LINGUISTIC ENGINEERING AND ITS APPLICABILITY TO BUSINESS INTELLIGENCE

Towards an integrated framework

Otten, S.F.J[1], Spruit, M.R[1]

[1]*Institute of Information an Computer Sciences,Utrecht University, Princetonplein 5, Utrecht, the Netherlands*

otten@csb-system.nl, m.r.spruit@cs.uu.nl

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

In recent years the internet has been evolving from a static source of information to a dynamic and versatile source of information available to everyone. Furthermore, the concept of WEB 2.0 emerged and is in fact a variety of concepts and terms including social networking, social media, and mash-ups (Anderson, 2007). For example, Facebook has over 500 million unique users. Another example is Twitter, a micro blog site, allowing users to share their thoughts in a message with a maximum of 140 characters. Today Twitter has more than 75 million users who express their thoughts and with that, generate valuable information (Kleinberg, 2008). The information is out in the open. Implicit thoughts are made explicit as well as personal preferences and information. However, only a limited amount of organizations leverage the knowledge hidden in these social networks. This is largely due to a lack of knowhow in fields like data mining, text mining and opinion mining (Pang & Lee, 2008).

In the early 1970s Decision Support Systems (DSS) were introduced. A DSS supports the decision making process within an organization and allows for better judgment when deciding on strategic decisions and changes (Watson & Wixom, 2007). DSS evolved over time and in the late 90s Business Intelligence (BI) saw the light of day to satisfy a manager's request for efficiently and effectively analyzing enterprise data in order to better understand the context of their business and improve decision making (Willen, 2002). Negash and Gray (2008) define BI as "*systems which combine data gathering, data storage, and knowledge management with analytical tools to present complex internal and competitive information to planners and decision makers"*. The data that is used in most organizations is already structured and comes from a variety of data sources within an organization such as ERP and CRM systems. The operational data is iteratively aggregated and mutated until it reaches the desired form to support strategic decision making (from a bottom-up approach) (Moody & Kortink, 2000). Unstructured data is in most organizations an untapped field that can generate a lot of potential valuable knowledge to an organization and thereby providing it with a competitive advantage. Before unstructured data can be used to gain a competitive advantage it has to be structured, processed, and analyzed with the help of Linguistic Engineering (LE) techniques (och Dag, Regnell, Gervasi, & Brinkkemper, 2005).

In this paper we aim to provide a framework that provides insight into how linguistic engineering can complement business intelligence 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, data warehousing, and linguistic engineering. Section 3 provides an overview on how this research is conducted. Within section 4 the results of the research are presented, and section 5 completes the paper with a discussion and conclusions.

2 RELATED LITERATURE

In this section we use related literature to position our research. As the terms Business Intelligence (BI) and Linguistic Engineering (LE) are umbrella terms we divided the literature in two parts, BI and LE respectively.

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 and was born in the early 90s and preceded by DSS (Willen, 2002; Chung, Chen, & Numumaker Jr, 2003; Watson & Wixom, 2007). DSS has evolved from two main areas of research – the theoretical studies of organizational decision making conducted during the late 1950s and the technical work carried out at MIT in the 1960s (Keen & Morton, 1978). Classic DSS tool design comprises of components for (1) database management capabilities with access to internal and external data, (2) powerful modeling functions accessed by a model management system, and (3) powerful yet simple user interface designs that enable interactive queries, reporting and graphing functions (Shim, Warkentin, Courtney, Power, Sharda, & Carlsson, 2002). DSS came to evolve over time and in the late 80s and early 90s Executive Information Systems (EIS) where introduced and extended the scope of DSS from small group use to corporate level usage. From there on out new perspectives emerged and one of them called organizational knowledge management, introduced around 1990-1995 is now beginning to mature and is strongly related and intertwined with BI (Paradice & Courtney, 1989). In order to monitor an organization's performance one has to introduce new techniques when it comes to BI. The following techniques should be introduced: data warehouses, OLAP, and data mining.

