7th IFAC Conference on Manufacturing Modelling, Management,
and Control
International Federation of Automatic Control
June 19-21, 2013. Saint Petersburg, Russia

Incorporating Text in Enterprise Information Systems

J.C. (“Hans”) Wortmann*. A. Ittoo**

**Faculty of Economics and Business, University of Groningen, 9747 AE Groningen
The Netherlands (Tel: 31 50 363 6598; e-mail: j.c.wortmann@ rug.nl).*
***Faculty of Economics and Business, University of Groningen, 9747 AE Groningen
The Netherlands (Tel: 31 50 363 3853; e-mail: r.a.ittoo@ rug.nl).*

Abstract: Enterprise Information Systems (EIS), the core ICT backbone of organizations, are based on structured data, which are stored in relational databases. These databases may contain text fields as attributes of objects, but lack functionalities to analyze text data. As a result, the considerable amount of valuable texts that is contained in enterprise systems’ database cannot be exploited to enrich corporate activities and processes. At the same time, Natural Language Processing (NLP) techniques have been developed to analyze texts from other sources, such as emails and social media. However, these techniques fail to leverage upon the high quality additional information that is inherent in the structure or schema of the EIS database in order to improve their performance. In this paper, we reconcile the seemingly dichotomous worlds of EIS and NLP. We posit that our approach allows to enrich and incorporate text in enterprise systems.

Keywords: Enterprise Information Systems, ERP, Database Management Systems, Natural Language Processing, Text Miming.

1. INTRODUCTION

Most organizations employ Enterprise Information Systems (EIS) (Finney & Corbett, 2007) (Jacobs & Weston, 2007). These systems provide the transactional backbone for elementary functions such as general ledger (financial bookkeeping), human resources management, and support of operations (Shang & Seddon, 2002). EIS are often realized by standard software such as ERP. Over time, they have been extended with many other technologies, as will be elaborated in this article.

EIS are traditionally based on structured data, organized in relational databases. However, these systems are often inadequate to support the analysis of unstructured texts. This is a rather paradoxical situation given the increasingly important need for text analytics. According to recent studies, an overwhelming 80% of corporate data exists in the form of texts, with the remaining 20% being accounted for by traditional structured data (Blumberg & Atre, 2003) (Russom, 2007).

Most EIS represent text data as attributes (fields) of objects, but provide no other functionalities to discover and exploit the meaningful information nuggets inherent in these texts for supporting business activities, such as getting customer feedback for quality management or collecting knowledge for new product development. In short, textual data is treated within organizations as a “stepchild”, in spite of the prominence of textual sources, such as social media (Grabot, et al., 2012), and the importance of text analytics.

This paper investigates why hardly any use of text data can be observed in EIS. After all, there has been much progress in Natural Language Processing (NLP) techniques over the past years. There are even enterprise search techniques dedicated to (text) mining of short texts as encountered in EIS (Ittoo, 2012). Therefore it is worthwhile to search for reasons why these techniques are not adopted in the context of EIS.

The analysis later in this paper will identify three drawbacks of existing techniques. In the first place, the current mining techniques are not designed to embed their results in a way which can be used in EIS – text remains an alien body. Secondly, these techniques ignore the ontology and knowledge base stored in the EIS. Thirdly, EIS fail to adopt the knowledge extracted from NLP techniques.

Consequently, in this article, we aim to bridge the gap between the state of the art in current EIS technology and the academic and professional developments NLP. Accordingly, this paper is structured as follows. In Section 2, the state of the art in EIS and NLP will be reviewed. Subsequently, we will analyse in Section 3 the occurrence of textual information in these systems. It will be argued that EIS tend to neglect textual information, despite the fact that their databases contain text fields. Also, if NLP techniques are used, they ignore the relation of the texts with the structured data. This analysis leads to the problem statement aimed at in this paper, viz. how can textual information not only be mined and explored, but naturally incorporated in EIS; and vice versa, i.e. how can EIS be used in mining available texts? In section 4 the problem is elaborated by an example, which leads to a direction for solution described in section 5. Section 6 concludes the paper.

978-3-902823-35-9/2013 © IFAC

10.3182/20130619-3-RU-3018.00129

2. LITERATURE REVIEW AND STATE OF THE ART

2.1 Transactional Systems and Relational Databases

The basic of EIS is to support organizations in transaction processing functionality. These systems use structured data as their underlying paradigm. Accordingly, they are based on relational foundations, in such a way that the inherent relational properties of structured data remain unaltered (Jacobs & Weston, 2007). It should be noted that for the discussion in this paper, object-oriented databases are also considered as "structured" databases, because they also have a strict separation between type and instance.

