International Journal of Information Technology and Business Management
29th July 2014. Vol.27 No.1

ISSN 2304-0777 www.jitbm.com

PREPARATION OF DATA FOR KNOWLEDGE BASE

Marian Švalec[1], Juraj Branický[2], Vladimír Hanušniak[3] and Katarína Zábovská[4]

[1,2,3]Department of Informatics, Faculty of Management Science and Informatics, University of Žilina, Slovakia
[4]Department of Macro and Micro Economics, Faculty of Management Science and Informatics, University of Žilina, Slovakia

{juraj.branicky, vladimir.hanusniak, marian.svalec, katarina.zabovska}@uniza.sk

Abstract

Knowledge bases are developed for the needs of access to consolidated data. They contain data along with their relationship and semantics, in the understandable form for computer, allowing higher degree of automation. Not only data stored in relational databases are of great importance for knowledge bases, also metadata and unstructured data sources need to be processed. This article covers possible approaches towards filling the knowledge base from unstructured data sources using text and data mining methods and also processing of extensive sets of data with aid of decomposition and parallelism.

Keywords: knowledge base, data mining, unstructured data, large volumes of data, data decomposition

INTRODCUTION

People realised their need for storing of information long time ago, in prehistory. With painting on walls in caverns they tried to capture the reality and keep it for the next generations. The spoken language and then hand writing as a mean of recording, exchanging and delivering of thoughts and events was created to fulfill the same need. Large amount of text documents lead to creation of libraries and archives as a centralized storage for recorded information.

As the information technologies were improved and began to be spread in early 60-ties a new form of digital data recording was used. The request of automatic data processing and better usage of data was met by electronical storages. A key criterion for effective data processing is their organization. The selection of storage and the structure of data depends on the aim of specific information system or an application.

Database systems are the most used ones and they are based on relational databases. Such systems allow effective manipulation with relatively large amount of data and after applying a suitable analytical method new valuable information can be revealed. The down side is in the very principle of relational databases, where only the values alone are stored. A conceptual model is used for covering the meaning of these values, because it describes the individual concepts in the domain and their relationships. However, conceptual model covers the meaning of data only partially. The greater part of their semantics is implicitly transferred into the logics of an application. In case of the data analysis process, a database of a specific information system is accessed from an external application and it is of vital importance to know the semantics of accessed data as well. In such situation, the conceptual model is not sufficient. This is why the data processing needs the domain experts and basically, it is impossible to be automated.

In addition, for the complex view of an organization lifecycle it is necessary to explore its data with respect to all actions that are performed within it. If the organization works with diverse applications and

International Journal of Information Technology and Business Management
29th July 2014. Vol.27 No.1

ISSN 2304-0777 www.jitbm.com

information systems, which are focused on different areas and have differently structured data, the access to consolidated data becomes even more complicated. One of the possibilities to solve this problem is the use of knowledge bases, which allow to connect individual data with metadata in a single common concept. The metadata are understood as an expansion of "raw" data and describes their semantics and interconnections. It is very crucial that they are readable by computer. Intelligibility of data for computer allows a higher degree of automation. [1]

KNOWLEDGE BASES

The knowledge base does not represent a concrete software solution, but a principle. It is not the subject of this article to specify a particular solution. It was partly suggested in article [2]. The knowledge base offers comfortable work with integrated and comprehensible data and puts higher demand on the modelling and the data preparation part.

The "raw" data stored in relational databases are no longer sufficient for the needs of knowledge base. As we mentioned above, semantics and relationships between the values are required. A description of their meaning which is based only on consultations with a specialist is inefficient and inadequate. Important thing is to find an automatic methodology that would reveal hidden relationships between the data and hence their semantics by analyzing them. This process is called mining from data or data mining as well.
The notion of data mining includes a series of steps, which may lead to revelation of links between the data and patterns (typical symptoms) of a system. On this basis it is possible to estimate the future behavior of the system or the behavior of the same system under changed conditions. Data mining consists of several stages:

- data selection,
- data preprocessing,
- data transformation,
- data classification and
- interpretation of data.