A data warehouse is a repository where all data relevant to the management of an organization is stored and from which knowledge emerges. The purpose of a data warehouse is to support all levels of management decision-making processes through the acquisition, integration, transformation and interpretation of internal and external data (March & Hevner, 2007). It comprises of data acquired from multiple structured data sources (internal or external) on an operational level within one or more organizations. Via Extract-Transform-Load (ETL) processes and tools, data are first being extracted from the data sources, secondly they are being transformed in order to fit into the structure of the data warehouse, and finally they are loaded into the data warehouse (Sen & Sinha, 2005). However, even with the correct data into the warehouse, it does not support anything yet. It is not yet suited for supporting decision-making processes. Hence, Online Analytical Processing (OLAP) systems were introduced, allowing the aggregation, the drilldown, and slicing/dicing of the earlier acquired data to support decision-making processes and thereby BI (Inmon W. H., 2002).

Another approach to discover useful knowledge in large sets of data in databases is known as data mining. In contrast to the previous example, which uses explicit information, data mining can be used to retrieve implicit, previously unknown information and transform it into useful knowledge (i.e. knowledge rules, constraints, patterns, association rules). One of the most interesting applications of data mining to-date is to find *association rules* in transactional or relational databases (Agrawal, Imieliski, & Swami, 1993; Kotsiantis & Kanellopoulos, 2006). The task is to derive a set of strong association rules in the format "X_1^….^X_m ==> Y_1^…^Y_n" where X_i *(*for $i \in \{1, \dots, m\}$) and Y_j (for $j \in \{1, \dots, n\}$) are sets of attribute-values, from the relevant data sets in a database. With BI, one can utilize technology such as OLAP and data mining to view and analyze internally available data alongside different perspectives and gain new insights allowing for more informed decision making. In order to leverage these new insights and perspectives, an organization needs to have certain objectives, goals and targets in place to compare the actual organizational performance (derived from transactional and operational databases via OLAP and Data mining). In order to do so, an organization can set up Key Performance Indicators (KPIs). Parmenter (2007) defines KPIs as a set of measures focusing on those aspects of organizational performance that are the most critical for the current and future success of the organization. Comparing actual internal data with the defined KPI was made easy due to the fact that data had already been structured and quantified, and metrics within the KPIs have been defined. If a target is not met in a certain period of time, senior management can, and will have to, take action when a critical component in an organization is lacking behind and thereby

posing a threat to the success and continuity of the organization.

Review of scientific literature on this matter reveals that at the moment BI mostly uses data which is internally available, structured and quantified. From there on the data are being aggregated, mutated and analyzed until they are transformed into information and valuable knowledge supporting SDM. Still this knowledge is based on internal data residing in a plethora of electronic data sources (i.e. databases, data warehouses) and organizations are missing out on very important data available in external data sources (i.e. the internet). Section 2.2 provides an overview of techniques and disciplines allowing an organization to leverage unstructured data on the web by converting them into structured data, thereby complementing the existing BI-process in an organization and allowing management to respond more rapidly to changes in the environment.

2.2 Linguistic Engineering (LE)

Linguistic engineering is the discipline concerned with the computational processing of unstructured text and speech in order to derive knowledge from it. Several similar disciplines such as Computational Linguistics (CL) and Natural Language Processing (NLP) exist (Cunningham, 1999). CL research concentrates on "studying natural languages, just as traditional linguistics does, but using computers as a tool to model (and, verify or falsify) fragments of linguistic theories deemed of particular interest" (Boguraev, Garigliano, & Tait, 1995). NLP is a branch of computer science that studies computer systems for processing natural languages. It includes the development of algorithms for parsing, generation, and acquisition of linguistic knowledge (Gazdar, 1996). Taking a closer look at both the definitions of CL and NLP we can conclude that the difference between both of these disciplines lies within its outcome. Whereas the first seems to analyze a natural language in order to verify or falsify its underlying structure, the latter is burdened with the actual extraction of implicit knowledge by using computer systems and turning the implicit knowledge into explicit knowledge. As mentioned earlier in this section, LE is an umbrella term comprising a subset of disciplines and techniques and is concerned with deriving knowledge from unstructured natural language.