Examples of data objects in EIS include commercial purchasing contracts and orders and employee's contracts (payroll). Objects can also be operational in nature as they depend on the organization's core business. For example, in a factory, they are materials and products. In a hospital, operational objects are patients and treatments. In most cases the resources and the flow of money is also captured by transactional systems (Devaraj, et al., 2013).

2.2 Developments beyond Transaction Processing

Although EIS are rooted in transactional systems, they have developed in many respects beyond this basic functionality (Jacobs & Weston, 2007). There are a number of developments that should be mentioned here (and many developments have to be ignored for conciseness reasons, such as the inclusion of enterprise application integration suites).

A first development to be mentioned is that EIS generally constitute the basis for reports to management and stakeholders on the state of affairs in businesses. A well-known example is the periodic financial statements, published for owners, business partners, tax authorities and other stakeholders. However, there is also an abundant need for other reports in most organizations, such as for management decision making, for compliance with legislation, for public relations. These requirements fit well with the fact that the underlying transactional systems are based on structured data. Structured data allows further enrichments towards data warehouses, which are also in tabular form, and which capture, for example, the change of the EIS' database over time. Much of the external data imported by organizations, such as prices levels, market developments, and economic indicators share the same tabular form, and can be merged with data from transactional systems. The usage of data warehouses and reporting is a second development to be mentioned. A data warehousing system (DWS) integrates data from potentially heterogeneous operational or transaction processing systems. The main technological underpinning of a DWS is the Extract-Load-Transform process, commonly known as ETL. It involves 1) extracting the data from some transaction processing system, 2) transforming the data, and 3) loading the transformed data into a single repository, which in effect is the data warehouse. Typically, the ETL process is performed on a regular basis to populate and/or to update the data warehouse. Thus, a data warehouse can be viewed as a more advanced form of the traditional relational database systems. It is multi-dimensional in the sense that its contents can be analysed across different dimensions (e.g. sale volumes and time), and it is subject-oriented such that all its contents pertain to the same real-world events. Another important characteristic of a data warehouse is that its data are never overwritten or deleted. Once committed the data are static and retained for reporting (Inmon, 2002) (Inmon et al. 2001)(Kimball et al., 2011). A third development is support for professional decision making and analysis, using specialised algorithms that rely extensively on structured data. Data warehouses, in the form of simple spread-sheets or more advanced technologies, are also abundant. Indeed, one of the main advantages of a DWS is its ability to perform on-the-fly, integrated, multi-dimensional analyses on enterprise data. These analyses are then used for further reporting and decision making activities. However, data warehouses exhibit a number of shortcomings. An important limitation of data warehouse technologies is that they are optimally designed to handle factual (structured) data, typically expressed in numerical formats, for e.g. sales volumes (Inmon, 2002). Furthermore, data warehouses rely on well-structured data schemas. Thus, they are unable to handle data expressed in natural language texts. For example, text data can be ascribed to multiple differing interpretations, which cannot always be formulated into clearly defined facts. Handling of these different interpretations is a major challenge to extant DWS. Another important shortcoming of data warehouses is their requirement that the data sources remain static over time in order to ensure that consistency is maintained when integrating data across different sources. However, novel sources of text data, such as Web 2.0 applications like Facebook and Twitter, exhibit a high degree of dynamicity, and thus, cannot be appropriately dealt with by traditional data warehouses (Ribeiro, 2013).To conclude this brief review of EIS, the reader will understand that these systems are not only rooted in structured relational data, but that recent developments tend to strengthen and confirm this feature.

2.3 Natural Language Processing

In Natural Language Processing (NLP), several algorithms have been developed for extracting meaningful information from large text collections. Typical text collections include Wikipedia articles, consumer reviews on forums or even text fields in EIS. These algorithms employ sophisticated linguistic and statistical techniques in order to identify the relevant information nuggets.

Information extraction algorithms can be broadly classified as term extraction (TE) or semantic relation extraction (RE) algorithms. TE algorithms, such as (Ittoo & Bouma, 2013a) (Dagan & Church, 1994), are used to extract domain-specific concepts, e.g. genes or products, which are mentioned in documents. RE techniques are concerned with detecting higher-order information, in particular, semantically related terms, such as part-wholes (Ittoo & Bouma, 2013b) (Girju et al. 2006), and cause-effect (Ittoo & Bouma, 2011) (Girju, 2003).