At the selection stage, only potentially useful data are selected from available sources and repositories, in which we want to look for hidden knowledge. Preprocessing of data consists of completing the incomplete data, corrections of mistakes and inconsistencies and in eliminating duplicates. Within transformation of preprocessed data a series of rules and functions are applied. Data classification is based on applying the algorithms which should provide enough information for the interpretation stage. Interpretation provides answers to modeled problem, or encourages the formulation of new questions and models.

STRUCTURED AND UNSTRUCTURED DATA

Data Mining is usually associated with the processing of structured data that make up one of the two subsets of all data. The second subset is composed of unstructured data. The boundary between structured and unstructured data is not entirely clear, in our article we will stick to the following definitions:
Structured data are data stored in fixed fields within a record or a file. The structured data term is mostly understood as information with certain degree of organization, such as inclusion in a relational database and the ability to be searched through using simple and clear searching algorithms. According to [3], complex and big data requires specific approaches to be implemented for better handling and processing. There are mentioned some commonly used approaches for large database processing such as SQL query optimization, distributed database systems, map-reduce concept and advanced techniques.
Unstructured data are the opposite of structured ones and they form such a wide area that, depending on the source, comprise from 70% to 90% of all data, as for instance mentioned in [4]. Unstructured data have no predefined model and are not organized in any known manner. They usually contain a significant amount of text, but also facts. These assumptions mean that data of this type are difficult to be processed automatically using a computer. In unstructured data we include: audio files, plain text, rich text (text mixed with multimedia - figures, charts, videos), video, figures, email (especially its body), websites.
Whilst the amount of structured data grows approximately linearly, unstructured data are growing in volume almost exponentially.

International Journal of Information Technology and Business Management
29th July 2014. Vol.27 No.1

ISSN 2304-0777 www.jitbm.com

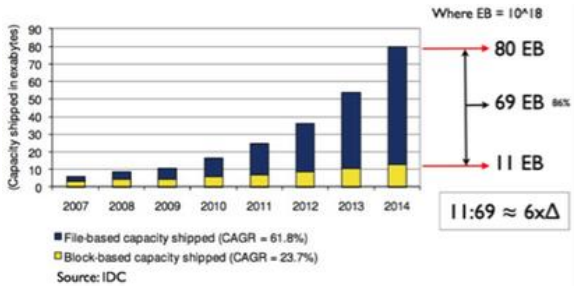

Figure 1: The growth of data in recent years, source: IDC

The graph shows that in 2014, 69 exabytes are expected to be sent in form of unstructured data, that means, almost 7 times more than structured. From the graph it is also evident that the increase in unstructured data, compared with the increase of structured data, is enormous and rises every year. The expected size increase is from 40% to 50% every year.

THE MINING OF UNSTRUCTURED DATA

Unstructured data is a source of knowledge, which has great potential for the knowledge base. They carry a lot of information that is not evident at first sight and in general it is difficult to process them using a computer.

However, there are methods that help us to mine knowledge from such data. In our article we will focus mainly on text and web pages. Data in the form of various audio-visual formats and their importance for the knowledge base will be omitted.

Text varies considerably depending on the specification and not only stylistically, but also linguistically. It contains a lot of confusions (for example synonyms) that human mind is able to differentiate based on context and semantics. Especially the semantics of words and a variety of metadata about the text are essential for the knowledge base.

That is why there are methods that can extract information from unstructured texts. The most known one is text mining and its variations, such as web mining, which mostly deals with websites and works with their specific structure.

The text mining process itself begins with filtering of words that have a little informational value. The remaining terms are transformed and attributes that characterize the text are created. These attributes are formed based on the semantic relationships of individual words in sentences and are already considered to be structured information. As such, they can be processed further - using data mining or can be entered into the database and so on.

Regarding the filling of the knowledge base, a large information potential lies in documentation, articles and publications of domain experts, which contain their knowledge in text form. Using *text mining* they could be processed automatically and by doing so the base would be expanded by other relevant knowledge developed by experts. Papers written by experts contain, except the evident knowledge, also the unique thought of every author. These thoughts, especially, provide valuable additional metadata for knowledge base.

