ISSN 2320-9984

Volume 4 Issue 1 | March 2016

ISSN: 2320-9984 (Online)

International Journal of Modern Engineering & Management Research

Website: www.ijmemr.org

Natural Language Processing & Ant Clustering Based Information Retrieval from Temporal Databases

Kavindra Markam
Research Scholar
Shree Ram Group of Institutions
Jabalpur (M.P.), [INDIA]
Email: kavir.mark@gmail.com

Anupam Choudhary
Lecturer,
Kalaniketan Polytechnic College
Jabalpur (M.P.), [INDIA]
Email:chowdharyanupam7@yahoo.com

Sapna Choudhary
Assistant Professor,
Department of Computer Science and Engineering.
Shree Ram Group of Institutions
Jabalpur (M.P.), [INDIA]
Email: choudharysapnajain@gmail.com

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. 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.

In this work, a new algorithm is being proposed which will allow increasing the efficiency of the natural language processing through application of ANT Clustering, which is expected to give best performance due to application of ANTs for searching and matching of the tokens. Since ANTs will work in parallel because they will be applied using multithreaded environment therefore the overall efficiency of the newly proposed system shall be much better than the existing works.

Keywords:*— Natural Language Processing, Ant Clustering, Temporal Databases, Human Language Query Processing.*

1. INTRODUCTION

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.

2. ANT CLUSTERING

Data clustering, or just clustering, is an explorative task that seeks to identify groups of similar objects based on the values of their attributes. Clustering works on the inherent characteristics of the data and attempts to discover distinct boundaries to divide the data set into meaningful partitions. Deneubourg et al. proposed a basic model which generalized the clustering behavior of ants into two simple actions: i) picking up an isolated item, and ii) dropping the item where more similar items are present. Assuming the ants or agents can only handle one item at a time and only one type of item exists in the environment, they defined each action in terms of a probabilistic function.

As such, the item will be moved until the agent reaches a more dense region. Deneubourg et al.'s model was extended by Lumer & Faieta to include a distance function, d between data objects, hence removing the need for assuming type homogeniety and making it more generalized for the purposes of exploratory data analysis. They provided an algorithm, which models the inherent similarity of data objects onto a lower dimensional space, e.g. 2-D. As in a connectionist's self-organizing map, such mapping would be able to capture the similarity relationships between higher-dimensional objects and project it on a 2-D grid.

Other similar work includes the AntCluss clustering algorithm, which is a combination of an ant colony with the partitional K-Means algorithm. The ant colony of AntClass differs from Lumer & Faieta's model as ants are allowed to carry more than a single object at a time, have local memory and other heterogeneous features.

Ant-based techniques, in the computer sciences, are designed for those who take biological inspirations on the behavior of these social insects. Data-clustering techniques are classification algorithms that have a wide range of applications, from Biology to Image processing and Data presentation. Since real life ants do perform clustering and sorting of objects among their many activities, we expect that a study of ant colonies can provide new insights for clustering techniques.

Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset share some common trait. Clustering is used as a data processing technique in many different areas of application, such as bioinformatics, data mining, image analysis, etc.

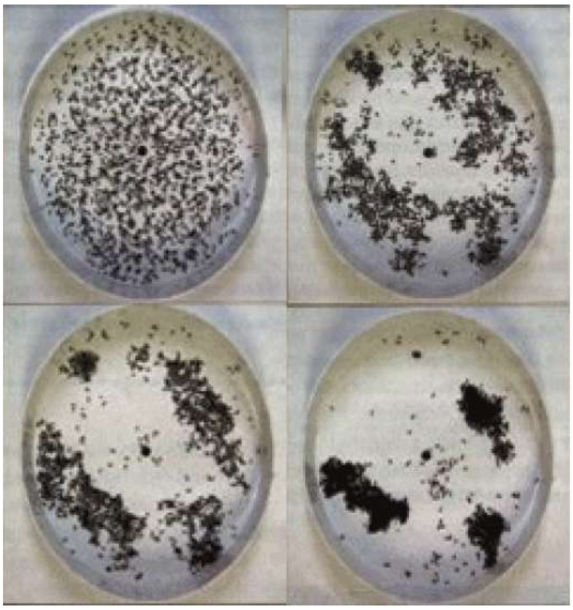

Figure 1: Ant colony (Messaor Sancta)