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 allowing the unstructured data (i.e. text in natural language form) to be analyzed and derive new knowledge from it and leverage it for a variety of purposes (Hu & Liu, 2004; Yang, Dia, Cheng, & Lin, 2006). Text Mining (TM) or Knowledge Discovery from Text (KDT) was first mentioned by Feldman and Dagan (1995) and deals with the machine supported analysis of text. It uses techniques from Information Retrieval (IR), Information Extraction (IE) as well as NLP and connects them with the algorithms and methods of Knowledge Discovery in Databases (KDD), data mining, machine learning and statistics (Hotho, Nurnberger, & Paass, 2005). Before TM can commence on large document collections a necessity exists to pre-process each text document and store the data in a structured manner, which allows easier and more accurate processing than using an unstructured text document. The pre-processing of text in the TM-domain, to obtain a so-called bag-of-words representation, entails (1) *tokenization*, which splits a text document in a stream of words by removing all punctuation marks replacing tabs and other non-text characters by single white spaces; (2) *filtering*, used to reduce the size of the dictionary and thus the dimensionality of the description of documents within the collection (i.e. stop word filtering, which removes words that have little to no meaning); (3) *lemmatization*, attempts to map verb forms to the infinite tense and nouns to the singular form. However, the word form has to be known in order to achieve lemmatization; (4) *stemming*, which tries to derive the basic form of words, i.e. strip plural 's' from nouns, the 'ing' from verbs, or other affixes (Banko, 1992; Hull, 1996; Feldman & Sanger, 2007). After pre-processing the text documents one can start data mining the outcome by using data mining algorithms found in the KDD-process with the purpose to classify or cluster documents.

Social mining can be defined as *"techniques/processes for developing and applying analytics to social content in order to derive and make sense of knowledge, behavior, affiliations and tendencies of web communities"* (Harris & Valdes, 2008). Another sub discipline of social mining is called Social Network Analysis (SNA) and comprises the analysis and visualization of an online social network or online community (Erétéo, Buffa, Gandon, Grohan, Leitzelman, & Sander, 2008). SNA allows for the examination of an online social network´s structure and thereby identifying sub communities and related key actors like *brokers* and *liaisons* who facilitate knowledge transfer between sub communities (Helms, 2007).

The knowledge to be leveraged, with the help of social mining, in these online social networks has

numerous applications, ranging from constructing customer profiles to defining sub communities and applying targeted marketing and advertising (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 documents or other text-based content, to determine whether comments and reviews are positive, negative or neutral (Liu, 2007). The main tasks as mentioned earlier by Liu (2007) are to (1) find product features that have been commented on by reviewers and (2) decide whether the comments are positive or negative. Determining whether an opinion or review is positive, negative or neutral is called *semantic orientation*. Several methods exists to determine a document's *semantic orientation*, ranging from simple word filtering and counting to context-aware algorithms (Hu & Liu, 2004; Ding, Liu, & Yu, 2008). Opinion mining from the web is a technique which has not yet reached its full potential; in most cases it is limited to deriving meaning from comments on product features of an object (Popescu & Etzioni, 2005; Kanayama & Nasukawa, 2006; Ding & Liu, 2007; Ding, Liu, & Yu, 2008).

Based on the conducted literature review as presented in this section, we now understand that LE is an umbrella term which houses a multitude of disciplines regarding the analysis and processing of natural language (Feldman & Dagan, 1995; Mika, 2005; Liu, 2007). 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. Despite the high degree of applicability of each technique in a BI-environment. In most cases it is not yet implemented in an organization's BI-environment resulting in a loss of valuable information and knowledge about customer profiles, targeted marketing and customer feedback on products (Steinbach, Ertoz, & Kumar, 2003; Yang, Dia, Cheng, & Lin, 2006; Ding, Liu, & Yu, 2008)

2.3 Definition of concepts

In this section we provide an overview of concepts and their definitions. The following concepts will be elaborated on in more detail: decisions support system, business intelligence, data warehouse, structured data, unstructured data, data mining, linguistic engineering, text mining, social mining, and opinion mining.

