LINGUISTIC ENGINEERING AND ITS APPLICABILITY TO BUSINESS INTELLIGENCE

Towards an Integrated Framework

S. F. J. Otten and M. R. Spruit
Institute of Information an Computer Sciences, Utrecht University, Princetonplein 5, Utrecht, The Netherlands

Keywords: Business intelligence, Linguistic engineering, Social network, Knowledge discovery, Unstructured data, Framework.

Abstract: This paper investigates how linguistic techniques on unstructured text data can contribute to business intelligence processes. Through a literature study covering 99 relevant papers, we identified key business intelligence techniques such as text mining, social mining and opinion mining. The Linguistic Engineering for Business Intelligence (LEBI) framework incorporates these techniques and can be used as a guide or reference for combining techniques on unstructured and structured data.

1 INTRODUCTION

The internet has been evolving from a static source of information to a dynamic and versatile source of information available to everyone. The concept of WEB 2.0 emerged (i.e. Facebook, Twitter) and contains valuable information (Anderson, 2007; Kleinberg, 2008). A limited amount of organizations leverage the knowledge hidden in these social networks due to a lack of knowhow (Pang & Lee, 2008). Business Intelligence (BI) is used to satisfy a manager's request analyzing enterprise data to improve decision making (Willen, 2002). The data is mutated until it reaches the desired form to support strategic decision making (Moody & Kortink, 2000). However unstructured data is in most organizations untapped. But it has to be structured, processed, and analyzed with by using Linguistic Engineering (LE) techniques (och Dag, Regnell, Gervasi, & Brinkkemper, 2005).In this paper we propose a framework that provides insight into how LE can complement BI for achieving a competitive advantage. Hence, the research question of this paper is:

In which ways can linguistic techniques such as text mining, social mining and opinion mining contribute to business intelligence processes?

The remainder of this paper is structured as follows. Section 2 provides a theoretical background concerning business intelligence and linguistic engineering. Section 3 provides the research approach. Within section 4 the framework is presented, and section 5 comprises a discussion and section 6 conclusions.

2 RELATED LITERATURE

In this section we explore the main concepts of our research: Business Intelligence (BI) and Linguistic Engineering (LE).

2.1 Business Intelligence (BI)

BI enables organizations to understand their internal and external environment through the systematic acquisition, collation, analysis, interpretation and exploitation of information (Chung, Chen, & Numumaker Jr, 2003). BI uses the following techniques: data warehouses, OLAP, and data mining. A data warehouse comprises data acquired from multiple structured data sources on an operational (Inmon, 2002; Sen & Sinha, 2005; Kotsiantis & Kanellopoulos, 2006). In order to leverage these new insights, an organization needs to have certain goals targets to compare the actual organizational performance. An organization can define Key Performance Indicators (KPIs) (Parmenter, 2007) to do so.

Otten S. and Spruit M..
LINGUISTIC ENGINEERING AND ITS APPLICABILITY TO BUSINESS INTELLIGENCE - Towards an Integrated Framework.
DOI: 10.5220/0003661704520456
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2011), pages 452-456
ISBN: 978-989-8425-79-9

2.2 Linguistic Engineering (LE)

Linguistic engineering is concerned with the computational processing of unstructured text and speech in order to derive knowledge from it. A technique used within the automated document summarization process is text mining. Another popular application of LE in recent years, with the growing popularity of Social Network Sites (SNS) (Heer & Boyd, 2005), is its ability to derive knowledge from social media via social mining and opinion mining (Hu & Liu, 2004; Yang, Dia, Cheng, & Lin, 2006). Text Mining (TM) was first mentioned by Feldman and Dagan (1995) and deals with the machine supported analysis of text. Before TM can commence a necessity exists to pre-process each text document and store the data in a structured manner. The pre-processing of text in the TM-domain entails (1) *tokenization*; (2) *filtering*; (3) *lemmatization*; (4) *stemming*. (Feldman & Sanger, 2007). Data mining is used with the purpose to classify or cluster documents. A discipline in social mining is called Social Network Analysis (SNA) and comprises the analysis and visualization of an online social network (Erétéo, Buffa, Gandon, Grohan, Leitzelman, & Sander, 2008). The knowledge to be leveraged, with the help of social mining has numerous applications (Adomavicius & Tuzhilin, 2002; Yang, Dia, Cheng, & Lin, 2006; Yang & Dia, 2008). Opinion mining aims to extract attributes and components of the object that has been commented on in a set of text-based content, to determine whether they are positive, negative or neutral and is called *semantic orientation* (Hu & Liu, 2004; Liu, 2007; Ding, Liu, & Yu, 2008). Based on the conducted literature, we understand that today's use of LE—in particular the disciplines of text mining, social mining and opinion mining—is not yet as established as one would hope. In most cases it is not yet implemented in an organization's BI-environment, resulting in a loss of valuable information and knowledge.