However, the majority of NLP algorithms developed to date have primarily focused on performance improvements, measured according to the precision and recall metrics. They have largely overlooked the crucial issue of how to leverage upon the valuable information extracted from texts to support and advance classical EIS and related applications, such as business intelligence and decision support. In particular, the output of NLP algorithms, such as terms or semantic relations, are often viewed and interpreted in isolation. Further research is needed to determine whether these outputs are compliant with the ontology underlying the EIS, and whether they can indeed be integrated with the EIS (Cimiano, et al., 2007).

3. ANALYSIS AND PROBLEM STATEMENT

Despite of the fact that EIS are fully rooted in structured data, it often occurs that text fields are added as attributes to objects. For example, doctors may add short paragraphs of text to an object entitled *Consultation*, which is linked to an object *Patient*. Similarly, teachers may enter some textual data in an object *Examination* related to objects *Student* and *Course*. Later in this paper we will elaborate on an example where there is an object *Complaint* which is linked to objects such as *Product*, and which consists of an attribute complaint text. In practice, most EIS offer such attributes with textual values. However, their business logic often do not provide rich functionalities for processing these attributes, and for extracting valuable information from their contents.

The popularity of text attributes in EIS can be attributed to convenience and pragmatism. Users will often create their texts in dedicated environments on PCs or other devices, but, for ease of storage and retrieval, these texts are stored in EIS. This can be done directly or via automated tools, which insert textual data in the corresponding EIS attributes.

Accordingly, the problem statement of this paper is twofold:

- Why is text data handled in EIS as a “stepchild”, despite the progress in text analytic techniques, such as Natural Language Processing (NLP)?
- What should be done to give text data its legitimate role in EIS?

In the remainder of this paper we will focus on text data, which is inserted in EIS via automated tools. This problem is often encountered in practice, and reduces the more general problem indicated in the Introduction to a more manageable size, which can be investigated in the context of a paper.

As already noted, although text data can be inserted into EIS, these systems, including the associated data warehouses, do not provide any means to analyse the data. In particular, NLP techniques are not (yet) a de-facto standard in EIS.

On the other hand, there is a need for information enrichment, just as in the case of structured data. For example, doctors may be interested in clustering their consultation texts in terms of diagnoses given, investigations ordered, treatments and other interventions.

If such text analysis is required, the most common way to proceed with current tools is to download the values of the textual attributes into a set of documents, which can be analysed as a corpus by a suite of NLP techniques. These techniques could be used for purposes as syntactic and semantic analysis, ontology building and document clustering, among others.

However, such an analysis of textual information has two major drawbacks:

- The resulting information cannot be easily fed back into the EIS: the semantics, ontologies, clusters or other results remain autarchic and cannot be (re-) used by the EIS. Therefore, this textual information remains an alien body (Fremdkörper) the EIS
- The knowledge inherent in the EIS (in particular the schema as an ontology but also the database as a knowledge base) is not used to improve the results of the NLP analysis. It is actually a waste of available knowledge if NLP text processing has to always start from each scratch (“ground zero”) each time a particular set of text data is analysed.

Our problem statement, articulated in this section, highlights the increasingly important need to integrate EIS and NLP technologies. To this aim, we propose a framework, depicted in Fig. 1.

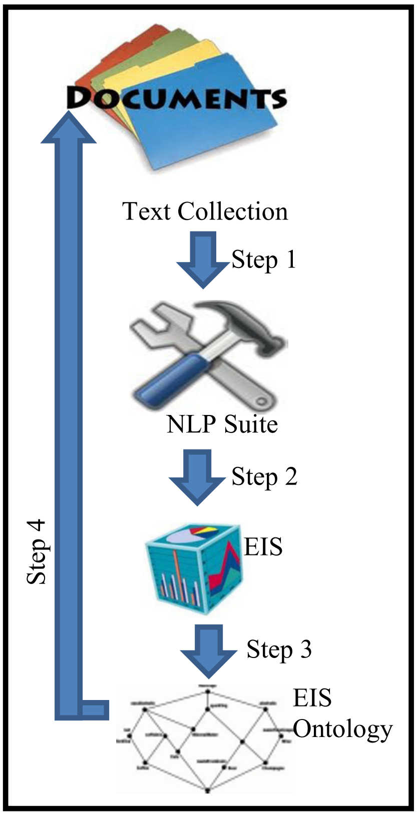

Fig. 1. Overall framework.

