, KNN, Text Alignment Method, Classification, BIS.

1. INTRODUCTION

Data mining is the process of extracting hidden information from the huge amount of datasets usually it comes from data warehouses, big organizations, social networks, etc. Nowadays these information's are used for business intelligence [1][2]. These data are heterogeneous and different data format. Moreover, these data are very noisy and irrelevant. All these data are cleaned and integrated as an individual model. This integrated data is used the big organization to take the decision. In recent days the role of data mining is enormous in telecommunication, education, healthcare, weather prediction, etc. In our proposed work different sources of heterogeneous data are integrated and a new business model is developed. All the data from different sources and are of different format. This process is called multisource data integration. Fig1 shows the proposed architecture of the novel heterogeneous data integration framework. Proposed architecture consists of three main modules. Data Genuineness Module, Data Integration Module, and Data classification module. For this multisource data integration, very famous three datasets are used: Facebook, Twitter, and News Channel.

A. *Problem Definition*

Nowadays social networks have become an important aspect in the daily life of humans. Adults and grown-up use social media alike. So, the genuineness of the data is viewed as an important concern. In modern business changes over to internet commerce platforms. Many products are marketed online. For business purpose some organization misuse the current marketing strategy by given fake reviews and ratings and forwarding it's to others. So, the online data genuineness is the major issue. The chief aim of this research is to develop a genuine multisource data integration framework.

The remainder of this paper is organized as follows: Section II describes the literature review. Section III presents the proposed methodology in detail. Section IV introduces the experimental results and Section V concludes the paper.

2. LITERATURE REVIEW

Jinga et al, 2008 developed the Study of Integration of Multi-Sources Heterogeneous Data Based on the Ontology, [3]. This paper addresses the issue remained in the spatial information technology field which is the heterogeneous data integration. Ontology concept was implemented to understand the integration of various data in the semantic side and solve the diverse semantic issue. Ontology followed 3 important roles while integration of data. They are which defines the fundamental idea and its connection; ontology defines the data sources and makes it be transparency; relations among the data sources are inherited which improves the efficiency while handling the large databases.

Michał Chromiak and Marcin Grabowiecki, 2015 published Heterogeneous Data Integration Architecture-Challenging Integration Issues [4]. Differential large data sources are residing everywhere and data processing system finds out to be hard enough to deal with it. Current fashion seeks a huge

amount of data in any way. Present systems of data integration seem to be unfriendly inviting problems. So the flaws of data integration design and architecture models are reconsidered and bring out new solutions for an efficient workout. Meta-model along with Fast Access Method (FAM) is implemented. It's the query retrieves the data in the fastest possible way from huge differed data sets. Meta-model represents the patterns of integration.

Paulo R. S. Costa et al developed Towards Integrating Online Social Networks and Business Intelligence [5]. The corporate domain is interested in the rising usage of Online Social Networks. Sectors involved in sentimental and social network analysis to find the relationship among the people and their process of decision making. Proposed the OSNBIA, an architecture dedicated to business intelligence to overcome lacking factors in integration. Twitter data is extracted and go through link mining and opinion mining by business intelligence techniques and the result is stored in the data warehouse. It gives new options to deal with data of OSN.

Mokrane Bouzeghoub and et al published Heterogeneous Data Source Integration and Evolution [6]. Data integration plays an important role in furnishing the heterogeneous data which allows users to define queries on them unknowingly. The mediator helps in apprehensions the user requirements and maps between distributed data sources and mediation schema. GAV and LAV mapping approaches are introduced. GAV and LA express each object as a query, that is, view. Both of them allow rewriting process which is the transformation of user-defined queries. GAV addresses the problem of generation, integration, and maintenance of mediation schema views. LAV's flexibility reduces the reformation of queries.

Xiang Ji, 2014 proposed Social Data Integration and Analytics for Health Intelligence[7]. This paper proposed the framework for social data analytics based on the healthcare system. Social media data distributes netizens' experiences and opinions towards the medical field and conditions. Medical care system and researchers would highly beneficial from this. The framework has 3 components as data integration: SPARQL model, predictive analytics: neighborhood pickup function; top-N selection method and population analytics: batch processing concept starts from data preprocessing and ends with analysis result representation. Its a thumbs up to a framework for medical people to have effective results.

Wael Shehab and et al, 2016 developed ROHDIP: Resource Oriented Heterogeneous Data Integration Platform [8]. The uprising of social network data from sites like Instagram, Facebook, etc. led to a vast amount of heterogeneous data and its corresponding stands and architectures. Such data can be in any possible database formats. ROHDIP delivers even access and integrates data from beyond the level. It universally assurances querying the reconciliation stage from wherever and whenever and getting the result. ROA gets any query outcome measure on an assortment of distributed data sources accomplishing the base response time by utilizing HTTP protocol dependent Restful service.

