ISSN 2320-9984 04 >
9 772320 998015

Volume 1, Issue 4, December 2013 ISSN: 2320-9984 (Online)

International Journal of Modern Engineering & Management Research

Website: www.ijmemr.org

Natural Language Processing Based Article Mining From Web Portal

Priyanka Patel
M.Tech Scholar
Department of Computer Science and Engineering
Shri Ram Institute of Science and Technology,
Jabalpur (M.P.) [INDIA]
Email: priyanka.patel789@gmail.com

Prof. Mahendra Rai
Professor & Guide,
Department of Computer Science and Engineering
Shri Ram Institute of Science and Technology,
Jabalpur, (M.P.) [INDIA]

Abstract—*The Human language Query processing will play vital role in computer interaction. It is a part of Artificial Intelligence which has information retrieval, Machine translation and Analysis. The main aim of Human Language Query Processing (HLQP) is to enable the novice user to interact Database by avoiding the complex command and function. The Conventional Database systems are responsible for the storage and processing of huge amounts of information. The data stored in these database systems refers to information valid at present time. High-level query languages such as SQL are available in commercial Databases. These are easy for those users with thorough understanding of programming concepts, database schema and relational algebra.*

DUE to the rapid growth of digital data made available in recent years, knowledge discovery and data mining have attracted a great deal of attention with an imminent need for turning such data into useful information and knowledge. Over the years, people have often held the hypothesis that phrase-based approaches could perform better than the term based ones, as phrases may carry more "semantics" like information. With a large number of patterns generated by using data mining approaches, how to effectively use and update these patterns is still an open research issue. In this work, we focus on the development of a knowledge discovery model to effectively use and update the discovered patterns and apply it to the field of text mining.

Keywords:— *HLQP, Data Mining, Pattern Discovery, SQL, Databases, Artificial Intelligence, Natural Language Processing*

1. INTRODUCTION

A. Natural Language Processing & its Application in Mining

The Human language Query processing will play vital role in computer interaction. It is a part of Artificial Intelligence which has information retrieval, Machine translation and Analysis. The main aim of Human Language Query Processing (HLQP) is to enable the novice user to interact Database by avoiding the complex command and function. This Human language Query processing make the people easy to learn and use the computer as well. This will make the user to enter the text message as they would pass to the person. The interactive with computer is very essential and also more effective. Nowadays computerization is implemented in almost all the fields. Particularly in Medical Field if the Doctor wants to interact with Database, he should know the complex command as well as procedure. But this Human Language Query processing made everyone to access the Database easily.

The Conventional Database systems are responsible for the storage and processing of huge amounts of information. The data stored in these database systems refers to information valid at present time. The conventional Database does not provide models to support and process the past and future data. The Temporal database stores data relating to time instances. It offers Temporal Data types and stores information related to past, present and future time. In Temporal Database the time period is added to express when it should be valid and when it is stored.

A Database that can store and retrieve temporal data, that is, data that depends on time in some way, is termed as a Temporal Database. The Conventional Database is generally two dimensional, and contains only current data. The two dimensions are rows and columns that interact with each other at cells containing particular value whereas temporal databases are three-dimensional with time interval as the third dimension. Temporal Databases can also be referred to as time-oriented Databases, time varying databases, or historical databases. A true temporal database is a bi-temporal database that supports both valid time and transaction time.

Transaction time is the actual time recorded in the database at which the data is entered and the time is known as the Timestamp. Time-stamps can include either only the date or both the date and clock time. Time-stamps cannot be changed. The other major type in Temporal Database is the valid time. Valid time is the actual or real world time at which point the data is valid. Conventional Databases represent the state of an enterprise at a single moment of time. The conventional database holds the snapshot data.

There is a growing interest in applying Database methods for version control and design management in e-commerce applications, requiring capabilities to store and process time dependent data. Moreover, many applications such as Medical Diagnosis System, Forest Information Systems, Weather Monitoring Systems and Population Statistics Systems have been forced to manage temporal information in an adhoc manner and support the storage and querying of information that varies over time temporal database holds time varying information, required by the above-mentioned applications. In the present scenario, writing better database queries for databases pertaining to an organization involves a significant amount of time and expertise. It has become a research issue now to increase the service capability of the database systems to help novice users to formulate a query for database access.

