Text Knowledge Mining: An Alternative to Text Data Mining

D. Sánchez, M.J. Martín-Bautista, I. Blanco
Dept. Computer Science and A.I., University of Granada
E.T.S.I.I.T., C/Periodista Daniel Saucedo Aranda s/n
18071 Granada, Spain
{daniel,mbautis,iblanco}@decsai.ugr.es

C. Justicia de la Torre
Center of Informatics and Communication Networks
University of Granada, 18071 Granada, Spain
consuelo@ugr.es

Abstract

In this paper we introduced an alternative view of text mining and we review several alternative views proposed by different authors. We propose a classification of text mining techniques into two main groups: techniques based on inductive inference*, that we call* text data mining *(TDM, comprising most of the existing proposals in the literature), and techniques based on* deductive *or* abductive inference*, that we call* text knowledge mining *(TKM). To our knowledge, the TKM view of text mining is new though, as we shall show, several existing techniques could be considered in this group. We discuss about the possibilities and challenges of TKM techniques. We also discuss about the application of existing theories in possible future research in this field.*

1 INTRODUCTION

Text mining refers to the discovery of non-trivial, previously unknown, and potentially useful knowledge from a collection of texts. Since its origin, text mining has been considered an analog of data mining (interpreted as Knowledge Discovery in Databases, or KDD) applied to text repositories. Text mining is very important since nowadays, around 80% of the information stored in computers (not considering audio, video, and images) consists of text.

What seems to give text mining a clear distinction from data mining is that the latter deals with structured data, whereas text presents special characteristics and its explicit appearance is basically unstructured. Following this vision of text mining as data mining on unstructured data, most of the approaches to text mining have been mainly concerned with obtaining structured datasets from text, called *intermediate forms*, on which usual data mining techniques are applied. Table 1 shows the most usual intermediate forms that have been proposed and the minimal text units (atomic data) they are comprised of. For example, the *bag of words* represents a piece of text by means of a set of weighted terms, like that employed in Information Retrieval. Data mining techniques employed on the intermediate forms of Table 1 include clustering, dependence analysis, and association rules among others.

Text Unit	Intermediate Form
Word	Bag-of-Words
	N-grams
Concept	Concept Hierarchy
	Conceptual Graph
	Semantic Graph
	Conceptual Dependence
Phrase	N-phrases
	Multi-term text phrases
	Trends
Paragraph	Paragraph
	N-phrases
	Multi-term text phrases
	Trends
Document	Document

Table 1. Minimal text units and their related Intermediate forms

However, textual data is not inherently unstructured. On the contrary, text is characterized by a very complex implicit structure that has defeated many representation attempts,

978-0-7695-3503-6/08 $25.00 © 2008 IEEE
DOI 10.1109/ICDMW.2008.57

with a very rich semantics. On this basis, several authors disagree with the most usual vision of text mining as applying data mining techniques, because these are designed to work on very simple data structures with a very limited expressive power. By employing these data structures as intermediate forms, we are losing most of the semantics of text; in addition, when using data mining techniques on richer semantic structures, we are exploiting a very small part of their expressive power.

Following this discussion, Hearst [7] distinguishes two types of text mining; first, the aforementioned vision of text mining as data mining techniques providing simple patterns and regularities. Second, what she calls *real text mining*, that obtains new and useful knowledge pieces with semantics as richer as that of text. As an example of the latter, she cites the work of Swanson and Smalheiser [20] in the Arrowsmith project. They provide a software tool able to employ the "causal" relationships that appear in the medical literature, as appearing in titles of papers in the MEDLINE database, in order to generate new, previously unknown hypothesis. The following example is a quote from Hearst [7]:

> "For example, when investigating causes of migraine headaches, he [Swanson] extracted various pieces of evidence from titles of articles in the biomedical literature. Some of these clues can be paraphrased as follows:
>
> - stress is associated with migraines
> - stress can lead to loss of magnesium
> - calcium channel blockers prevent some migraines
> - magnesium is a natural calcium channel blocker
> - spreading cortical depression (SCD) is implicated in some migraines
> - high levels of magnesium inhibit SCD
> - migraine patients have high platelet aggregability
> - magnesium can suppress platelet aggregability
>
> These clues suggest that magnesium deficiency may play a role in some kinds of migraine headache; a hypothesis which did not exist in the literature at the time Swanson found these links."