Step 1: the text collection is fed to a suite of NLP applications. These applications extract semantically rich information nuggets from the textual contents, such as customers' emotions and opinions, causes of product failures and customer complaints.

Step 2: the extracted nuggets (from Step 1) are used to enrich the structured data contents of an EIS. For example, products in a Sales database can be augmented with customer sentiments about them.

Step 3: the ontology underlying the EIS definition is obtained. For example, database tables correspond to ontological classes, columns to class attributes and records to ontology instances.

Step 4: the derived ontology (Step 3) is used as a knowledge-base to support the NLP tools in extracting more accurate information from new text collections.

The aforementioned 4 steps can then be executed recursively, such that the semantic information from NLP is used to augment structured EIS data (Steps 1,2), and the latter is used to support NLP task (Steps 3,4).

In this article, due to space constraints and to our objectives as articulated in Section 1, we concentrate on Steps 1 and 2, and elaborate on them in Section 4.

4. INTEGRATING EIS AND NLP

Consider the case of a medical equipment company, which enters product complaints on medical equipment in its EIS. Assume that complaints are stored in a class *Complaint*, which has a large text attribute to store the body (text content) of the complaint. Assume also that complaints are linked to a *Particular Product,* which may be a component of another *Particular Product*. Each *Particular Product* is linked to a *Product Type,* as shown in Figure 1.

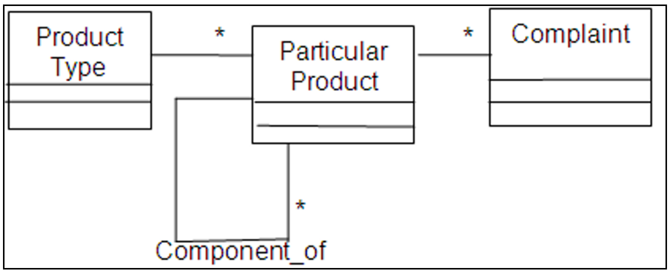

Fig. 2. Partial conceptual schema of EIS.

The texts of complaints are written locally in a PC by medical experts or hospital technicians. These texts have to be automatically fed into the EIS *(problem 1)*, in order to allow answers to later queries such as 'what are all the complaints on computer screens in Spain' *(problem 2)*.

Suppose that a hospital in Barcelona, Spain, has bought a medical system called 'Cardio Vascular 123', and that there is a complaint because the computer screen of the system fails to display images. The screen is denoted as 'Monitor 456' in the EIS. The complaint is in plain text could read like: *"The screen of our Cardio Vascular 123 does not display images"*.

Let us first elaborate problem 1. If the complaint has to be properly fed in the EIS, the textual complaint should be linked to the particular component 'screen'. However, there is no component called 'screen'. But there is a component called 'monitor 456'. Accordingly, problem 1 requires that there is a semantic translation from 'screen' to 'monitor'. Such a translation requires NLP technology: the semantic proximity of the terms 'screen' and 'monitor' in this context should established. However, in order to link the complaint correctly to the right particular product in the EIS database, we need database navigation:

- First, the particular product 'Cardio Vascular 123' has to be found and
- Subsequently the screen-like component 'monitor 456' should be identified.

Accordingly, EIS technology and NLP technology should be combined.

Now consider problem 2. A query such as 'what are all the complaints on computer screens in Spain' is not an SQL query. The problem has to be solved by

- Moving from class level to instance level in the class *Complaints* (EIS technology)

- Finding the complaints linked to screens (NLP technology-semantic proximity)
- Finding the monitors concerned and navigating to their medical systems (EIS technology)
- Finding the locations of these systems, and restricting to Spain (NLP technology based on geographic information retrieval)

The above scenario again shows the need to integrate EIS technology and NLP technologies.

5. DIRECTIONS FOR A SOLUTION

5.1 Improved EIS Technology

From the above discussion and as illustrated in the previous example, it becomes clear that EIS can benefit from adopting NLP technology. Rather than merely supporting queries and reports which are basically rooted in SQL[1], richer querying possibilities should be used which combine the SQL queries with NLP. The example of reporting is only one application where NLP may be used. However, there are many other examples where EIS could benefit from the automated interpretation of text messages, written by humans.

