Creating Data Warehouse For Natural Language Processing

[1]Namita Arora, Jaspreet Kaur Sahiwal, [2]Sanjeev Kumar Sharma
[1]LPU, Jalndhar
[2]B.I.S College of Engineering and Technology, Moga – 142001, India

Abstract: *Organizations be it industry or business or even educational institutes, need to improve their information inventory system so as to survive in the competitive environment. The organizations have to increase their efficiency and effectiveness in maintaining the cycle of activities, in their planning, decision-making processes, and analytical needs. There are several ways to acquire this goal; one of it is with data mining which is able to make a prediction using existing data in their database in order to forecast future demand. All most all the NLP application are based on the data mining techniques. So there is need to apply the data warehouse technique in natural language processing field.*

Introduction

A data warehouse (DW) can be defined as a database used for reporting as well as for analysis. The data stored in the warehouse is uploaded from the operational systems. A data warehouse is a central repository for all or significant parts of the data that an enterprise's various

business systems collect. Data warehousing is combining data from multiple and usually varied sources into one comprehensive and easily manipulated database. Common accessing systems of data warehousing include queries, analysis and reporting.

Because data warehousing creates one database in the end, the number of sources can be anything you want it to be, provided that the system can handle the volume, of course. The final result, however, is homogeneous data, which can be more easily manipulated and from which the useful information con be easily extract. Usually a data warehouse is either a single computer or many computers (servers) tied together to create one giant computer system. Data consists of raw data or formatted data. It can be on various types of topics including the organization's sales, salaries, operational data, summaries of data including reports, copies of data, human resource data, inventory data, external data to provide simulations and analysis, etc. In addition to a relational database, a data warehouse environment includes an extraction, transportation, transformation, and loading (ETL) solution, an online analytical processing (OLAP) engine, client analysis tools, and other applications that manage the process of gathering data and delivering it to business users.

Data Warehousing Schemas

A schema is a collection of database objects, including tables, views, indexes, and synonyms. You can arrange schema objects in the schema models designed for data warehousing in a variety of ways. Most data warehouses use a dimensional model. The model of your source data and the requirements of your users help you design the data warehouse schema. You can sometimes get the source model from your company's enterprise data model and reverse-engineer the logical data model for the data warehouse from this.

Star Schemas

The star schema is the simplest data warehouse schema. It is called a star schema because the diagram resembles a star, with points radiating from a center. The center of the star consists of one or more fact tables and the points of the star are the dimension tables. The most natural way to model a data warehouse is as a star schema, only one join establishes the relationship between the fact table and any one of the dimension tables. A star schema optimizes performance by keeping queries simple and providing fast response time. All the information about each level is stored in one row.

Operational Database System Versus Data Warehouses

Since most people are familiar with commercial relational database systems, it is easy to understand what a data warehouse is by comparing these two kinds of systems. The major task of online operational database systems is to perform online transaction and query processing. These systems are called online transaction processing (OLTP) systems, and cover most of the day to- day operations of an organization such as purchasing, inventory, manufacturing, banking, payroll, registration, and accounting. Data warehouse systems, on the other hand, serve users or\knowledge workers in the role of data analysis and decision making. Such systems can organize and present data in various formats in order to accommodate the diverse needs of the different users. These systems are known as online analytical processing (OLAP) systems. The major distinguishing features between OLTP and OLAP are:

(i) Users and system orientation: An OLTP system is customer-oriented and used for normal transaction like issue of book, and query processing by librarians, teachers, and information technology professionals. An OLAP system is more of planning and decision-making oriented and is primarily used by knowledge workers including managers,
Executives, and analysts.

(ii) Data contents: An OLTP system manages current data that, typically, are too detailed and used to run the day-to-day business of the institution library. An OLAP system manages large amounts of historical data, provides facilities for summarization and aggregation, and stores and manages information at different levels of granularity. These features make the data easier for use in informed decision making.

(iii) Database design: An OLTP system usually adopts an entity-relationship (ER) data model and an application-oriented database design. An OLAP system typically adopts either a star or snowflake model and a subject-oriented database design.