A *Decision Support System* (DSS), according to Shim et al. (2002), can be defined as "*a computer technology solution that can be used to support complex decision making and problem solving".* Over time the concept evolved into *Business Intelligence* which is "*a system which combines data gathering, data storage, and knowledge management with analytical tools to present complex internal and competitive information to planners and decision makers"* (Negash & Gray, 2008). In order to present complete internal and competitive information to planners and decision makers the data have to be stored. The data are stored in a Data Warehouse, which is defined as "a s*ubject-oriented, integrated, time-invariant, non-updateable collection of data used to support management decision-making processes and business intelligence"* (Hackathorn & Inmon, 1994).

The data stored in a Data Warehouse can be *structured data* which is defined as *"any set of data values conforming to a common schema or type"* (Arasu & Garcia-Molina, 2003) or in can be categorized as *unstructured data* which is defined by Weglarz (2004) as "*any data stored in an unstructured format at an atomic level. That is, in the unstructured content, there is no conceptual definition and no data type definition*".

On data stored in a Data Warehouse one can perform analysis. Such analysis is called *Data Mining* and is defined as "*the process of nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases"* (Piatetsky-Shapiro & Frawley, 1991).

Linguistic Engineering (LE) is a contraction of the terms *linguistic* and *engineering.* To fully understand the term *linguistic engineering* a decomposition of the term is needed and each sub-concept should be examined separately. *Linguistic* is a synonym for the term *language* and can be defined as "a *system for the expression of thoughts, feelings, etc., by the use of spoken sounds or conventional symbols"* (Makins, 1991). A definition of engineering and in particular the sub discipline software engineering can be defined as "*the disciplined application of engineering, scientific, and mathematical principles and methods to economical production of quality software"* (Humphrey, 1988). Having examined both terms separately, we can now define the term *Linguistic Engineering* as "*the process concerned with the computational processing of text and speech by applying scientific and mathematical principles to derive knowledge from unstructured text and spoken language".* Sub disciplines of LE are text mining, social mining and opinion mining. *Text mining* is defined as "*the application of algorithms and methods from the field of machine learning and statistics to texts with the goal of finding useful patterns"* (Hotho, Nurnberger, & Paass, 2005). *Social mining* can be defined as "*techniques/processes for developing and applying*

analytics to social content in order to derive and make sense of knowledge, behavior, affiliations and tendencies of web communities" (Harris & Valdes, 2008). *Opinion mining* can be defined as "*the analysis of a set of text documents D that contain opinions (or sentiments) about an object*" (Liu, 2007). Opinion mining aims to extract attributes and components of the object that have been commented on in each document $d \in D$ to determine whether comments/ reviews are positive, negative or neutral.

3 RESEARCH DESIGN

Our research entails the exploration of the possible uses of Linguistic Engineering (text mining, social mining, opinion mining) in combination with already in place Business Intelligence processes in corporate 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?"

To answer our research question we formulated two additional (sub) research questions allowing us to answer the main research question:

1) What are the main possible applications of text mining, social mining and opinion mining to-date?
2) Can the applications of text mining, social mining and opinion mining complement Business Intelligence?

Departing from these research questions we conducted a literature study comprising the exploration of the main concepts "Business Intelligence" and "Linguistic Engineering" regarding its history and todays use in organizations. Furthermore we explored the sub disciplines, methods and techniques subject to both concepts, in order to get a firm grasp of each concept's capabilities and applicability. We only included scientific publications (i.e. books, articles, and conference proceedings) and did not exclude any specific types of scientific publications.

The literature referenced in this paper was found through an online literature search, mostly using Google Scholar and Omega. Google Scholar is a "general purpose" academic-oriented search engine, which provides the user with selected scientific articles. Omega, our second search engine, covers over more than 16 billion full-text papers in a variety of digital journals of different disciplines. Publishers included are Elsevier, JStor, and Springerlink.

Table 1 provides an overview of keywords used in search-queries for this literature study. 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 performed by checking the title and abstract of each source and quick-scanning the publication for relevant information regarding our research. In total we found 99 relevant scientific publications which could be used in our research.