The discussion about what is and what is not text mining has been also considered by other authors, together with reflections about the importance of knowledge and its representation in text mining. In this paper we offer our own view of text mining. In our view, text mining techniques can be classified into two main groups: techniques based on *inductive inference*, that we call *text data mining* (TDM, comprising most of the existing proposals in the literature), and techniques based on *deductive* or *abductive inference*, that we call *text knowledge mining* (TKM). Usual data mining techniques, based on induction, fall in the first group. The technique for generating new hypothesis proposed by Swanson and Smalheiser can be considered, as we shall see, as based on abduction; in this sense, is an example of technique of the second group.

To our knowledge, the TKM view of text mining is new though, as we shall show, several existing techniques could be considered in this group. We discuss about the possibilities and challenges of TKM techniques. We also discuss about the application of existing theories in possible future research in this field.

2 TEXT KNOWLEDGE MINING

2.1 MOTIVATION AND DEFINITION

Data mining techniques for KDD are based on inductive inference, i.e., starting from a set of particular cases described by some (structured) data model, they obtain general patterns consistent with the particular examples. These patterns are subject to objective and/or subjective assessment; the former is based on statistics in general, while the latter takes into account the background knowledge in the domain in order to assess the novelty, usefulness, and non-triviality of the patterns, in accordance with the objectives of KDD. Inductive inference is the natural way to obtain new knowledge from databases since they contain particular examples, simple facts, arranged in data structures.

As we have seen, the usual view of text mining consists of obtaining a structured dataset from text and then applying on it the usual inductive data mining techniques. However, in our opinion, this strong association between *mining* and *inductive learning* that seems to dominate the research in text mining, comes from the fact that data mining was the first "mining" paradigm. Remarkably, the word *induction* does not appear in the definition of text mining. Text mining is concerned with obtaining new, non-trivial, and potentially useful knowledge from text repositories stored in computers. Though inductive approaches have been shown to be very useful, they do not seem to be the "natural" choice for text mining as they are for data mining.

In our view, the main difference between data and text is that the former is represented by means of data models whilst the latter is represented by natural language. And these representations have very different characteristics. Data models have a limited expressive power, mainly restricted to the structured representation of simple facts. On the contrary, natural language has a far richer expressive

power, and it is able to represent not only simple facts like those stored in a database, but general knowledge of almost any kind, from relations like rules to complex procedures. Unfortunately, natural language has some well-known disadvantages like lack of a simple, computationally easy to manage structure, vagueness, ambiguity, etc.

Summarizing, and leaving apart the issue of structure and manageability, the main difference between data and text can be stated as follows:

- Data represents a collection of simple facts, particular cases, from where knowledge can be obtained in a natural way by means of inductive inference.
- On the contrary, text expressed in natural language is a representation of general knowledge, a knowledge base [7], whose semantics is usually much richer than simple facts, on which the natural way of obtaining new knowledge is by means of deductive or abductive inference.

Let us point out that

- the idea of using deductive and/or abductive inference does not contradict the definition of text mining, since there is no restriction about what kind of techniques should be employed to obtain the new knowledge, and
- though we consider deductive and abductive inference to be the natural choice for text mining, this does not means that text mining based on inductive inference is *not* text mining. Almost all the text mining approaches existing in the literature, that have been shown to be very useful in practice, are based on induction.

That is, by proposing text mining based on non-inductive inference, we propose a new direction that, as we shall show, leads to techniques able to provide new knowledge from text that cannot be obtained by approaches based on induction. In summary, we propose to classify text mining techniques into two groups:

- **Text Data Mining (TDM) techniques** obtain new knowledge by applying inductive inference on a structured data representation obtained from text. Most existing text mining techniques fall in this category.
- **Text Knowledge Mining (TKM) techniques** obtain new knowledge by applying deductive and/or abductive inference on a structured knowledge representation of the semantic content of text.