High-level query languages such as SQL are available in commercial Databases. These are easy for those users with thorough understanding of programming concepts, database schema and relational algebra. To help non-expert users to perform query, a natural language front end is required. For those users who feel SQL difficult to use and for novice users who would like to retrieve data without having to learn querying mechanism such as SQL, a temporal natural language querying mechanism has been provided to access data from temporal databases. The Natural Language Interface helps the distribution of the thought process from the human query users to the system. Doing so helps reducing the effort spent by the query users in forming the queries.

B. Text Mining & Data Mining

DUE to the rapid growth of digital data made available in recent years, knowledge discovery and data mining have attracted a great deal of attention with an imminent need for turning such data into useful information and knowledge. Many applications, such as market analysis and business management, can benefit by the use of the information and knowledge extracted from a large amount of data. Knowledge discovery can be viewed as the process of nontrivial extraction of information from large databases, information that is implicitly presented in the data, previously unknown and potentially useful for users. Data mining is therefore an essential

step in the process of knowledge discovery in databases.

In the past decade, a significant number of data mining techniques have been presented in order to perform different knowledge tasks. These techniques include association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. Most of them are proposed for the purpose of developing efficient mining algorithms to find particular patterns within a reasonable and acceptable time frame.

With a large number of patterns generated by using data mining approaches, how to effectively use and update these patterns is still an open research issue. In this work, we focus on the development of a knowledge discovery model to effectively use and update the discovered patterns and apply it to the field of text mining.

Text mining is the discovery of interesting knowledge in text documents. It is a challenging issue to find accurate knowledge (or features) in text documents to help users to find what they want. In the beginning, Information Retrieval (IR) provided many term -based methods to solve this challenge, such as Rocchio and probabilistic models, rough set models, BM25 and support vector machine (SVM) based filtering models.

Over the years, people have often held the hypothesis that phrase-based approaches could perform better than the term based ones, as phrases may carry more "semantics" like information. This hypothesis has not fared too well in the history of IR. Although phrases are less ambiguous and more discriminative than individual terms, the likely reasons for the discouraging performance include:

1. Phrases have inferior statistical properties to terms,
2. They have low frequency of occurrence, and
3. There are large numbers of redundant and noisy phrases among them.

Classification of articles as a Text: As number of stories are there in a daily newspaper, the users would like to see the stories in which names of different persons are involved for some place, organization etc. Manually doing such a task is a tedious job, text mining method i.e. Information extraction can be used to perform such a task, which would retrieve templates containing different entities and their relationship with one another in the structured format, which would be put in the databases, on which Data Mining techniques can be applied for retrieving the interesting patterns.

A. Demerits a/ Text Mining

(i) No programs can be made in order to analyse the unstructured text directly, to mine the text for information or knowledge.

(ii) The information which is initially needed is no where written.

B. Merits a/Text Mining

(i) As database can store less amount of information, this problem has been solved through Text Mining.

(ii) Using the technique such as information extraction, the names of different entities, relationship between them can easily be found from the corpus of documents set.

(iii) Text mining has solved the problem of managing such a great amount of unstructured information for extracting patterns easily; otherwise it would have been a great challenge.

2. EXISTING SYSTEM

Nowadays interaction with computer is essential, effective process and also the storing and retrieving of data from database will play vital role in the database application. To access the Database the user should have a strong knowledge in SQL command and procedures. But this is not possible for all users. So in this we present Human Language Query Processing for Temporal Database. This will help the novice user to interact Temporal Database in their Native language (English), without using any SQL command or procedures.

In this paper Human Language Query Processing for Temporal Database has been designed and implemented to access Temporal Database. This lets the novice user to formulate their queries in their native language. The main purpose of this system is focused for Medical domain, but this is a generalized system i.e. it also supports Population system, Accounting System, Banking System, etc. In this system we used Temporal Database, as it is a time varying database we can formulate the historical data and also the data validity.

