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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Multi-Dimensional data Extraction using Iterative Dichotomiser 3 Algorithm

Ramyashree, Prathibha

Dept. of Computer Science, RVCE College, Bangalore, India

Dept. of Computer Science, RVCE College, Bangalore, India

ABSTRACT: in this paper proposed Multi dimensional data extraction Unstructured data like texts, documents, or SNS messages has been increasingly being used in many applications rather than structured data consisting of simple numbers or characters. Thus it becomes more important to analysis unstructured text data to extract valuable information for users decision making. Multidimensional analysis is a data analysis process that groups data into two categories: data dimensions and measurements. For example, a data set consisting of the number of wins for a single football team at each of several years is a single-dimensional (in this case, longitudinal) data set. A data set consisting of the number of wins for several football teams in a single year is also a single-dimensional (in this case, cross-sectional) data set. A data set consisting of the number of wins for several football teams over several years is a two-dimensional data set.

KEYWORDS: OLAP, Multi dimensional analysis, text cube, data cube, text database, information retrieval, TE-IDF

I. INTRODUCTION

As the amount of data grows very fast inside and outside of an enterprise, it is getting important to seamlessly analyze both of them for getting total business intelligence. The data can be classified into two categories: structured and unstructured. Especially, as most of valuable business information is encoded in the unstructured text documents including Web pages in Internet, we need a specialized Text OLAP solution to perform multidimensional analysis on text documents in the same way as on structured relational data. Since the technologies of text mining and information retrieval are major technologies handling text data, we first review the representative works selected for demonstrating how they can be applied for Text OLAP. And then, we survey the representative works selected for demonstrating how we can associate and consolidate both unstructured text documents and structured relation data for obtaining total business intelligence. Finally, we present architecture for a total business intelligence platform incorporating structured and unstructured data. We expect the proposed architecture, which integrates information retrieval, text mining, and information extraction technologies all together as well as relational OLAP technologies, would make an effective platform toward total business intelligence.

Online analytical processing (OLAP) technology is generally used for multidimensional analysis of a vast amount of data from many perspectives [7]. In this paper, we call the multidimensional analysis on text documents using OLAP technology as Text OLAP. For Text OLAP we employ three major text handling technologies: Text Mining (TM), Information Retrieval (IR), and Information Extraction (IE). From a given document set, a TM system does such as extracting top keywords, summarizing, classifying or clustering documents; an IR system retrieves the ones containing the keywords given as a user query; an IE system extracts the structured information according to the schema given by a user.

DOI: 10.15680/IJIRCCE.2016. 0406239

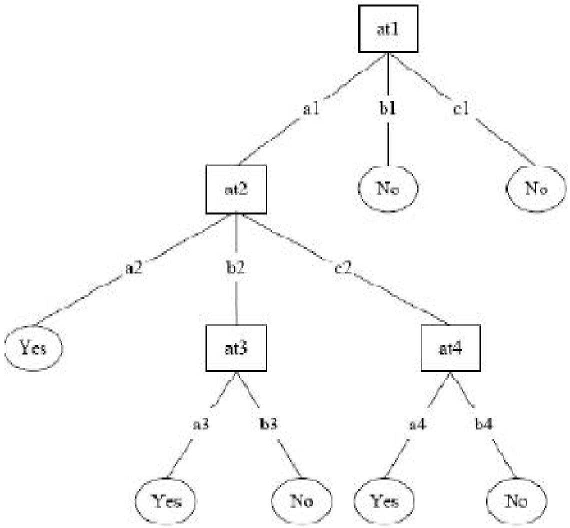

Figure 1: A typical decision tree that contains the attribute a1 and decision values b1 and c1

A set of text documents contains semantically rich and valuable information within them. We have two challenging works here: one is how to do a multidimensional analysis on the vast amount of information themselves contained in the documents, and the other how to connect the information contained in the documents to the data stored in the relational database for doing a multidimensional analysis in a consolidated way. For this purpose, we need a total business intelligence platform that integrates TM, IR, and IE technologies as well as relational OLAP technology. We describe the representative Text OLAP approaches based on TM, IR, and IR technology respectively, and discuss the research issues for obtaining total business intelligence.

II. LITERATURE SURVEY

Maha Azabou et al [1] this Method proposes for on-Line Analytical Processing (OLAP) has generated methodologies for the analysis of structured data. However, it is not appropriate to handle document content analysis. Because of the fast growing of this type of data, there is a need for new approaches abling to manage textual content of data. Generally, these data exist in XML format an approach of construction of our Diamond multidimensional model, which includes semantic dimension to better consider the semantics of textual data In addition, we propose new aggregation operators for textual data in OLAP environment.