Notice that we employ TKM meaning "mining by reasoning using knowledge contained in text", i.e., the process of obtaining new knowledge taking the knowledge contained in the texts as the starting point. We understand TKM as a particular case of what we call *knowledge mining*, the latter defined as "obtaining non-trivial, previously unknown, and potentially useful knowledge from knowledge repositories". Notice that the main difference between our definition of knowledge mining and the usual definition of data mining is that the former intends to discover knowledge from *knowledge* repositories, whilst the latter tries to do the same from *data* repositories. This difference in the starting point, the basic material from which new knowledge is to be obtained, makes clear that the natural way to knowledge mining is deductive or abductive inference, whilst the natural way to data mining is inductive inference.

We think that TKM techniques are examples of what Hearst [7] call "real text mining" because i) the real knowledge contained in text is employed instead of text metadata or intermediate forms with a poorer semantic content, and ii) new, non trivial knowledge is derived from data, with the same expressive power.

The phases of the TKM process are exactly the same as those of TDM, i.e., first text refining for obtaining a computationally manageable intermediate form representing the text, a mining procedure to obtain new knowledge, and a final step for assessing and filtering the knowledge obtained. However, there are substantial differences between TKM and TDM in the way these procedures are performed and the challenges they pose. We shall go into details in section 3.

2.2 KM: WHAT WE DO NOT MEAN

The term "knowledge mining" has been employed by some authors in a different sense from that introduced in the previous section. In [16], knowledge mining is employed to specify that the intermediate form employed contains knowledge richer than simple data structures (though inductive data mining techniques are used on the intermediate forms). In [9], knowledge mining is employed to indicate that background knowledge and user knowledge is incorporated in the mining process in order to ensure that the intermediate form and the final patterns contain only those concepts interesting for the user; again, inductive data mining techniques are used. In [18] knowledge mining refers to using background knowledge to evaluate the novel and interesting patterns after an inductive process.

Another term appearing in the literature with a different meaning with respect to our proposal is "deductive mining". We have not found a proper definition, but references to [10]. In this paper, deductive mining refer to a group of text mining techniques, where the better known example is said to be Information Extraction. Information extraction is frequently referred as "the mapping of natural language texts into predefined, structured representation, or templates, which, when filled, represent an extract of key

information from the original text"; or alternatively "the process of filling the fields and records of a database from unstructured text".

Information extraction appears in relation to text mining in many papers [8, 14, 15], but it is concerned mainly with the task of obtaining structured intermediate representations from text, that is, with translating the knowledge existing in text into another, structured, representation; however, no new knowledge is derived. The term "deductive" is applicable to information extraction since its starting point is some knowledge about the kind of structures to be filled and rules about how to find instances of values in text to fill them.

However, the knowledge employed for the filling task is not the knowledge contained in the text, but *metaknowledge* about text and how to translate it into a predefined structure, and the deductive inference performed is translating from one representation to another, but no new knowledge is discovered. This fact is pointed out by Hearst [7] in relation to other techniques coming from computational linguistics and applied to particular problems like automatic augmentation of Wordnet relations. In [6], metaknowledge is employed for these purposes for (so-called) knowledge mining. In our view, information extraction performs TDM but not TKM; however, it can play a very important role in obtaining intermediate representations for TKM.

The use of deduction in combination with data mining has been proposed by some authors referring to the incorporation of deductive capabilities to mining tools in order to improve the assessment of the knowledge obtained [19, 5]. In [12], the term "deductive data mining" is employed to indicate that the knowledge should be obtained mathematically and taking into account the hidden assumptions about data. Finally, in [17], a "deductive data mining" paradigm is proposed, where the term "deductive" refer to the fact that the discovery process is performed on the result of user queries (seen as the result of a deductive process) that limit the possibility to work on corrupt data. In all these cases, the generation of new knowledge is purely inductive.