A more detailed analysis of the found literature, by fully reading the articles, reduced the total from 99 to 70 relevant publications. Due to the conceptual and process-oriented nature of this research we omitted most scientific publications which were mainly mathematically oriented, mostly found in publications with topics like linguistic engineering, text mining and opinion mining, data mining. 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 whereas 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.

Using the 70 relevant publications we began researching the origins of each main concept followed by the identification of popular and validated sub disciplines and / or techniques, concluding with its possible application to-date and determine if it is viable to combine with already in place business intelligence-processes.

4 LEBI FRAMEWORK

Based on the literature study conducted in section 2 of this paper we propose a framework with the aim of bridging the gap between the unstructured data and structured data, allowing an organization to utilize the full potential of information freely available and thereby creating new knowledge.

From literature we derived six main stages (stage I through stage VI) and incorporated them into the proposed framework. Figure 2 depicts a comprehensive overview of the framework showing how each stage is interlinked with one another. These six stages are:

Stage I: Determine business needs required by (senior) management on which they base their decisions.

Stage II: Define the (unstructured and structured) data sources required to effectively retrieve the required data / information.

Stage III: Develop ETL-procedures, allowing the data to be cleansed and mutated as desired.

Stage IV: Setup data warehouse / marts to store the cleansed data and develop data cubes for analysis.

Stage V: Conduct analysis on data residing in data cubes via a variety of techniques (e.g. mining, OLAP analysis) and generate reports with results.

Stage VI: (Senior) management uses the generated reports to decide on a variety of matters and formulate concrete objectives coherent with these decisions and corresponding action plans for the whole organization. Figure 1 is a high-level overview of the LEBI-framework and has a lot of resemblance with already existing frameworks like the KDD framework, CRISP-DM framework and the Three-Phases Framework regarding the defined stages (Vleugel, Spruit, & van Daal, 2010). The difference with these existing frameworks and the LEBI-framework is the addition of the processing and analyzing steps for unstructured data. A more comprehensive view of the LEBI-framework is depicted in figure 2.

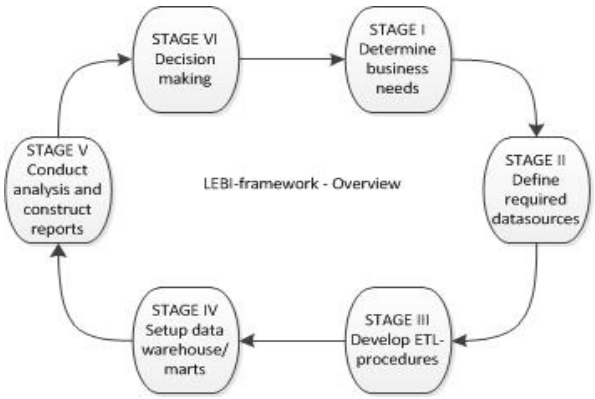

Figure 1: Linguistic Engineering for Business Intelligence framework - overview

As can be seen in figure 2, the LEBI Framework comprises two separate processes in stages II and III, where the unstructured data and structured data are separated until both are properly processed for usage in the data warehouse.

Stage I – Business needs

Determine business needs required by (senior) management on which they base their decisions. 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, one (possibly) has to determine the data sources containing unstructured data (i.e. social networks, text documents) and (possibly) data sources containing structured data (i.e. ERP systems).

Stage III – ETL procedures

Extract: When the data sources with unstructured data are defined, the data has to be extracted. Due to data being unstructured , one has to develop its own interface in order to retrieve data. One can use public available Application Programming Interfaces (APIs) to develop an interface. However, when a data source contains structured data it is often easier to gain access to the data by already available interfaces (i.e. ODBC-connections)

Transform: After data has been extracted, the unstructured data has to be transformed before it can be loaded in the data warehouse. Especially in the case of unstructured data we stress the importance of cleansing the data. For unstructured data (in

particular text) steps like (1) tokenization, (2) filtering, (3) lemmatization, and (4) stemming are important before one can actually load them into the data warehouse/mart. the structured data has to be transformed before it can be loaded in the data warehouse for analysis. Due to the structured nature of the data one can easily aggregate, mutate, group, order, cleanse or filter the data via simple operations (i.e. grouping sales figures per region) in so called ETL-software-packages (i.e. Microsoft Business Intelligence Development Studio). Or one can develop custom queries in their own Database Management System (DBMS).