3 RESEARCH DESIGN

Our research entails the exploration of the possible uses of Linguistic Engineering in combination with already defined Business Intelligence processes in organizations. Hence, our research question:

"In which ways can linguistic techniques such as text mining, social mining and opinion mining contribute to business intelligence processes?"

The literature referenced in this paper was found through an online literature search, by using Google Scholar and Omega. Table 1 provides an overview of keywords used in search-queries. The keywords were also combined to find literature in a specific context.

Table 1: Keywords.

Keywords			
Subject	Keywords	Publications	Validated Publications
Business Intelligence		28	6
	Business Intelligence	7	
	Data warehouse	8	
	Data mining techniques	5	
	Decision Support Systems	2	
	Strategic Decisions Making	6	
Linguistic Engineering		71	26
	Linguistic Engineering	15	
	Text mining	25	
	Social mining	18	
	Opinion mining	13	
Total		99	32

With each search-query we only analyzed the top 100 items, under the assumption that each search engine lists the most relevant results first. In our search queries we used the keywords with and without quotations; no significant differences in the top 100 items were discovered. Our analyses of the search results were done by checking the title and abstract of each source and quick-scanning the publication for its relevance regarding our research. In total we found 99 relevant scientific publications which could be used in our research. A more detailed analysis of the literature, by fully reading the articles, reduced the total from 99 to 70 relevant publications. From the 70 relevant publications 32 were either empirically or expert validated, 34 were not validated, and 4 could not be determined as "validated" or "not validated". The validated papers all proposed a new technique/algorithm for handling unstructured or structured data. The 34 that were not validated were literature studies, surveys or framework descriptions. The remaining five relevant publications all described the evolution of theory development on several concepts.

4 LEBI FRAMEWORK

From literature we derived six stages (stage I through stage VI) and used them in the LEBI framework. Figure 1 depicts a comprehensive overview of the framework. The difference with existing frameworks and the LEBI-framework is the addition of the processing and analyzing steps for unstructured data.

Stage I – Business Needs: Determine business needs required by management. In order to create and determine the desire reports which server as input for the SDM-process in stage VI.

Stage II – Data Sources: Based on business needs defined in stage I, data sources are identified, containing unstructured data and (possibly) data sources containing structured data (i.e. ERP systems).

Stage III – ETL Procedures: *Extract:* The data has to be extracted from the data source. (APIs). *Transform:* After data has been extracted, the data has to be transformed and cleansed. *Load:* After the transformation activity is completed on both sides one has to load the created data in the data warehouse. Depending on the format in which the new data is stored a choice has to be made on how databases and tables are designed and stored (Inmon, 2002).

Stage IV – Data Warehouse Storage: *Data warehouse development:* The data warehouse has to be developed in order to allow storage of both types of data. *Datamart development*: if required, before developing data cubes, datamarts can be designed which combines data from various databases and or tables for specific analysis purposes. *Cube development:* The development of cubes comprises selecting the right dimensions, measures and the right data.

Stage V – Analysis: *Mining:* in case of unstructured data after development of a particular cube, one needs to analyze the data and turn it into useful information by applying certain algorithms on the data. Mining structured data is done to recognize implicit patterns within explicit data making it explicit. *OLAP-analysis:* OLAP-analysis is mostly used for structured data. With OLAP-analysis the dimensions serve as perspectives and the data being viewed originates from the fact table. *Construct report:* A report is generated comprising a combination of results from both OLAP and DM.

Stage VI – Decision Making: Management uses the report as a tool for making strategic decisions.