PROCESSING OF LARGE DATA SETS

The unstructured resources and their elaborate processing is just one of the problems associated with the data. Because of nowadays sufficient storage capacity all available data are collected with all the details, in a fine granularity and their analysis is postponed. Thus, there are extensive set of data that conventional analytical methods are not able to process in an acceptable time. Therefore, processing of complex tasks or large amounts of data cannot be done without *parallelism* - processing one task simultaneously on multiple computing nodes. Solving such tasks in terms of parallelism can be divided into two approaches. [5]

First approach is suitable for computationally demanding tasks over a small amount data. Task is divided into several partial sub-tasks. Along with data sub-tasks are processed parallelly on multiple computing nodes. We can say that *the data are moved to the processing*. This approach is called *task decomposition*.

Second approach is called *data decomposition*, for less computationally demanding tasks over large volumes of data. Within the focus of this article, we assume rather large amount of data than computationally demanding algorithms. Here, a significant role plays the speed of the disk. Individual disks, normally used for storage, allow reading with speed of several tens of megabytes per second. Discs connected in disk arrays reach reading speed several times higher,-which is, however, still deficient. This problem can be solved using a distributed file system and a programming model that allows parallel

ISSN 2304-0777 www.jitbm.com

reading and data processing on multiple computing nodes simultaneously. A computing node provides storage capacity and also computing power for data processing. It processes locally stored data, eliminating the overhead of transferring amounts of data for processing. In this case, we can talk about *moving the processing to the data*.

Currently, there are tools that address the problem of storing and processing large volumes of data. The most famous tool is framework *Apache Hadoop.* Its main components are distributed repository HDFS (Hadoop Distributed File System) and MapReduce programming model. HDFS deals with storage of large files, provides scalability and high data availability. MapReduce allows distributed execution of the desired task over the data stored in HDFS. Hadoop can contain from tens to hundreds of computing nodes. Its big advantage is that it works over conventional, affordable hardware, as it is described in [6].

CONCLUSION – THE USAGE OF KNOWLEDGE BASE

The benefits from large amounts of data, which constantly increases and contains great hidden potential, can be maximized by combining approaches mentioned in this article. The transformation of extracted information and knowledge into a common repository, such that a semantic lost is minimized, is the main motivation of our work. The common semantic repository is a good basis for automation of processes and artificial intelligence, which we want to pay attention in the future.

ACKNOWLEDGEMENT

This paper is supported by the following project: University Science Park of the University of Žilina (ITMS: 26220220184) supported by the Research & Development Operational Program funded by the European Regional Development Fund.

REFERENCES

1. Lukasová, A., Habiballa, H., Telnarová, Z., "Formální reprezentace znalostí", V Ostravě: Ostravská univerzita, 2010, ISBN 978-80-7368-900-1
2. Joštiak, M., Švalec, M., Zábovská, K., "Baza znalostí v ontologií", http://www.sssi.sk/ojs/index.php/CasopisSI/
3. Zábovský, M., Matiaško, K., Zábovská, K.: Database Exploration Using Metrics and Visualization, 7th Scientific Conference „Internet in Information Society“ 2012, In: Zastosowania Internetu : praca zbiorowa. - Dąbrowa Górnicza: Wyższa Szkoła Biznesu, 2012. - ISBN 978-83-62897-08-7. - S. 159-168. - (Prace naukowe Wyższej Szkoły Biznesu w Dąbrowie Górniczej)
4. International Data Corporation (IDC) - http://www.idc.com//home.jsp
5. Hadoop: The Definitive Guide, 3rd Edition, O'Reilly Media, Inc., 2012, ISBN 978-1-4493-1152-0
6. Shvachko, K., Kuang, H., Radia, S., Chansler, R., The Hadoop Distributed File System, In Proceedings of MSST2010, May 2010.
7. Mařík, V., Štěpánková, O., Lažanský, J., "Umňelá intelligence 4", Academia, 2003, ISBN 80-200-1044-0