Load: After the transformation activity is completed on both sides one has to load the newly created data (unstructured and structured) in the designated tables residing in the data warehouse. Depending on the format in which the new data is stored (i.e. XML, CSV, Database) a choice has to be made on how data is loaded into the data warehouse and how databases and tables are designed and stored.

In most cases unstructured data and structured data are still not combined in tables but are stored separately for deeper analyses (more on this in stage IV).

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 in databases and related tables. A decision has to be made regarding the dimensional design/ typology of the data warehouse (snowflake- or star-typology) (Kimball & Ross, 2002), depending on the degree of standardization of the data.

Datamart development: A datamart according to Imnon (1998) is a logical subset (view) or physical subset of a larger data warehouse. Possibly, if required, before developing data cubes, datamarts can be developed which combine data from various databases and or tables residing in the data warehouse for specific analysis purposes.

Cube development: The development of cubes comprises selecting the right dimensions, measures (data to be displayed) and the right combination of data. It is possible to create a cube by (1) selecting data directly from the data warehouse, (2) selecting data from a pre-defined datamart, or (3) use a combination thereof.

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. A variety of algorithms to map a layout and discover possible hidden sub communities.

In case of text documents / product reviews two techniques are used more and more, namely, auto summarization/classification and opinion mining to derive meaning from product reviews. Mining structured data is used to recognizing implicit patterns between explicit data residing in different databases/tables and making it explicit (see section 2.1 for more information)

OLAP-analysis: OLAP-analysis is mostly used for structured data. It allows to view data from different point of views (dimensions) and thereby, possibly, providing new insights. With OLAP-analysis the dimension tables in the data warehouse serve as these point of views and the data being viewed originates from the fact table (see section 2.1.1 for more information) (i.e. view sales per region)

Construct report: After performing the mining-activity and OLAP-analysis on both structured and unstructured data a report should be constructed / generated comprising a combination of results from both activities resulting in new valuable knowledge for (senior) management. (i.e. opinion mining combined with sales figures of a (new) product in a specific time period)

Stage VI – Decision making

The report should be introduced in the SDM-process of (senior) management. Departing from that notion (senior) management can use the report as a supportive tool for making strategic decisions and translating them into objectives and action plans.

Note that it is not required to use the LEBI-framework as a whole. It is possible to use a specific part (unstructured-/ structured-part) of the framework for a variety of purposes. However, when using this framework in a practical environment, we must stress the fact that the order of the stages is sequential and cannot be interchanged with one another at both sides.

5 DISCUSSION

This paper focused on the problem of organizations not being able to utilize freely 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 external unstructured data sources (i.e. social media) into their analyses.

Departing from this observation this research has developed 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 CONCLUSION

This paper was aimed at developing a framework which combined sub disciplines of linguistic engineering with the already existing business intelligence discipline. This implied that the framework should support the extraction, transformation, loading and analyzing of unstructured data as well as structured data. Based on our literature study we found that before conducting any analysis the unstructured data should undergo extreme cleansing (tokenization, filtering, lemmatization, stemming). We also found that the sub 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 the conduction of 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" which aims to derive implicit data from explicit data and turns them into valuable explicit knowledge. 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, for the time being, can already be useful as a structured guide or reference, although leaving ample room for improvement and the implementation of new or existing techniques.

7 REFERENCES

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

Agrawal, R., Imieliski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record, 22*(2), 207-216.

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

Arasu, A., & Garcia-Molina, H. (2003). Extracting structured data from web pages. *Proceedings of the 2003 ACM SIGMOD international conference on Management of data* (pp. 337-348). San Diego, California, United States of America: ACM.

Banko, M. (1992). Lemmatization Algorithms for Dictionary Users. A Case Study. *International Journal of Lexicography, 5*(3), 1-10.