5 DISCUSSION

This paper focused on the problem of organizations not being able to utilize available data on the internet. For most organizations it remains an untapped source of valuable data which can be turned into valuable information and knowledge aiding the organization's competitive advantage. In most organizations, where Business Intelligence systems are in place, the data sources being utilized are all characterized as structured data sources and internally available. These systems do not seem to be capable of incorporating unstructured data sources into their analyses. Departing from that observation we propose a framework which can be used as a guide or reference for combining unstructured- and structured data and transforming it into useful knowledge. However, the framework still needs empirical and expert validation due to it being only based on an extensive literature study. As mentioned earlier in section 4, it is also possible to use just one side (structured or unstructured) of the framework as a guide or reference. However, the stages are sequential and cannot be interchanged with one another on both sides.

6 CONCLUSIONS

This paper was aimed at developing a framework which combined disciplines of linguistic engineering and already existing business intelligence disciplines. This implied that the framework should support the extraction, transformation, loading and analyzing of unstructured data as well as structured data. We found that before conducting any analysis the unstructured data should undergo extreme cleansing. We also found that the disciplines text mining, social mining, and opinion mining were most suited for analyzing unstructured data and most complementary to existing business intelligence processes.

Regarding structured data we found a plethora of sub disciplines, methods and techniques for the extraction, transformation, loading, and analysis, due to it being a more grounded object of research. We incorporated the most used processes of structured data handling in our framework (KDD, CRISP-DM). We found that one of the most useful and popular data mining techniques is "association rule mining". Therefore we incorporated this mining technique in

our framework. For further analysis of structured data we choose to incorporate OLAP analysis as it is a widely accepted technique for viewing data alongside different perspectives (dimensions).

We conclude that this framework can already be useful as a structured guide or reference, but leaving ample room for improvement and the implementation of new or existing techniques.

REFERENCES

Adomavicius, G., & Tuzhilin, A. (2002). Using data mining methods to build customer profiles. Computer, 34(2), 74-82.

Anderson, P. (2007). What is web 2.0. Ideas, technologies and implications for education, 60, 1-10.

Chung, W., Chen, H., & Numumaker Jr, J. F. (2003). Business Intelligence Explorer: A knowledge map framework for discovering business intelligence on the Web. *Proceedings of the 36th Annual Hawaii International Conference on System Sciences (pp. 1-10). Honolulu, Hawaii, USA: IEEE.*

Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. *Proceedings of the international conference on Web search and web data mining (pp. 231-240).* Stanford, California, United States of America: ACM.

Erétéo, G., Buffa, M., Gandon, F., Grohan, P., Leitzelman, M., & Sander, P. (2008). A state of the art on social network analysis and its applications on a semantic web. *Proceedings of the 7th International Semantic Web Conference (pp. 1-6). Karlsruhe, Germany: Citeseer.*

Feldman, R., & Dagan, I. (1995). Knowledge discovery in textual databases (KDT). *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (pp. 112-117). Montreal, Quebec, Canada: ACM.*

Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. New York: *Cambridge University Press.*

Heer, J., & Boyd, D. (2005). Vizster: Visualizing online social networks. *Proceedings of the 2005 IEEE Symposium on Information Visualization (pp. 1-5). Minneapolis, Minnesota, USA: IEEE Computer Society.*

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.*

Inmon, W. H. (2002). Building the Data Warehouse (3rd ed.). New York: Wiley.

Kleinberg, J. (2008). *The convergence of social and technological network. Communications of the ACM, 51(11), 66-72.*

Kotsiantis, S., & Kanellopoulos, D. (2006). Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering, 32(1), 71-82.*

Liu, B. (2007). Opinion Mining. In B. Lui, & B. Liu (Ed.), Web data mining: *Exploring hyperlinks, contents, and usage data (pp. 411-447). Springer.*

Moody, D. L., & Kortink, M. A. (2000). From enterprise models to dimensional models: a methodology for data warehouse and data mart design. *Proceedings of the Second International Workshop on Design and Management. 28, pp. 1-12. Stockholm, Sweden: Citeseer.*

och Dag, N., Regnell, B., Gervasi, V., & Brinkkemper, S. (2005). *A linguistic-engineering approach to large-scale requirements management. Software, IEEE, 22(1), 32-39.*

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2(1), 1-135.*

Parmenter, D. (2007). *Key Performance Indicators. Hoboken, New Jersey, USA: John Wiley & Sons*.