(iv) View: An OLTP system focuses mainly on the current data within an enterprise or department, without referring to historical data or data in different organizations. In contrast, an OLAP system often spans multiple versions of a database schema, due to the evolutionary process of an organization. OLAP systems also deal with information that originates from different organizations, integrating information from many data stores. Because of their huge volume, OLA data are stored on multiple storage media.

(v) Access patterns: The access patterns of an OLTP system consist mainly of short, atomic transactions. Such a system requires concurrency control and recovery mechanisms. However, accesses to OLAP systems are mostly read-only operations (since most data warehouses store historical rather than up to-date information), although many could be complex queries. Other features which distinguish between OLTP and OLAP systems include database size, frequency of operations, and performance metrics.

Natural Language Processing

Natural Language Processing (NLP) is the computerized approach to analyzing text that is based on both a set of theories and a set of technologies and, being a very active area of research and development.

Goal

The goal of NLP as stated above is "to accomplish human-like language processing". The choice of the word 'processing' is very deliberate, and should not be replaced with 'understanding'. For although the field of NLP was originally referred to as Natural Language Understanding (NLU) in the early days of AI, it is well agreed today that while the goal of NLP is true NLU, that goal has not yet been accomplished. A full NLU System would be able to:

1. Paraphrase an input text
2. Translate the text into another language
3. Answer questions about the contents of the text
4. Draw inferences from the text

While NLP has made serious inroads into accomplishing goals 1 to 3, the fact that NLP systems cannot, of themselves, draw inferences from text, NLU still remains the goal of NLP. There are more practical goals for NLP, many related to the particular application for which it is being utilized. For example, an NLP-based IR system has the goal of providing more precise, complete information in response to a user's real information need. The goal of the NLP system here is to represent the true meaning and intent of the user's query, which can be expressed as naturally in everyday language as if they were speaking to a reference librarian. Also, the contents of the documents that are being searched will be represented at all their levels of meaning so that a true match between need and response can be found, no matter how either are expressed in their surface form. Origins

As most modern disciplines, the lineage of NLP is indeed mixed, and still today has strong emphases by different groups whose backgrounds are more influenced by one or another of the disciplines. Key among the contributors to the discipline and practice of NLP are: Linguistics - focuses on formal, structural models of language and the discovery of language universals - in fact the field of NLP was originally referred to as Computational Linguistics; Computer Science - is concerned with developing internal

representations of data and efficient processing of these structures, and; Cognitive Psychology - looks at language usage as a window into human cognitive processes, and has the goal of modeling the use of language in a psychologically plausible way. Divisions

While the entire field is referred to as Natural Language Processing, there are in fact two distinct focuses – language processing and language generation. The first of these refers to the analysis of language for the purpose of producing a meaningful representation, while the latter refers to the production of language from a representation. The task of Natural Language Processing is equivalent to the role of reader/listener, while the task of Natural Language Generation is that of the writer/speaker. While much of the theory and technology are shared by these two divisions, Natural Language Generation also requires a planning capability. That is, the generation system requires a plan or model of the goal of the interaction in order to decide what the system should generate at each point in an interaction. We will focus on the task of natural language analysis, as this is most relevant to Library and Information Science.

Natrual Lanaguage Processing Applications

Natural language processing provides both theory and implementations for a range of applications. In fact, any application that utilizes text is a candidate for NLP. The most frequent applications utilizing NLP include the following:

• **Information Retrieval** – Given the significant presence of text in this application, it is surprising that so few implementations utilize NLP. Recently, statistical approaches for accomplishing NLP have seen more utilization.

• **Information Extraction (IE)** – A more recent application area, IE focuses on the recognition, tagging, and extraction into a structured representation, certain key elements of information, e.g. persons, companies, locations, organizations, from large collections of text. These extractions can then be

utilized for a range of applications including question-answering, visualization, and data mining.

• **Question-Answering –** In contrast to Information Retrieval, which provides a list of potentially relevant documents in response to a user's query, question-answering provides the user with either just the text of the answer itself or answer-providing passages.

• **Summarization –** The higher levels of NLP, particularly the discourse level, can empower an implementation that reduces a larger text into a shorter, yet richly constituted abbreviated narrative representation of the original document.

• **Machine Translation –** Perhaps the oldest of all NLP applications, various levels of NLP have been utilized in MT systems, ranging from the 'word-based' approach to applications that include higher levels of analysis.