Boguraev, B., Garigliano, R., & Tait, J. (1995). Natural Language Engineering. *Natural Language Engineering, 1*(1), 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.

Cunningham, H. (1999). A definition and short history of Language Engineering. *Natural Language Engineering, 5*(1), 1-16.

Ding, X., & Liu, B. (2007). The utility of linguistic rules in opinion mining. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 811-812). Amsterdam, The Netherlands: ACM.

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.

Gazdar, G. (1996). Paradigm merger in natural language processing. In R. Milner, & I. Wand, *Computing Tomorrow: Future Research Directions in Computer Science* (pp. 88-109). Cambridge: Cambridge University Press.

Hackathorn, R. D., & Inmon, W. H. (1994). *Using the Data Warehouse.* New York: Wiley.

Harris, K., & Valdes, R. (2008). *Hype Cycle for Social Software.* Stamford: Gartner.

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

Helms, R. W. (2007). Redesigning Communities of Practice using Knowledge Network Analysis. In A. S. Kazi, L. Wohlfart, & W. P, *Hands-On Knowledge Co-Creation and Sharing: Practical Methods and Techniques* (pp. 253-273). Knowledgeboard.

Hotho, A., Nurnberger, A., & Paass, G. (2005). A brief survey of text mining. *Journal for Computational Linguistics and Language Technology, 20*(1), 19-62.

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.

Hull, D. (1996). Stemming algorithms. *Journal of the American Society for Information Science, 47*(1), 70-84.

Humphrey, W. S. (1988). The software engineering process: definition and scope. *Proceedings of the 4th international software process workshop on Representing and enacting the software process* (pp. 82-83). ACM.

Inmon, B. (1998). Data mart does not equal data warehouse. *DM Review, 1*(5), 1-10.

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

Kanayama, H., & Nasukawa, T. (2006). Fully automatic lexicon expansion for domain-oriented sentiment analysis. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 355-363). Sydney, Australia: Association for Computational Linguistics.

Keen, P. G., & Morton, M. S. (1978). *Decision support systems: an organizational perspective* (Vol. 35). Reading: Addison-Wesley Publishing.

Kimball, R., & Ross, M. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling* (2nd 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.

Makins, M. (1991). *Collins English Dictionary.* London: Harper Collins.

March, S. T., & Hevner, A. R. (2007). Integrated decision support systems: A data warehousing perspective. *Decision Support Systems, 43*(3), 1031-1043.

Mika, P. (2005). Social networks and the semantic web. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 285-291). Bejing, China: IEEE.

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.

Negash, S., & Gray, P. (2008). Business Intelligence. *Handbook on Decision Support Systems, 2*, 175-193.

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.

Paradice, D. B., & Courtney, J. F. (1989). Organizational knowledge management. *Information Resource Management Journal, 2*(3), 1--14.

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

Piatetsky-Shapiro, G., & Frawley, W. J. (1991). *Knowledge Discovery in Databases.* Cambridge, MA, USA: MIT Press.

Popescu, A. M., & Etzioni, O. (2005). Extracting product features and opinions from reviews. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 339-346). Vancouver, B.C, Canada: Association for Computational Linguistics.

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

Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R., & Carlsson, C. (2002). Past, present, and future of decision support technology. *Decision support systems, 33*(2), 111-126.

Steinbach, M., Ertoz, L., & Kumar, V. (2003). Challenges of clustering high dimensional data. In L. T. Wille, *New Vistas in Statistical Physics - Applications in Econophysics, Bioinformatics, and Pattern Recognition* (pp. 1-273). Springer-Verlag.

Vleugel, A., Spruit, M., & van Daal, A. (2010). Historical Data Analysis through Data Mining From an Outsourcing Perspective: The Three-Phases Model. *International Journal of Business Intelligence Research, 1*(3), 42-65.

Watson, H. J., & Wixom, B. H. (2007). The current state of business intelligence. *Computer, 40*(9), 96-99.