Sen, A., & Sinha, A. P. (2005). A comparison of data warehousing methodologies. *Communications of the ACM, 48(3), 79-84.*

Willen, C. (2002). Airborne opportunities. *Intelligent Enterprise, 5(2), 11-12.*

Yang, W. S., & Dia, J. B. (2008). *Discovering cohesive subgroups from social networks for targeted advertising. Expert Systems with Applications, 34(3), 2029-2038.*

Yang, W. S., Dia, J. B., Cheng, H. C., & Lin, H. T. (2006). Mining social networks for targeted advertising. *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (pp. 1-10). Honolulu, Hawaii, United States of America: IEEE.*

APPENDIX

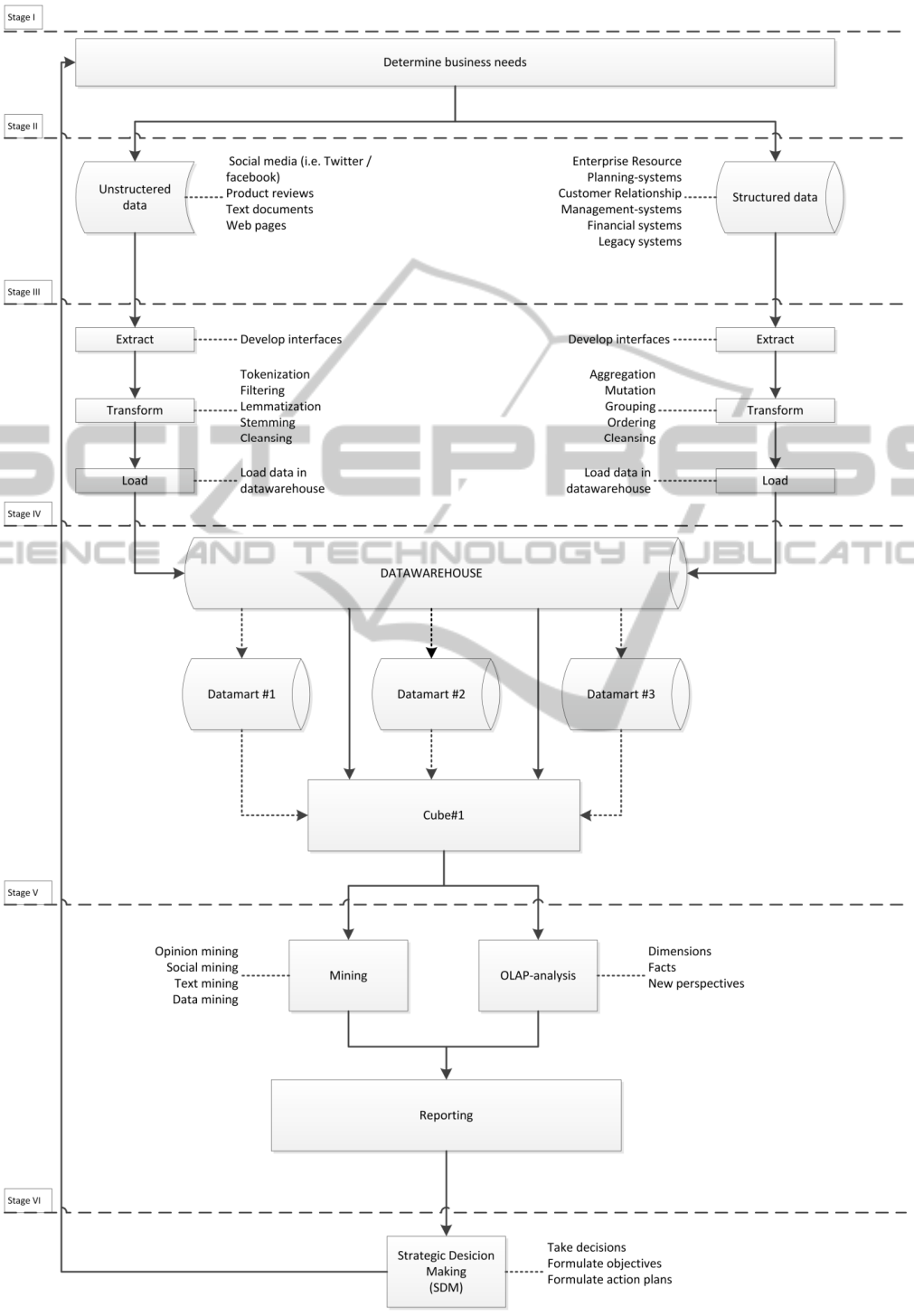

Figure 1: Linguistic Engineering for Business Intelligence framework: comprehensive overview.