M. Catherine McCabe et al [2] described a method of searching text collections that takes advantage of hierarchical information within documents and integrates searches of structured and unstructured data. Thiere show that Multidimensional databases (MDB), designed for accessing data along hierarchical dimensions, and are effective for information retrieval. It demonstrates a method of using On-Line Analytic Processing (OLAP) techniques on a text collection. This combines traditional information retrieval and the slicing, dicing, drill-down, and roll-up of OLAP. And also demonstrate use of a prototype for searching documents from the TREC collection.

Duo Zhang et al [3] Proposes for a Method as the amount of textual information grows explosively in various kinds of business systems, it becomes more and more desirable to analyze both structured data records and unstructured text data simultaneously. While online analytical processing (OLAP) techniques have been proven very useful for analyzing and mining structured data, they face challenges in handling text data. On the other hand, probabilistic topic models are among the most effective approaches to latent topic analysis and mining on text data also its propose a new data model called topic cube to combine OLAP with probabilistic topic modeling and enable OLAP on the dimension of text data in a multidimensional text database. Topic cube extends the traditional data cube to cope with a topic hierarchy and store probabilistic content measures of text documents learned through a probabilistic topic model. To materialize topic cubes efficiently, a heuristic method to speed up the iterative EM algorithm for estimating topic models by leveraging the models learned on component data cells to choose a good starting point for iteration.

Experiment results show that this heuristic method is much faster than the baseline method of computing each topic cube from scratch. We also discuss potential uses of topic cube and show sample experimental results

Biswadeep Nag et al [4] Approaches Multi-dimensional data analysis and online analytical processing are standard querying techniques applied on today's data warehouses. Data mining algorithms, on the other hand, are mostly run in stand-alone, batch mode on flat files extracted from relational databases. In this paper we propose a general querying model combining the power of relational databases, SQL, multi-dimensional querying and data mining. Using this model allows data mining to leverage much of the extensive infrastructure that has already been built for data warehouses including many of the highly successful query processing strategies designed for OLAP. We present an integrated, chunk-based caching scheme that is central to the design of an interactive, multi-dimensional data mining system and conclude with an experimental evaluation of three different cache replacement algorithms.

Vedika Gupta et al [5] describe structured data and unstructured data are handled as two distinct information entities. This often results in failure of decision management as information embedded in unstructured data (USD) can play a vital role in making business decisions for the fact being that around 80% of information resides in unstructured format in the organizations. So there is a need of a framework that meaningfully relates and integrates structured and unstructured data and would act as a total data warehouse (TDW) that may serve as the foundation business intelligence. This way data would be treated purely as a bunch of information for gaining business insight, irrespective of the structure of the data

III. PROPOSED SYSTEM

Below figure shows the overview of text cube model. Here considered 3 dimensional databases. Multi-dimension is represented as hierarchical structure. For example keywords like name of person, city, hotel name, id, etc are represented in single dimension but keywords like date, country, etc are represented as hierarchical structure. User can give query for keyword information according to dimension.

a. TF-IDF

Tf-idf stands for term frequency-inverse document frequency, and the tf-idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the tf-idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query.

One of the simplest ranking functions is computed by summing the tf-idf for each query term; many more sophisticated ranking functions are variants of this simple model.

Tf-idf can be successfully used for stop-words filtering in various subject fields including text summarization and classification.

TF:

Term Frequency, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization.

TF (t) = (Number of times term t appears in a document) / (Total number of terms in the document).

"A high weight in TF–IDF requires a combination of a high term frequency and a low frequency of documents that contain the term among the whole collection of documents; the weights hence tend to filter out common terms." TF-IDF counts different words to provide independent evidence of similarity. One of the advantages of using the TF-IDF method is that it does not require large training data sets in order to distinguish between various documents. By representing a document through a vector space model computed via TF-IDF, comparing a document to other documents or queries is simply achieved through the application of a similarity function.