Weglarz, G. (2004). Two Worlds Data-Unstructured and Structured. *DM Review, 14*(1), 19-23.

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

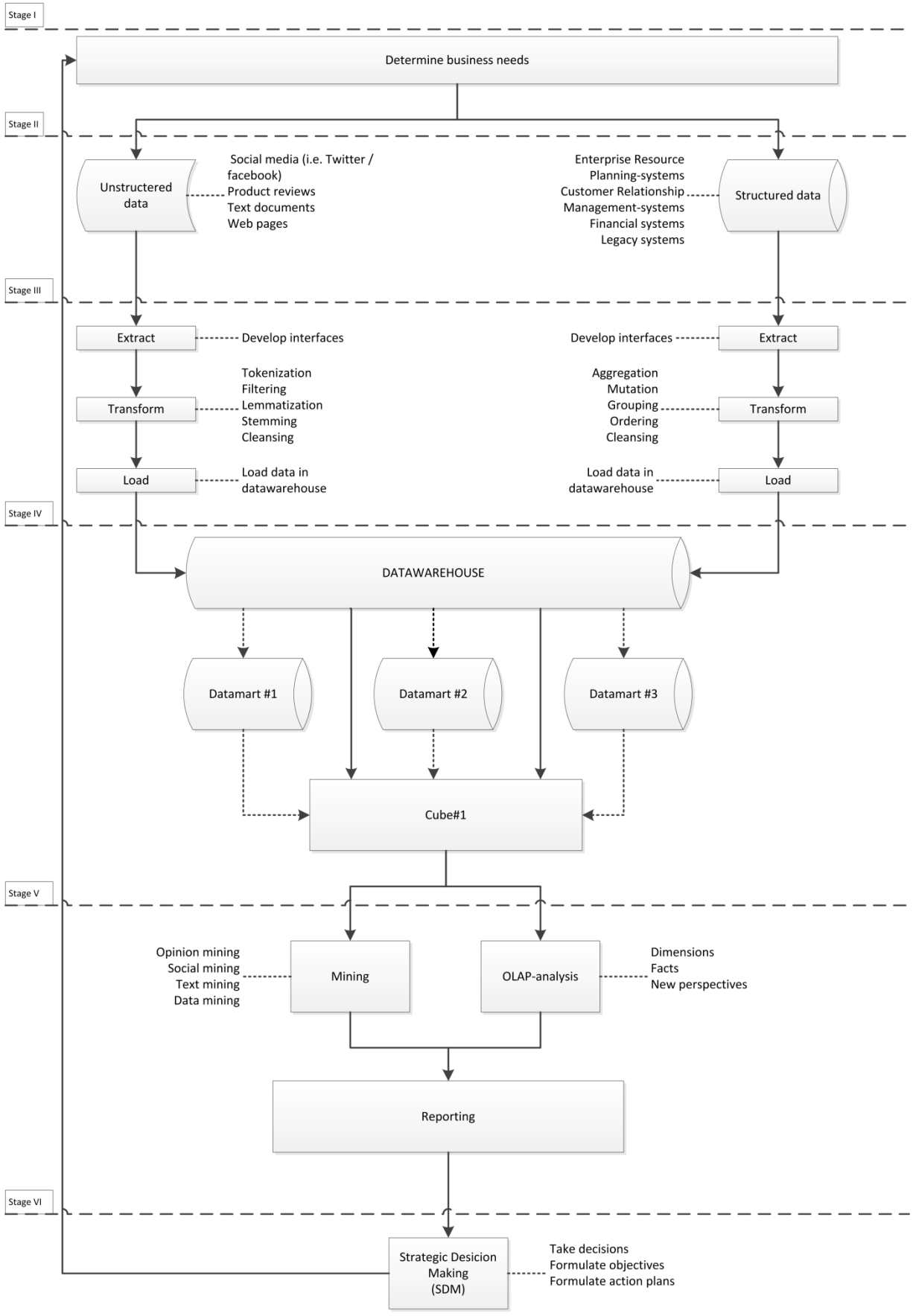

Figure 2: Linguistic Engineering for Business Intelligence framework: comprehensive overview