Classic approaches to test input generation – such as dynamic symbolic execution and search-based testing – are commonly driven by a test adequacy criterion such as branch coverage. However, there is no guarantee that these techniques will generate meaningful and realistic inputs, particularly in the case of string test data. Also, these techniques have trouble handling path conditions involving string operations that are inherently complex in nature.

This paper has presented an approach for generating values for String data types by using tailored web searches, dynamic regular expressions and NLP techniques. The empirical study showed that the valid values can be obtained using the approach. Another benefit is that the generated values are also realistic rather than arbitrary-looking – as often the case with the most automatic test data generation techniques. This is because the values are obtained from the Internet which is a rich source of human-recognizable data. When an automated oracle is non-existent, the test cases using such values help in reducing human-oracle cost [23] in terms of time and effort involved in interpreting results. More empirical evidence is required to hold the claim that is planned in the future work.

The quality of decisions made in business and government relates directly to the quality of the information used to formulate the decision. This information may be retrieved from an organization's knowledge base (Intranet) or from the World Wide Web. Intelligence services Intranet held information can be efficiently manipulated by technologies based upon either semantics such as ontologies, or statistics such as meaning-based computing. These technologies require complex processing of large amount of textual information. However, they cannot currently be effectively applied to Web-based search due to various obstacles, such as lack of semantic tagging. A new approach proposed in this paper supports Web-based search for intelligence information utilizing evidence-based natural language processing (NLP).

This paper presented a new framework for Web-based intelligence information acquisition and formation of a textual knowledge base. The major strength of this framework lies in the combination of existent NLP techniques, grounded theory and evidential analysis to automatically extract unknown unknowns from Web-based textual content and form a knowledge base that can be effectively manipulated by analysts to find facts (names) and associations between them (events).

The proposed similarity estimation has provided encouraging results in comparing large amounts of texts due to a higher frequency of word-concept co-occurrence, making it possible to disambiguate a sense that each word has within its context. Extracting the word sense will allow manipulation with distributional profiles of concepts that contain measures for strength of association between

each word used in each of its senses (categories from the thesaurus) co-occurring with other categories, i.e. strength of association between concepts only rather than concepts and words.

Web portals are a major class of web-based content management systems. They can provide users with a single point of access to a multitude of content sources and applications. However, further analysis of content brokered through a portal is not supported by current portal systems, leaving it to their users to deal with information overload. We present the first work examining the integration of natural language processing into web portals to provide users with semantic assistance in analyzing and interpreting content. This integration is based on the portal standard JSR286 and open source NLP frameworks. Two application scenarios, news analysis and biocuration, highlight the feasibility and usefulness of our approach.

In this paper, we presented a novel technical solution and interaction patterns for the integration of NLP tools with portal technology, in order to provide semantic assistance to users of web portals. Our approach allows users to benefit from a broad spectrum of various NLP services, such as named entity extractors, summarizers, indexers, and others. Users can use these services on a multitude of content and applications delivered through modern portals.

Natural language processing relies heavily on resources. Most common usage scenarios include using the resources for automated lexical tagging or named entity recognition. Also manually annotated language resources are used for benchmarking of new automated approaches. To allow any processing on a large scale and considering the complexity of natural language (words can have multiple meanings within the same general context) the resources have to be quite large. In this paper we focus on lexical resources in ontology form.

Statistical natural language processing relies heavily on resources. Most common usage scenarios include using the resources for automated lexical tagging or named entity recognition. Also manually annotated language resources are used for benchmarking of new automated approaches. To allow any processing on a large scale and considering the complexity of natural language (words can have 3 or more possible meanings within the same general context) the resources have to be quite large. Current lexical resources are created in many different formats. This problem was previously addressed with the Text Encoding Initiative. However using of semantic technologies provides many advantages for lexical resources.

This paper describes the modules of natural language processing (NLP) engine which can be used with Hungarian input. There are many standard NLP engines which have tokenization, part-of-speech (POS) tagging, named entity recognition, parsing modules. Most of them work for universal languages like English. Processing of Hungarian language is a much more difficult and there cannot be found such a complete NLP system which satisfies all tasks for syntactic and semantic analysis of incoming inputs. This paper summarizes the existing solutions and techniques and gives a brief description of a proposed NLP engine using for inputs in Hungarian language.