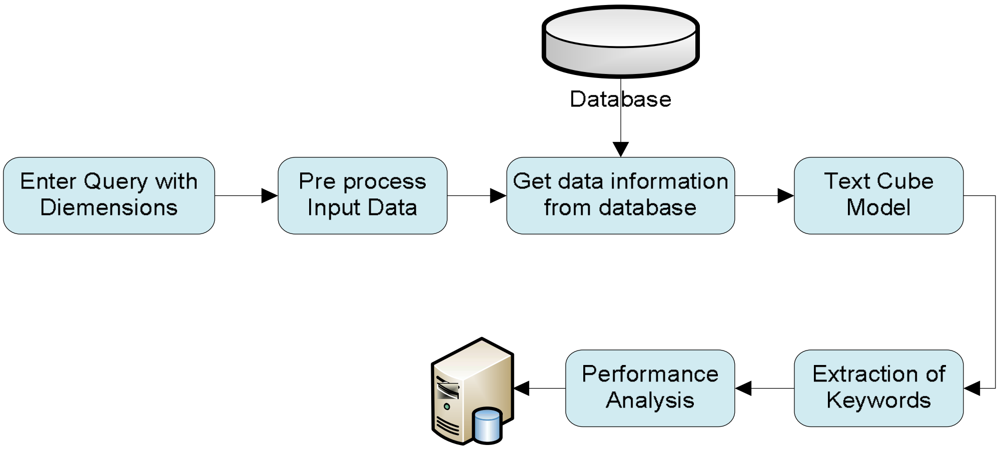

Fig 1. Architecture of Proposed System

Term weight function

The most common forms for the weights of the term are binary and TF-IDF.

$$U_{id} = \begin{cases} 0 & c(d,t) = 0 \\ 1 & c(d,t) > 0 \end{cases}$$

Where $\mathbf{c}(\mathbf{d}, t_i)$is the number of occurrences of term ti in document d

b. ID3 Decision Tree

Decision tree learning is a procedure for calculating the target value having discrete function. The function that has been learned is symbolized by a decision tree. For the inductive inference the decision tree learning is one of the most commonly and broadly used methods which are practical in nature [1][3]. The decision tree learning algorithms are mainly used because of the three reasons:

1. Decision tree is a good infer from the particular cases that are unobserved instance.
2. The calculations in these methods are efficient and are proportional to the instances that are observed.
3. At the final, the decision tree which is produced is easily understood by the human.

IV. RESULTS

Tree based decision tree helps in extracting query data from the database. The whole database is reformed hierarchically. And in each level dimensional values are matched. Based on the matched result it would select the child node as decision. As shown in figure 2.

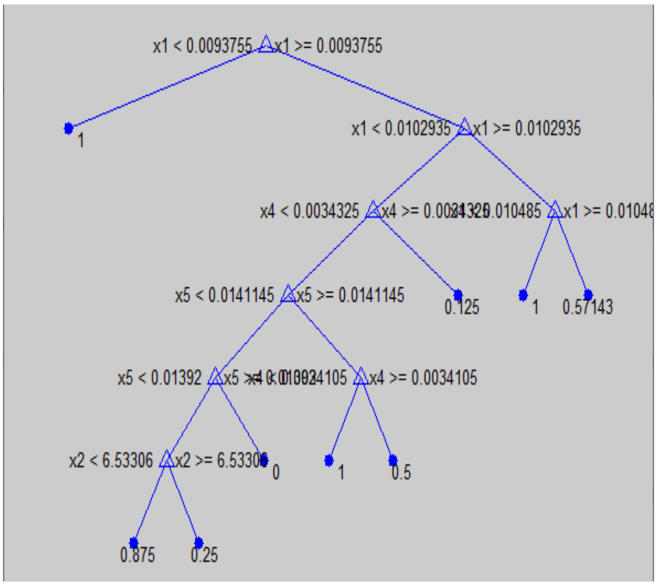

Figure 3: Regression tree viewer

Performance is evaluated by computing execution time. Both TF-IDF and our proposed ID3 algorithm are compared. Our decision tree is giving more efficient result compared to TF-IDF based data extraction.

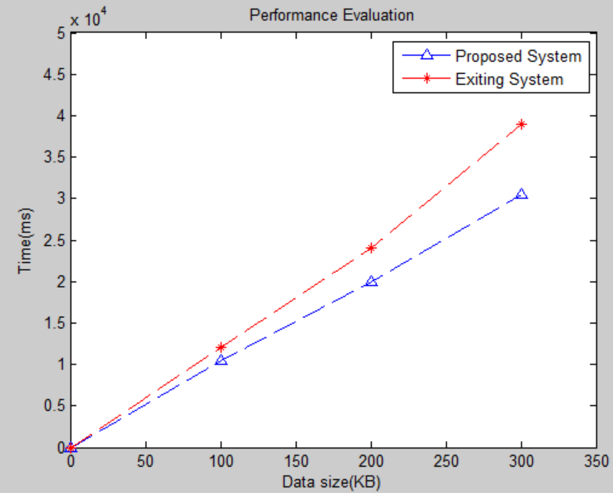

Figure 4: Comparison graph of TF-IDF and Proposed ID3 based approach. Where x-axis represents the data size to be extracted and y axis represents the time taken extract the data in milliseconds