5.2 Improved NLP Technology

In current NLP applications, the focus is on algorithms which improve performance (i.e. precision and recall) of a particular NLP task. The test for improving performance is usually based on a well-defined corpus and starting from a greenfield ("starting from scratch") situation. The current challenge of NLP is largely in automated discovery of semantics and in building ontologies (Turney & Pantel, 2010) (Buitelaar, et al., 2005). Application of this technology can benefit by leveraging upon the information which already resides in the database of EIS. In particular, not only the database schema (such as the UML class diagram depicted in Figure 2) can be used for ontology building, but also the database content (such as the particular products referred to earlier). These constitute reliable information, which is useful for improving the performance of NLP techniques.

6. CONCLUSIONS

In this paper, we described the gap between EIS, text data and NLP techniques. EIS are primarily designed to operate on structured data as found in relational databases. This gives rise to a rather paradoxical situation: the majority of corporate data is currently in the form of unstructured, natural language text, such as in email messages and in social media communications. However, to date, most EIS have largely ignored these textual sources.

This paper identifies the disjoint usage of the above technologies as the root cause of the paradox. It posits that EIS technology can be extended with NLP technology for better (text) data entry and for improved query processing. It also argues that the application of NLP in the context of EIS would benefit from maximal re-use of information which exists in the database of the enterprise system. This information can comprise both the schema (or metadata) and the content of these databases.

[1] Although in EIS practice SQL will be avoided for performance reasons

REFERENCES

Blumberg, R. and Atre, S. (2003). The problem with unstructured data. *DM REVIEW,* Volume 13, pp. 42--49.

Buitelaar, P., Cimiano, P. and Magnini, B. (2005). *Ontology learning from text: methods, evaluation and applications* IOS Press.

Cimiano, P., Hasse, P., Herold, M., Buiterlaar, P. (2007). LexOnto A Model for Ontology Lexicons for Ontology-based NLP. *Proceedings on OntoLex 2007.*

Dagan, I. and Church, K., 1994. Termight Identifying and translating technical terminology.

Devaraj, S., Ow, T. and Kohli, R., (2013). Examining the impact of information technology and patient flow on healthcare performance: A Theory of Swift and Even Flow (TSEF) perspective. *Journal of Operations Management*, 31(4), pp. 181--192.

Finney, S. and Corbett, M. (2007). ERP implementation: a compilation and analysis of critical success factors. *Business Process Management Journal,* 13(2), pp. 329--347.

Girju, R., (2003). Automatic detection of causal relations for question answering. *ACL 2003 workshop on Multilingual summarization and question answering.*

Girju, R., Badulescu, A. and Moldovan, D., (2006). Automatic discovery of part-whole relations. Computational Linguistics, 32(1), pp. 83--135.

Grabot, B., Houé, R., Lauroua, F. and Mayère, A. (2012). *Introducing 2.0 functionalities in an ERP.* RAPMS 2012, Rhodes.

Inmon, W. H., (2002). Building the data warehouse. Wiley.

Inmon, W., Imhoff, C. and Sousa, R., (2001). *Corporate information factory*. Wiley.

Ittoo, A. (2012). *Natural Language Processing meets Business,*Chapter 1. SOM, Groningen.

Ittoo, A. and Bouma, G., (2011). Extracting explicit and implicit causal relations from sparse, domain-specific texts. In: Lecture Notes in Computer Science - Natural Language Processing and Information Systems. Springer, pp. 52--63.

Ittoo, A. and Bouma, G., (2013). Minimally-supervised extraction of domain-specific part–whole relations using Wikipedia as knowledge-base. *Data* (Cimiano, et al., 2007) *Knowledge Engineering*, Volume 85, pp. 57--79.

Ittoo, A. and Bouma, G., (2013). Term Extraction from Sparse, Ungrammatical Domain-Specific. *Expert Systems with Applications*, 40(7), pp. 2530--2540.

Jacobs, F. R and Weston, F., 2007. Enterprise resource planning (ERP)—A brief history. *Journal of Operations Management*, 25(1), pp. 357--363.

Kimball, R., Ross, M., Thornthwaite, W., Becker, B., and Mundy, J., (2011). *The data warehouse lifecycle toolkit*. Wiley.

Ribeiro, J., (2013). *Multidimensional Process Discovery*, Beta - Research School for Operations Management and Logistics.

Russom, P. (2007). BI Search and Text Analytics. TDWI Best Practices Report, pp. 9--11.

Shang, S. and Seddon, P. (2002). Assessing and managing the benefits of enterprise systems: the business manager's perspective. *Information Systems Journal*, 12(4), pp. 271--29.

Turney, P. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1), pp. 141--188.