• **Dialogue Systems –** Perhaps the omnipresent application of the future, in the systems envisioned by large providers of end-user applications. Dialogue systems, which usually focus on a narrowly defined application (e.g. your refrigerator or home sound system), currently utilize the phonetic and lexical levels of language. It is believed that utilization of all the levels of language processing explained above offer the potential for truly habitable dialogue

Utilization of Data Warehouse in NLP

Almost all the NLP applications are based upon the statistical techniques. The statistical techniques require a huge amount of information in the form of data. For example if we have to develop a POS tagger based on the statistical technique and using viterby algorithm then we will be required to have a huge amount of data available to get the transition probability etc. now such a huge data can be collected from a no of sources which may include internet, books and magazines etc. also this type of data is collected by different persons from different locations. So there is need to store this

data at a single location. Also there are chances of ambiguity in such type of data. So we proposed a new system to store this type of data in a data warehouse. So that it could be directly used to develop such applications. Also the data warehouse can be fed from different sources and from different locations. As per our knowledge no such type of efforts have been done till date.

Conclusions and Future Work

We used SQL server 2005 and visual studio 2008 for developing a data warehouse and we filled it with more than one lakh of words extracted from four difference file sources. This can be further enhanced by adding more data in the form of grammatical information of the words and other different type of data that is in the NLP applications

References

1. Abrahams P. W. et al. “The LISP 2 Programming Language and System”, in proceedings of FJCC, No. 29, USA, 1966, pp. 661– 676.
2. T. Amble, “BusTUC - A Natural Language Bus Route Oracle.” 6th Applied Natural Language Processing Conference, Seattle, Washington, USA, 2000.
3. A. Arun and F. Keller, “Lexicalization in Cross linguistic Probabilistic Parsing “, in Proceedings of the 43rd Annual Meeting of the ACL, 2005, pp. 306 – 313.
4. R. Elmasri, and S. Navathe, (2007). “Fundamentals of Database System”, 5th ed. Addison Wesley, USA.
5. B. Grosz, A. Joshi and S. Weinstein, “Providing a unified account of definite noun phrases in discourse” In Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics, Cambridge, MA USA, 1983, pp. 44 – 50.
6. K. Hamilton and R. Miles “Learning UML 2.0. O’Reilly”, ISBN-10: 0-596-00982-8, 2006.

7. G. Hendrix, “The LIFER manual A guide to building practical natural language interfaces”, SRI Artificial Intelligence Center, Menlo Park, Calif. Tech. Note 138, 1977.
8. G. Hendrix, E. Sacrdoti, D. Sagalowicz, and J. Slocum, “Developing a natural language interface to complex data”, ACM Transactions on Database Systems, Volume 3, No. 2, USA, 1978, pp. 105 – 147.
9. G. Luger and W. Stubblefield, “Artificial Intelligence Structures and Strategies for Complex Problem Solving”, 3rd ed. Addison-Wesley, USA, 1999.
10. J. McCarthy, “LISP Programmers Manual, Handwritten Draft” MIT AI Lab., Vambridge, USA, 1959.
11. H. Nogami, Y. Yoshimura, and S. Amano, “Parsing with look-ahead in real-time on-line translation system”, Research and Development Center Toshiba Corporation Kawasaki-City, Japan Volume 1, 1989, pp. 488 –493.
12. M. Palmer, T. Finin, “Workshop on the Evaluation of Natural Language Processing Systems”, Computational Linguistics, MIT Press, Volume 16, USA, 1990, pp 175 – 181.
13. S. Sinan, “Guide To applying The UML. Springer-Verlag, New York, Inc, USA, 2002, ISBN 0-387-95209-8.
14. nmon, W.H. Buliding the dataware house. QED Technical Publishing Group, Wellesley,Massachutusetts.
15. Chaudhuri, S. & Dayal U. An overview of data warehousing and OLAP technology. SIGMOD Record 26:1, 1997, March 65-74.
16. Han, J. & Kamber, M. Data mining concepts and techniques, edited by Morgan Kaufmann, V. Harinarayan, A. Rajaraman, J.D.Ulman. *In* Implementing Data Cubes Efficiently. Proceedings of SIGMOD Record, **25(**2), 205-16.

* * * * *