2.3 RELATED WORK

Related to TKM, an interesting summary and discussion about [7] is provided in [11]. In this paper, the authors distinguish between *non-novel investigation*, like information retrieval, in which no new knowledge is generated; *semi-novel investigation*, like standard text mining and KDD, in which patterns/trends that already exist in data are discovered, and *novel investigation* or *knowledge creation*, in which new knowledge, telling something about the world outside the data collection itself, is created. The latter includes solutions to problems like, "how can the linguistic features of text be used to create knowledge about the outside world?", "does a new theme reflect reality?", and "which business decisions are implied?".

For achieving these goals of knowledge creation, the authors [11] propose what they call *Intelligent Text Mining*, i.e., the use of i) interaction between users and mining tools, and ii) AI techniques (no particular technique is specified). However, these techniques are proposed for the assessment of knowledge only. No indication about the kind of inference employed for obtaining the new knowledge is provided.

Some advanced text mining techniques that can be considered TKM are related to literature-based discovery [3, 20, 13, 22]. The objective of these techniques is to provide hypothesis in the form of possible relations between concepts appearing in the literature about a specific topic. In this sense, they can be seen as abduction-based TKM techniques, though some authors consider that their work is not about discovery but about assisting human experts in formulating new hypothesis by means of an interactive process [13, 22]. The process is performed usually on the title and abstract of papers in a bibliographic database. The main work has been made in the medical field by using the MEDLINE database [20, 13, 22], plus some additional background knowledge like the UMLS thesaurus in some cases [22]; however, there are results in other areas [3].

The basic idea of literature-based discovery [3, 20, 22] is the following: let $\{A_1, \ldots, A_n, B_1, \ldots, B_m, C\}$ be concepts and suppose we are interested in finding hypothesis involving possible relations between C and some of the A_i. Suppose that no A_i and C appear simultaneously in the collection, but they can appear simultaneously with some of the B_i. Then, if the set of B_j that appear simultaneously with A_i and C (obviously, in different documents) is large enough, a relation between A_i and C can be hypothesized. In [13], several statistical measures are employed. This procedure has provided new hypothesis that have been later confirmed by experimental studies, as that in the introduction of this paper or that in [22]. Let us remark that C and restrictions about what concepts can play the role A or B are to be specified in advance.

A key point in this approach is the identification of concepts to obtain the intermediate forms of documents. The main problem with this point is that a lot of background knowledge about the text is necessary, and varies a lot from one domain to another. In addition, let us remark that i) the relation between concepts is based in most of the cases on their simultaneous appearance in the same title or abstract only, and ii) the generation of hypothesis does not follow a formal procedure. However, it is inspired on one of the ways humans generate hypothesis. As stated in [22], using a better semantical analysis seems to be promising. Remarkably, some formal model for conjectures, consequences and hypothesis are available in the literature [21, 23], potentially providing a formal tool for literature-based discovery and

TKM.

A text mining technique that we also consider as TKM was proposed in [4] for discovering contradictions in a collection of texts. The starting point is a set of text documents $T = \{t_1, \ldots, t_n\}$ of arbitrary length. The objective is to discover possible contradictions that may be deduced from the content of these texts. For this purpose, a first order logic representation in clausal normal form of the content of text is employed as intermediate form, and a deductive inference procedure based on resolution is performed.

The intermediate form is obtained by a semi-automatic procedure in two steps. First, an interlingua representation based on United Nation's standard UNL [1, 4] is obtained from text. This has the advantages that there are UNL encoders available for different languages, hence eliminating the multilinguism problem typical of text mining, and that UNL is a computationally manageable representation. From the UNL representation, the first order clausal form is obtained by means of information extraction techniques. Background knowledge in the form of either text or clauses can be incorporated to the process as an additional document.

The mining process is performed by using a resolution-based deductive inference procedure. This is applied to every subset of texts looking for contradictions. The procedure is similar to the levelwise + candidate generation exploration in frequent itemsets discovery, but replacing the minimum support criteria by non-contradiction in the knowledge base obtained by the union of the intermediate forms of texts. First, individual text are evaluated to discard the possibility that they contain a contradiction. This corresponds to the exploration of the first level of the lattice of $\mathcal{P}(T)$ with respect to inclusion. Texts containing contradictions are not considered in the following level. Candidate generation for the next level is performed in the same way as in frequent itemset mining in the successive levels. The final result is a collection of subsets of texts in the collection that generate a contradiction. In addition, intermediate clauses obtained by the resolution process are also new pieces of knowledge obtained by this TKM technique.