Figure shows a real ant colony (Messor Sancta) performing clustering task with 1500 ant corpses over a period of 26 hours. Besides this patchwork sorting, the Leptothorax ant has been observed to cluster its brood into concentric annular structures, with eggs and micro-larvae in the center, and larger off-springs located in the outer side of the structure.

3. TEMPORAL DATABASE

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.

4. EXISTING WORK

From the various papers studied and analysed from IEEE, not much of the work has been done in the field of natural language processing, its problems and their solutions. Descriptions of the a few papers are as follows:

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. The conventional Database will give only current data not past or future data. But the Temporal data will support for past, present and future data. In temporal data we used third axis as time interval, which support both Transaction time as well valid time. The valid time is the actual or real world time at which the data is valid. The main aim of this system is that the human language is interpreted with Temporal Database and to produce appropriate results. This system is implemented in Java which can be used in any platform [1].

In this paper Hwnan 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 [2].

This paper presents a novel approach of finding valid values by collating suitable regular expressions dynamically that validate the format of the string values, such as an email address. The regular expressions are found using web searches that are driven by the identifiers appearing in the program, for example a string parameter called email Address. The identifier names are processed through natural language processing techniques to tailor the web queries. Once a regular expression has been found, a secondary web search is performed for strings matching the regular expression [2].

A topic-dependent-class (TDC)-based -gram language model (LM) is a topic-based LM that employs a semantic extraction method to reveal latent topic information extracted from noun-noun relations. A topic of a given word sequence is decided on the basis of most frequently occuring (weighted) noun classes in the context history through voting. Our previous work (W. Naptali, M. Tsuchiya, and S. Seiichi, “Topic-dependent language model with voting on noun history,” ACM Trans. Asian Language Information Processing (TALIP), vol. 9, no. 2, pp. 1–31, 2010) has shown that in terms of perplexity, TDCs outperform several state-of-the-art baselines, i.e., a word-based or class-based -gram LM and their interpolation, a cache-based LM, an n -gram-based topic-dependent LM, and a Latent Dirichlet Allocation (LDA)-based topic-dependent LM. This study is a follow up of our previous work and there are three key differences [3].

A TDC is a topic-dependent LM with unsupervised topic extraction employing semantic analysis and voting on nouns. We demonstrated that a TDC with soft clustering and/or soft voting in the training and/or test phases improved performances. Soft clustering solved the unreliable topic mapping while soft voting solved the shrinking data problem in the TDC. Soft clustering in the TDC should be performed in both the training and test phases. Soft voting yielded a larger improvement compared with soft clustering. Soft voting performed in only one phase (either the training or test phase) also produced good results [4].

We also demonstrated that incorporating a cache-based LM improved the TDC further. The cache-based LM helped the TDC capture an aspect of the language that was not covered, such as increasing the probability of co-occurring words. The evaluation of perplexity showed that the TDC achieved a 25.1% relative reduction in perplexity for the English corpus and a 25.7% relative reduction for the Japanese corpus compared with the baseline [4].

Natural language interfaces to ontologies allow users to input their queries in natural language to the system and retrieve their desired information from ontologies. Many natural language interfaces have been developed to date but their capability in handling negation queries is limited [5].

This paper proposes a negation query handling engine which is particularly designed to handle user queries with negation. The proposed engine is designed to understand the complexity of natural language queries with negation which was previously not catered effectively by the existing systems. The proposed engine effectively understands the intent of the user query on the basis of a sophisticated algorithm which is governed by a set of techniques and transformation rules [5].

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 [6].