3. PROPOSED METHODOLOGY

In this research, we introduce a novel framework for text mining. The name of the framework is heterogeneous data integration. It is shown in fig.1. The heterogeneous data integration is integrating data of different sources and extracting information from them. Our proposed framework is a dynamic text integration application using this we can integrate any number of text sources. In our proposed method 3 different types of text data source have been used. For this Facebook, Twitter, and BBC News channel data set are the data that include business information. Using data mining methods we have extracted important information (tags, business keywords, etc) form the data sets. All these tags are business related wordings. In our proposed method there are three main modules: Data Genuineness Module (Text pre-processing), Data Integration Module and Data Classification.

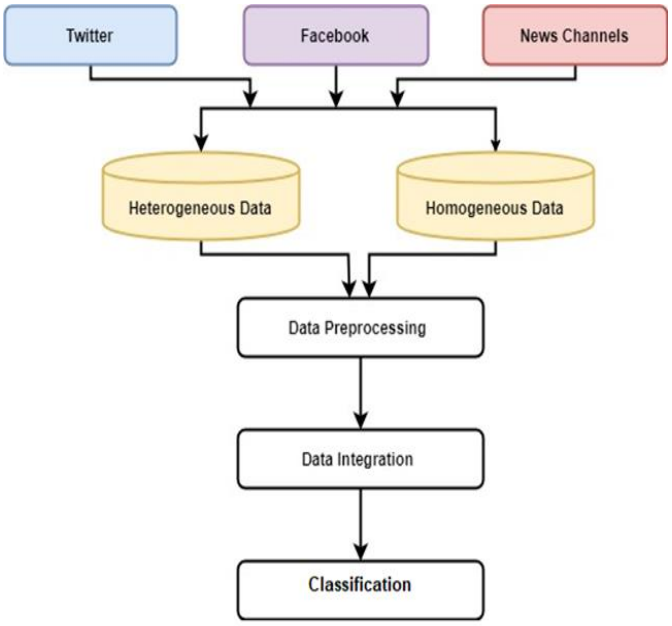

Fig1 Architecture of Proposed Model

A. Data pre-processing

1) In social networks, different types of data are posted every day. Based on the current technology anybody can post any data. For example, when a new product comes to the market the customers are defrauded that they are convinced using fake reviews. The chief aim of our proposed method is to select a genuine dataset before solving the text alignment problem. Hence we can provide a genuine business intelligent system to the customer. In our proposed method to filter the fake review pattern mining is used. The pattern mining filters the fake

review is a repetition of the same catchword, repetition of recommendation by the same customer, the frequency of the review, length of the sentence of the review, etc [9][10]. The filtration process comprises the following sub-processes: Configure the tokenizer and specify a stop words list.

2) **Configure the tokenizer**: This process converting the sentence into a set of features using a pattern mining algorithm.

3) **Specify a stop words list:** In this process, we want to remove the stop words before mining. Commonly used words are: ("and", "a", "of", "I", "you", "it"). Stop words removal reduce memory usage while filtering the datasets.

B. Data integration

The objective of this module is to merge different sources of texts efficiently. For text integration, we combine two methods that we have already used. First, the met path based method and the String-based similarity method. In our proposed architecture met path method is used to align the text and String-based similarity method is used for matching similar keywords. Met path based method working based on string similarity. In this method, P is the sequence of relation and R1, R2….Rn is the relational constraints. In String-based similarity-based method, two types of string similarity functions are used: token based similarity and character based similarity function. String-based similarity method two strings are taken and their edit distance is calculated and similar texts are identified.

C. Data Classification

In our proposed framework KNN is used for classifying the different business domains [11][12]. To classify these important features are taken and stored as a training set. Based on this training set it classified the data that we have already aligned. Fig2 explain the working process of KNN.

k-Nearest Neighbor
Classify $(\mathbf{X}, \mathbf{Y}, x)$ // $\mathbf{X}$: training data, $\mathbf{Y}$: class labels of $\mathbf{X}$, x: unknown sample
for $i = 1$ **to** m **do**
  Compute distance $d(\mathbf{X}_i, x)$
end for
Compute set I containing indices for the k smallest distances $d(\mathbf{X}_i, x)$.
return majority label for $\{\mathbf{Y}_i$ where $i \in I\}$

Fig 2 KNN Algorithm

4. FRAMEWORK EVALUATION

To develop the proposed method J2EE and MySQL are used. J2EE is used to develop front end and MySQL is used to store data set. A computer Intel I5 processor with 4GB ram and 400GB disc space to develop this. For the purposes of this paper, this section presents the results of framework testing over the sets of 500 online documents from different categories in the learning phase and classification. Evaluation of Framework was performed in several test phases:

- Speed text classification
- Classification sensitivity according to categories of documents

In this section, the already existing methods also have been compared. Our experimental results prove that the proposed method is better than the already existing methods. Table1 explain the comparison details with existing methods. Fig 3 shows the graphical representation of the performances with the existing methods.