More specifically, let $res(t_i)$ be a procedure performing resolution on the set of clauses obtained from t_i, giving the value $false$ in case it finds a contradiction. Let $t_i t_j$ be the concatenation of texts (the corresponding clauses) t_i and t_j. Let PT be the set of all the possible concatenations of subsets of T of any size. V represents the set of concatenations that do not contain contradictions. Then, the algorithm proposed in [4] for the discovery of contradictions is that of figure 1.

A problem with this approach is that first order logic representation is monotonic and has some well-known drawbacks. However, this basic procedure could be extended to other kind of logics and the corresponding reasoning for contradictions.

1. $V_1 \leftarrow \{t_i \in T \mid res(t_i) \neq false\}$
2. $V \leftarrow V_1$
3. For k=2 to n
   (a) $V_k \leftarrow \{t't'' \mid t' \in V_{k-1},\ t'' \in V_1, t^* t'' \in V \ \forall t^* \subset t'$ and $res(t't'') \neq false\}$
   (b) $V \leftarrow V \cup V_k$
4. $Contradictions \leftarrow PT \backslash V$

Figure 1. Algorithm for discovering contradictions proposed in [4].

Looking for contradictions is very useful. It can be employed to assess the consistency of the text collection, or to assess the validity of a new text to be incorporated, in terms of the knowledge already contained in the collection. In addition, if we consider collections in which contradictions are known to be (like for example repositories of papers containing opinions about topics), the search for contradictions can be employed in order to separate different groups of texts and, consequently, of authors in terms of consistency of their ideas. Finally, reasoning with ontologies can be reduced in the end to checking the consistency (i.e., non contradiction) of a knowledge base hence looking for contradictions can be seen as a basic procedure for performing different types of reasoning.

3 CHALLENGES OF TKM

As data mining owes a lot to machine learning techniques, statistics, and other previously existing areas, so TKM would benefit of the large amount of existing results in areas like knowledge representation, reasoning algorithms for performing deductive and abductive inference, and knowledge-based systems, as well as natural language processing techniques. As in the case of data mining, these techniques must be adapted for TKM; new specific research results are needed as well. In the following we detail some of the challenges that, in our opinion, TKM must address in several aspects.

3.1 OBJECTIVES

Knowledge-based systems and TKM systems have very different objectives that affect the way techniques coming from knowledge representation and reasoning can be adapted to TKM:

- In a knowledge-based system, a knowledge base is built containing the knowledge needed by the system to solve a specific problem. In TKM, the knowledge managed by the systems is collected from a collection of texts, each of one regarded as a particular knowledge base; however, these knowledge bases in texts are not generated for the specific purpose of building an effective knowledge-based system, but generally as historical or normative reports.

- Knowledge-based systems intend to give answer to *every possible question*, or to solve any possible problem posed to the system. On the contrary, the objective of TKM is just to provide new pieces of non-trivial, previously unknown, and potentially useful knowledge derived from the collection of text. A TKM is not required to obtain *every such piece of knowledge*; in particular, such set of knowledge pieces is unknown by definition.

- A knowledge-based systems performs a reasoning process as a result of a query, stated by the user of a consultive system, or by the necessity to perform an action in a certain situation in the case of intelligent agents. In this sense, the inference procedures employed by knowledge-based systems are in general oriented and optimized to give an answer to specific questions. On the contrary, a TKM system is most generally asked to find new knowledge without specifying a specific question. Of course, the TKM process can also be focused by means of specific queries (e.g. "find new knowledge involving X", with X being and object, a kind of action, or anything else).