5. PROBLEM STATEMENT

Nowadays, most of information saved in companies is as unstructured models. Retrieval and extraction of the information is essential works and importance in semantic web areas. Many of these requirements will be depend on the storage efficiency and unstructured data analysis. Merrill Lynch recently estimated that more than 80% of all potentially useful business information is unstructured data. The large number and complexity of unstructured data opens up many new possibilities for the analyst. We analyze both structured and unstructured data individually and collectively. Text mining and natural language processing are two techniques with their methods for knowledge discovery form textual context in documents.

Natural language processing (NLP) and related semantic technologies promise to support users in analyzing, transforming, and creating knowledge from large amounts of content. However, it is an open question how exactly these technologies can be combined with existing information system infrastructure like web portals, in a way that brings measurable improvements to their users.

The work done by the authors in the base paper suggests that increasing the size of the Dictionary and Grammar rules would decrease the efficiency. Global dictionary can be introduced for various domains. Further research in this will enhance for the complex queries and all types of Joins. The works done in the area of natural language processing will also be facing the problem of decreased efficiency as the data will increase which is an unbound phenomenon.

6. PROPOSED WORK

This work proposes a new algorithm based on natural processing and mining of the data from databases using various natural language processing algorithms and ant classification and mining. The various steps involved in the proposed work are:

Step 1: Load database of the data.

Step 2: Prepare dictionary of data words for extraction and matching of information from the natural language text inputted by the user.

Step 3: Creating an interface for user input.

Step 4: Dictionary Search and Grammar rules application for correctness of the user input.

Step 5: Tokenization of the user inputs.

Step 6: Creating ANTs for each token to find the existence in the data sets created in steps 1 and 2.

Step 7: Retrieve the occurrences of the tokens and use them to create the data base query for processing and retrieving user required information.

Step 8: Measuring performance of the proposed application and

comparing with the work in base paper.

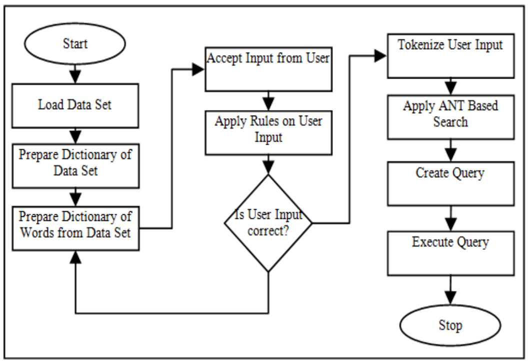

Figure 2: Flow Chart showing the flow of processing in proposed work

7. RESULTS & DISCUSSION

In this work, a new algorithm is being proposed which will allow increasing the efficiency of the natural language processing through application of ANT Clustering, which is expected to give best performance due to application of ANTs for searching and matching of the tokens. Since ANTs will work in parallel because they will be applied using multithreaded environment therefore the overall efficiency of the newly proposed system shall be much better than the existing works.

This work proposes to enhance the speed of processing and in turn enhanced efficiency by the introduction of data mining along with the natural language processing technique. This work uses ANT clustering mechanism to perform the searching and matching in the dataset in the various intermediate stages of the applications of natural language processing application so that the behavior of ANTs shall be used for fast speeds.

In this work, a new algorithm is being proposed which will allow increasing the efficiency of the natural language processing through application of ANT Clustering, which is expected to give best performance due to application of ANTs for searching and matching of the tokens. Since ANTs will work in parallel because they will be applied using multithreaded environment therefore the overall efficiency of the newly proposed system shall be much better than the existing works.

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 30, 31, 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] Welly Naptali, Masatoshi Tsuchiya, and Seiichi Nakagawa, “Topic-Dependent-Class-Based n-Gram Language Model”, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 5, JULY 2012, Digital Object Identifier 10.1109/TASL.2012.2183870, 1558-7916 © 2012 IEEE

[4] Rizwan Iqbal, Masrah Azrifah Azmi Murad, Mohd Hasan Selamat, Azreen Azman, “Negation Query Handling Engine for Natural Language Interfaces to Ontologies”, 978-1-4673-1090-1/12 © 2012 IEEE

[5] 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