Table 1 Performance comparison

Method	Speed (Milli Second)	Accuracy (%)
Proposed Method	0.12	83
Ontology Method[3]	0.21	81
HDIA Method[4]	0.19	79
OSNBIA[5]	0.83	78.3
SPARQL Model [7]	0.71	77.1
ROHDIP Model [8]	0.62	76.4

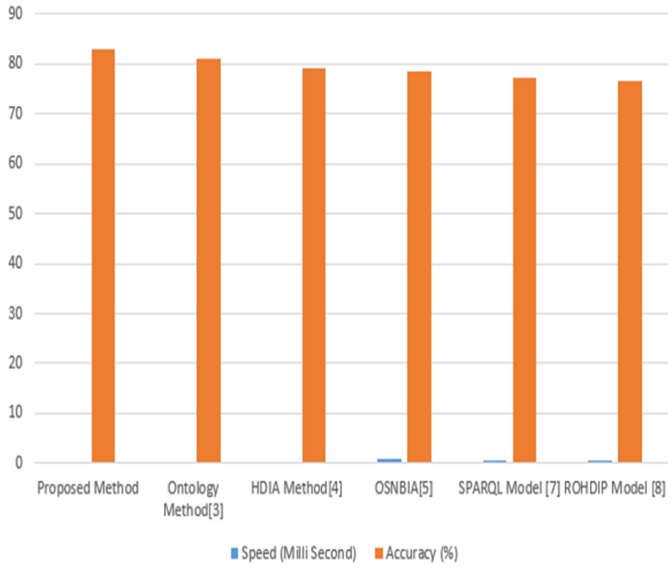

Fig 3 Performance comparison chart

5. CONCLUSIONS

In this paper, we have explained how multisource data can be integrated efficiently. The important problem in text mining is data integration problem. We have proved this integration process using real-time experiment. For the same, we have collected three types of real-world data sets: Twitter, Facebook, and News channel. Using the method we have developed its important features have been extracted. Based on experiment results our proposed method gives an accurate and efficient data mining model. It became a solution to many problems of business intelligence.

REFERENCES

[1] Xindong Wu; Xingquan Zhu; Gong-Qing Wu; Wei Ding, "Data mining with big data", IEEE Transactions on Knowledge and Data Engineering (Volume: 26, Issue: 1, Jan. 2014).

[2] Shivam Agarwal, "Data Mining: Data Mining Concepts and Techniques", 2013 International Conference on Machine Intelligence and Research Advancement.

[3] LUO Jinga, DANG An-rong, and, MAO Qi-Zhi, "The Study of Integration of Multi-Sources Heterogeneous Data Based on the Ontology," in The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B2. Beijing 2008.

[4] Michal Chromiak and Marcin Grabowiecki, "Heterogeneous Data Integration Architecture-Challenging Integration Issues," AI XV, 1 (2015) pp. 7 – 11, DOI: 10.1515/umcsinfo-2015-0001.

[5] Paulo R. S. Costa, Fernando F. Souza, Valéria C. Times, and Fabrício Benevenuto, "Towards Integrating Online Social Networks and Business Intelligence,"

[6] Mokrane Bouzeghoub, Bernadette Farias Lóscio, Zoubida Kedad, and Assia Soukane, "Heterogeneous Data Source Integration and Evolution," DEXA 2002, LNCS 2453, pp. 751-757, 2002.

[7] Xiang Ji, "Social Data Integration and Analytics for Health Intelligence," in the proceedings of the VLDB 2014.

[8] Wael Shehab, Sherin M. ElGokhy, and ElSayed Sallam, "ROHDIP: Resource Oriented Heterogeneous Data Integration Platform," in the proceedings of (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 9, pp. 104-109, 2016.

[9] O. Jamsheela; Raju G., "Frequent itemset mining algorithms: A literature survey", 2015 IEEE International Advance Computing Conference (IACC).

[10] V. Kavitha; B. G. Geetha, "Review on high utility itemset mining algorithms", 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave).

[11] Shweta Taneja; Charu Gupta; Kratika Goyal; Dharna Gureja, "An Enhanced K-Nearest Neighbor Algorithm Using Information Gain and Clustering", 2014 Fourth International Conference on Advanced Computing & Communication Technologies.

[12] Anjali Ganesh Jivani, "The Novel k Nearest Neighbor Algorithm", 2013 International Conference on Computer Communication and Informatics.