3.2 KNOWLEDGE REPRESENTATION

Text mining requires to translate text into a computationally manageable intermediate form. We have introduced some of the most employed intermediate forms for TDM in Table 1. Techniques coming form natural language processing and information retrieval have been employed for obtaining intermediate forms from text. This step of the text mining process is crucial and poses several challenges, common to TDM and TKM, like obtaining intermediate forms from multilingual text collections, defining structures for information extraction, identifying semantic links between new concepts, relating different identifiers, etc.

It is important to remark that the intermediate forms and inference procedures employed are obviously linked to each other, e.g., in order to perform logical deduction, a logic-based representation of the knowledge in texts is needed. Usually, abductive and (specially) deductive inference require a more expressive, semantically richer intermediate form than inductive inference, that are more difficult to obtain from texts in general. However, let us remark that many TDM techniques perform inductive inference on intermediate forms that are also suitable for TKM, as they contain more information about the semantics of texts, like semantic graphs and conceptual dependencies in Table 1, that allow certain kind of non-inductive inference.

A key problem with obtaining intermediate forms for TKM is that the current state-of-the-art techniques for translating text into those intermediate forms are mainly semi-automatic, i.e., they require the participation of human experts in order to verify that the translation is correct. On the other hand, several ongoing projects deal with the possibility of expressing all or at least key parts of the knowledge in text by means of knowledge representation models directly. This is the case for example of work on the Semantic Web, where ontologies represented by description logics or languages with the same expressive power are employed in order to represent the semantics of concepts and their relations [2]. TKM in the Semantic Web seems a promising direction since efficient deductive inference techniques are also available.

The objective though of an intermediate form for TKM is not to represent every possible aspect of the semantic content of text, but those related to the kind of inference we are interested in, that are on its turn determined by the kind of knowledge we want to discover. Even the lack of a text analyzer able to translate without loss from text to the IF of our choice has no more negative effect that, at most, preventing some pieces of knowledge from being discovered. However, as we discussed in the previous section, TKM is not expected to be exhaustive in obtaining new knowledge and, in this sense, this is not a major problem. Of course, better analyzers provide more effective TKM systems.

On the other hand, if the analyzer obtain an inexact representation of text in an intermediate form, this can affect the knowledge discovered, leading to false discoveries. Minimizing such misleading discoveries (that, in data mining and TDM, use to be the larger part of the patterns discovered, together with the uninteresting ones) is an important challenge. In this sense, not only the misfunctioning of text analyzers, but also the reliability of text itself and the adequacy of the intermediate forms introduce new problems in TKM. The knowledge in a knowledge-based system is assumed to be reliable and consistent, but the same cannot be assumed in general in a collection of texts, even before the intermediate form of each one is obtained. Coping with the reliability of knowledge in the inference process is a major challenge in TKM. In this point, existing techniques for the assessment of opinions, importance, reliability and confidence of knowledge may be very useful.

3.3 BACKGROUND KNOWLEDGE

The necessity of incorporating background knowledge in text mining applications is widely recognized. However, it is specially crucial in TKM, since usually text does not contain commonsense knowledge and specific domain knowledge that is necessary in order to perform TKM. For example, a text contaning "A is father of B" and "B is father of A" is not recognized as contradictory unless commonsense knowledge is considered. The recopilation and representation by means of intermediat forms of background knowledge is a challenging problem.

Background knowledge can contribute to TKM in some other very interesting aspects. First, it allows us to "create" new knowledge that is not fully contained in the collection of text, but can be derived from a combination of text and background knowledge. This was a requirement of *intelligent text mining* in [11]. Another interesting application of BK is in the assessment of knowledge, see section 3.5, in aspects like novelty and importance.

3.4 REASONING AND COMPLEXITY

Complexity is one of the important aspects of automatic inference in knowledge-based systems. Any knowledge representation paradigm, from which maybe ontologies and their representation by means of description logics is one of the most visible nowadays, must find a tradeoff between expresiveness and complexity of reasoning. Obviously, TKM systems are also affected by these facts. However, the problem is somehow less hard than in the case of knowledge-based systems or ontological reasoning.