The paper provides a detailed system requirement analysis on the internal structure of an NLP engine. The survey of existing solutions and proposals shows that the combination of standard and language dependent modules provides a realizable framework. The pilot system based on the presented model is in the test phase of the implementation.

3. PROPOSED WORK

This work proposes to perform text mining for web documents and main key area being used to show the implementation and

results is based on natural language processing techniques from the web portals.

This work will use mainly indexers and entity extractors & summarizers using frequency and pattern matching for performing mining of the text and will be used for classification of the articles from the web portal.

This work can be applied on fields of web resource cataloguing, finding user's area of interest and documents separations etc.

The proposed work is offering to perform classification of articles using natural language processing techniques as follows:

Step 1: Loading web documents in RAM for processing.

Step 2: Human language query processing technique for reading the user inputs, performing grammar check and using dictionary search for use of correct words in the query inputted by the user.

Step 3: Tokenization & Lexical Analysis of the words shall be done using String Tokenizer to find the relevant words.

Step 4: Semantic Analyser will be created to convert the query in the form of subjects and categories of the articles.

Step 5: Data Collection: various articles will be collected on the basis of the frequencies and availability of the tokens in the dataset of the articles.

Step 6: Data Purification: In this step mechanism of stopping and stemming shall be applied to perform preprocessing of the data and filter data for useful words available in the documents.

Step 7: Analysis of the proposed work using results and graphs. Following parameter shall be used to generate and compare the results:

- Accuracy of the Articles Classification
- Time Taken in processing
- Size of the data and documents involved in the system

4. CONCLUSION

The project work implementation will have the following screens:

- Creating Various Screens for user interfacing
- Loading available articles datasets for text mining into the project
- Taking user inputs and performing HLP techniques on it
- Tokenization and Lexical Analysis of the user input
- Semantic Analysis for checking grammar of the user inputs
- Data Collection & Purification
- Analysis and comparison of the results with the existing work

REFERENCES:

[1] K.Murugan, T. Ravichandran, "Human Language Query Processing in Temporal Database using Semantic Grammar", IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM -2012) March 3 0, 3 1, 2012, ISBN: 978-81-909042-2-3 ©2012 IEEE

[2] Muzammil Shahbaz, Phil McMinn, Mark Stevenson, "Automated Discovery of Valid Test Strings from the Web using Dynamic Regular Expressions Collation and Natural Language Processing", 2012 12th International Conference on Quality Software, 1550-6002 © 2012 IEEE DOI 10.1109/QSIC.2012.15

[3] Natalia Danilova, David Stupples, "Application of Natural language

Processing and Evidential Analysis to Web-Based Intelligence Information Acquisition" 2012 European Intelligence and Security Informatics Conference, 978-0-7695-4782-4 © 2012 IEEE DOI 10.1109/ EISIC.2012.41

[4] Fedor Bakalov, Bahar Sateli, Ren´e Witte, Marie-Jean Meurs, Birgitta K¨onig-Ries, "Natural Language Processing for Semantic Assistance in Web Portals" 2012 IEEE Sixth International Conference on Semantic Computing, 978-0-7695-4859-3 © 2012 IEEE, DOI 10.1109/ ICSC.2012.38-

[5] Sandi Pohorec, Ines _eh, Milan Zorman, Marjan Mernik, and Peter Kokol, "Natural Language Processing Resources: Using Semantic Web Technologies", Proceedings of the ITI 2012 34th, Int. Conf. on Information Technology Interface, June 25-28, 2012, Cavtat, Croatia, doi:10.2498/iti.2012.0386

[6] P. Barabás, L. Kovács, "Requirement Analysis of the Internal Modules of Natural Language Processing Engines ", SAMI 2012 10th IEEE Jubilee International Symposium on Applied Machine Intelligence and Informatics January 26-28, 2012 Herl'any, Slovakia, 978-1-4577-0197-9/ © 2011 IEEE