DOI: 10.15680/IJIRCCE.2016. 0406239

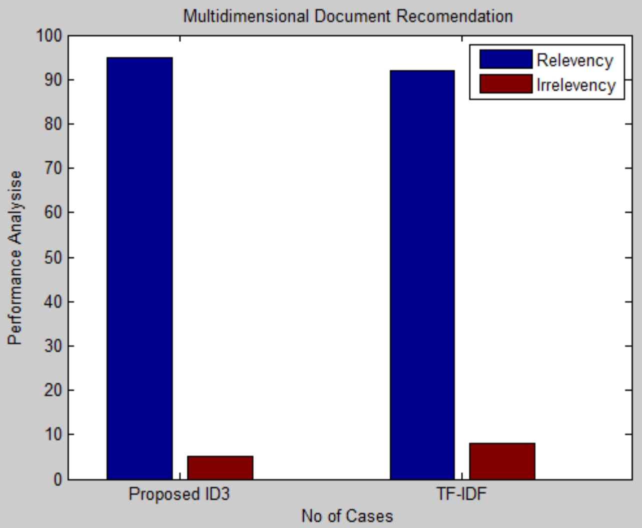

Figure 5: Graphs shows the accuracy and error rate of both existing and proposed methods. Accuracy is computed based on relevancy of the data that user obtained as a result

V. CONCLUSION

According to this paper conclude every organization knows they have rapidly expanding amounts of data. The challenge for users is to transform the knowledge inside that data into competitive advantages. Many people know that information is power; the ability to understand the buying patterns within a customer base can make a huge difference to the corporate bottom line. However, that transformation process can only happen if users can access data in the correct manner and format.

To allow users to intelligently query and analyze information organizations spend large amounts of time and resources adding structure and order to their relational database schemas. This structure is intended to simplify the analysis process for end users. The zenith of this path to structure and order is the multi-dimensional model. It allows user to access a data warehouse schema and quickly, simply and easily transform data into quality information. It also makes users self-supporting as their interaction with the database is done using business orientated language and terminology they use every day in their place of work.

REFERENCES

[1] Maha Azabou, Kaïs Khrouf, Jamel Feki, Chantal Soulé-Dupuy, Nathalie, “Diamond multidimensional model and aggregation operators for document OLAP”.

[2] M. Catherine McCabe, Jinho Lee, Abdur Chowdhury, David Grossman, Ophir Frieder, “On the Design and Evaluation of a Multi-dimensional Approach to Information Retrieval”.

[3] Duo Zhang Chengxiang Zhai Jiawei Han, “Topic Cube: Topic Modeling for OLAP on Multidimensional Text Databases”.

[4] Prof. Pramod Patil, Prini Kotian, Aishwarya Gaonkar, Sachin Wani, Pramod Gaikwad,” Map-Reduce for Cube Computation”. International Journal of Scientific Research Engineering & Vol 4 Issue 4, April 2015

[5] Maha Azabou, Kaïs Khrouf, Jamel Feki, Chantal Soulé-Dupuy, Nathalie Vallès, “Diamond multidimensional model and aggregation operators for document OLAP”.

[6] Byung-Kwon Park, Hyoil Han, and Il-Yeol Song, “XML-OLAP: A Multidimensional Analysis Framework for XML Warehouses”.

[7] Dattatray V. Meshram, D. M. Dakhane, “Unstructured Multidimensional Array Multimedia Retrival Model Based Xml Database”, IJRET: International Journal of Research in Engineering and Technology.

[8] Anand Bahety, “Extension and Evaluation of ID3 – Decision Tree Algorithm”. Wei Peng, Juhua Chen and Haiping Zhou, “An Implementation of ID3 --- Decision Tree Learning Algorithm”.

[9] Cindy Xide Lin, Bolin Ding, Jiawei Han, Feida Zhu, Bo Zhao, “Text Cube: Computing IR Measures for Multidimensional Text Database Analysis”.

[10] Dewi Puspa Suhana Ghazali1, Rohaya Latip1, 2, Masnida Hussin1 and Mohd Helmy Abd Wahab, “A Review Data Cube Analysis Method in Big Data Environment” Vol. 10, No 19, 2015.