First, text mining and data mining are inherently hard problems in terms of computational complexity. Researchers in these fields are used to cope with exponential algorithms, and many strategies for designing efficient mining algorithms are available. Second, as we have pointed out previously, TKM is not intended to be exhaustive in the derivation of new knowledge. Hence, time and/or space limitations can be stated in the algorithms so that the knowledge obtained is just the knowledge that has been possible to obtain with those resources.

3.5 ASSESSING THE RESULTS

In data mining, objective measures intend to assess the statistical significance of results as a first filter to discard uninteresting patterns; then, subjective assessments about the novelty and usefulness of the patterns obtained are employed as a final filter. Usually, this process is performed directly by human experts, though some methodologies for subjective pattern evaluation have been developed. We consider the following aspects regarding the assessments of results in TKM:

- The reliability of the results in TKM comes from several aspects. First, the validity and reliability of the texts; second, the inference procedures employed. Formal deductive procedures are mathematically sound and consistent, for instance; there are also formal models of abductive inference [21, 23]. Uncertainty in the intermediate forms as a result of the information extraction process may be propagated to the results as well.

- Non-triviality of the results is to be evaluated by the expert user. However, notice that deductive and abductive inference are not trivial reasoning procedures in general, so the results are not expected to be trivial in general.

- Novelty of the results may be evaluated in terms of the BK employed: the new knowledge obtained is novel if it was not previously contained neither in the text collection nor in the BK.

- Usefulness is to be evaluated by the experts. However, a possible way to look for useful knowledge is by leaving the expert to specify the topic (object, action, etc.) he is interested in obtaining knowledge about, prior to the mining process.

4 CONCLUSIONS

This paper introduces the idea of text knowledge mining (TKM) as a particular case of *knowledge mining*, the latter defined as "obtaining non-trivial, previously unknown, and potentially useful knowledge from knowledge repositories". To our knowledge, this idea does not exists in the literature, though the term "knowledge mining" has been employed by different authors in different senses.

Knowledge mining differs from data mining in that the former deals with explicit knowledge, whilst the latter deals with data in which the knowledge is implicit. As a consequence, the natural way to obtain new knowledge in knowledge mining is by means of deductive and abductive inference; on the contrary, inductive inference is the natural way to proceed in data mining.

Most of the existing text mining techniques transform text into data and/or perform inductive inference on the intermediate representation obtained in order to make explicit the implicit knowledge in data. This *text data mining* approach has proved to be very useful in many applications. However, though datasets can be obtained from it, text is primarily an explicit representation of knowledge in natural language. As a consequence, in this paper we have proposed text knowledge mining as an alternative and natural way to text mining that obtains new knowledge from the ex-

plicit knowledge expressed in the text, via deductive and/or abductive inference.

This new proposal opens a wide range of possibilities in text mining. The application of non-inductive inference techniques that, remarkably, take into account a richer representation of the semantic content of text, leads to new types of text mining applications that can take advantage of the extensive available literature in natural language processing, knowledge representation, and reasoning. In this paper we have briefly recalled some existing text mining approaches that, in our opinion, can be considered as TKM techniques. The adaptation to different objectives, the translation of natural language into manageable knowledge representation models, the inclusion of background knowledge, complexity issues of reasoning algorithms, and assessment of the inferred knowledge are some of the exciting challenges to be addressed in the future.

References

[1] *Web of the Institute of Advanced Studies, The United Nations University, Spanish Language Center- Universidad Politécnica de Madrid. http://www.unl.fi.upm.es/homepage.htm.*

[2] G. Antoniou and F. Van Harmelen. *A Semantic Web Primer*. The MIT Press, Cambridge, MA, 2004.

[3] K.A. Cory. Discovering hidden analogies in an online humanities database. *Computers and the Humanities*, 31:1–12, 1997.

[4] C. Justicia de la Torre, M. J. Martín-Bautista, D. Sánchez, and I. Blanco. Un enfoque deductivo para la minería de textos (translates to "a deductive approach to text mining"). In *Actas de Estylf 2006*, pages 241–246, 2006.

[5] S. Greco, E. Masciari, and L. Pontieri. Combining inductive and deductive tools for data analysis. *AI Communications*, 14:69–82, 2001.

[6] U. Hahn and K. Schnattinger. Knowledge mining from textual sources. In F.Golshani and K.Makki, editors, *CIKM'97 - Proceedings of the 6th Intl. Conference on Information and Knowledge Management.*, pages 83–90, ACM, NY, USA, 1997.

[7] M.A. Hearst. Untangling text data mining. In *Proc. 37 Annual Meeting of the Association for Computational Linguistics*, pages 20–26, 1999. http://citeseer.ist.psu.edu/367814.html.

[8] H. Karanikas, C. Tjortjis, and B. Theodoulidis. An approach to text mining using information extraction. In *Proc. KMTA 2000*, 2000.

[9] K. Kaufman and R.S. Michalski. From data mining to knowledge mining. In C.R. Rao, J.L. Solka, and E.J. Wegman, editors, *Handbook in Statistics, Vol. 24: Data Mining and Data Visualization*, pages 47–75. Elsevier/North Holland, 2005.

[10] Y. Kodratoff. Rating the interest of rules induced from data and within texts. In *Proc. 12th Int. DEXA Workshop*, pages 265–269. 2001.

[11] J.H. Kroeze, M.C. Matthee, and T.J.D. Bothma. Differentiating data- and text-mining terminology. In J. Eloff, A. Engelbrecht, P. Kotzé, and M. Eloff, editors, *Proceedings of the 2003 Annual Research Conference of the SAICSIT*, pages 93–101. SAICSIT, Republic of South Africa, 2003.

[12] T.Y. Lin. Deductive data mining: Mathematical foundation of database mining. In G. Wang, Q. Liu, Y. Yao, and A. Skowron, editors, *Proc. 9th International Conference, RSFDGrC. LNAI 2639.*, pages 403–404. Springer, 2003.

[13] R.K. Lindsay and M.D. Gordon. Literature-based discovery by lexical statistics. *JASIS*, 50:574–587, 1999.

[14] A. McCallum and D. Jensen. A note on the unification of information extraction and data mining using conditional-probability, relational models. In *IJCAI03 Workshop on Learning Statistical Models from Relational Data*, 2003. http://citeseer.ist.psu.edu/mccallum03note.html.

[15] R.J. Mooney and R. Bunescu. Mining knowledge from text using information extraction. *SIGKDD Explorations*, 7(1):3–10, 2005.

[16] T. Nasukawa and T. Nagano. Text analysis and knowledge mining system. *IBM Systems Journal*, 40(4):967–984, 2001.

[17] L. Russell. Deductive data mining: Uncertainty measures. In A. Borgida, V.K. Chaudhri, and M. Staudt, editors, *Proceedings of the 5th KRDB Workshop*, pages 15.1 – 15.5, 1998.

[18] M.S. Siadaty and W.A. Knaus. Locating previously unknown patterns in data-mining results: A dual data- and knowledge-mining method. *BMC Med Inform Decis Mak.*, 6(13), 2006. Available online doi: 10.1186/1472-6947-6-13.

[19] E. Simoudis, B. Livezey, and R. Kerber. Integrating inductive and deductive reasoning for data mining. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 353 – 373. AAAI Press, Menlo Park, CA, 1996.

[20] D.R. Swanson and N.R. Smalheiser. Implicit text linkages betwen medline records : Using arrowsmith as an aid to scientific discovery. *Library trends*, 48(1):48–59, 1999.

[21] E. Trillas, S. Cubillo, and E. Castiñeira. On conjectures in orthocomplemented lattices. *Artificial Intelligence*, 117:255–275, 2000.

[22] M. Weeber, H. Klein, L.T.W. de Jong-Van Den Berg, and R.Vos. Using concepts in literature-based discovery: Simulating Swansons Raynaud-fish oil and migraine-magnesium discoveries. *JASIST*, 52(7):548–557, 2001.

[23] M. Ying and H. Wang. Lattice-theoretic models of conjectures, hypothesis, and consequences. *Artificial Intelligence*, 139:253–267, 2002.