Available online at www.sciencedirect.com

ScienceDirect

Expert Systems with Applications 34 (2008) 1707–1720

Expert Systems with Applications

www.elsevier.com/locate/eswa

Seeding the survey and analysis of research literature with text mining

Dursun Delen *, Martin D. Crossland

Department of Management Science and Information Systems, William S. Spears School of Business, Oklahoma State University, Tulsa, OK 74106, United States

Abstract

Text mining is a semi-automated process of extracting knowledge from a large amount of unstructured data. Given that the amount of unstructured data being generated and stored is increasing rapidly, the need for automated means to process it is also increasing. In this study, we present, discuss and evaluate the techniques used to perform text mining on collections of textual information. A case study is presented using text mining to identify clusters and trends of related research topics from three major journals in the management information systems field. Based on the findings of this case study, it is proposed that this type of analysis could potentially be valuable for researchers in any field.

Keywords: Text mining; Data mining; Literature survey; Information extraction; Categorization; Clustering; Classification

1. Introduction

Researchers conducting searches and reviews of relevant literature have faced an increasingly complex and voluminous task. In extending the body of relevant knowledge, it has always been important to work hard to gather, organize, analyze, and assimilate existing pieces from the literature, particularly from the researcher's "home" discipline. With the increasing abundance of potentially significant research also reported in related fields, and even in what are traditionally deemed "non-related" fields of study, the researcher's task is evermore daunting, if a thorough job is desired.

In developing new streams of research, the researcher's task may be even more tedious and complex. Trying to ferret out relevant work that others have reported may be difficult at best, and perhaps even near impossible if traditional, largely manual reviews of published literature are required. Even with a legion of dedicated graduate students or other assisting colleagues, trying to cover all potentially relevant published work is problematic.

For the typical Ph.D. student, trying to find an appropriate dissertation topic can be a formidable task. Even if he or she has a good mentor as a dissertation chair, finding an appropriate topic, one that is both "new enough" and yet still considered "adding to an established body of knowledge" by being supported by the existing literature, can be difficult at best.

Many scholarly conferences take place every year. In addition to extending the body of knowledge of the current focus of a conference, organizers often desire to offer additional minitracks and workshops. In many cases, these additional events are intended to introduce the attendees to significant streams of research in related fields of study, and to try to identify the "next big thing" in terms of research interests and focus. Identifying reasonable candidate topics for such minitracks and workshops is often subjective rather than objectively derived from the existing and emerging research.

A number of journals now have a long history of publishing in their respective fields, and their editors work diligently at making the relevance of their contributions known. Often this includes documenting the history and

* Corresponding author. Tel.: +1 918 594 8283; fax: +1 918 594 8281. *E-mail address:* delen@okstate.edu (D. Delen).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.01.035

development of particular streams of research, and of the contributions made by these streams to the general body of knowledge. As an example, one may consider the journal Management Science, which recently published a series of articles that were summaries of major themes of research reported in that journal in its rich history (Banker & Kauffman, 2004). These types of studies are obviously resource-intensive, requiring one or more contributing authors and perhaps multiple assistants to comb through the history of published papers, trying to develop some notion of coherencies, topical clustering, and trends of research streams over during time.

In this paper, we propose a method to greatly assist and enhance the efforts of the researchers in each of these situations by enabling a semi-automated analysis of large volumes of such unstructured information, through the application of text mining. By accessing the extensive number of abstracts that are available online, we detail how we have used text mining to analyze related research, even across multiple subject domains.

Text mining is the automated or partially automated processing of text. It involves imposing structure upon text so that relevant information can be extracted from it (Miller, 2005; Romero & Ventura, 2007). According to Miller (2005), text mining (at least for common business applications) encompasses the following general classes of activities:

- Text categorization, where the initial objective is to organize and understand various volumes of textual data.
- Information retrieval, relating to searching to find the proverbial "needle in a haystack".
- Measurement, which is developing and defining text measures that may be used to convert contextual information into numeric information, so that it may be analyzed in ways similar to those employed for other business data.

In this study, we explore and utilize all three classes. We develop a method to apply text categorization to organize available research abstracts into logical categories, or topic clusters. The aim is to allow the researcher to consider objectively all the available contributions to a body of knowledge, then to apply repeatable, standardized techniques of categorization to that body of knowledge to identify reasonable and perhaps significant groupings of research papers and articles that "hang together" in objectively justifiable patterns and sets.

In this study, the potential use of text mining on the literature of management information systems is presented. Similar results could be obtained for any research-intensive, literature-rich field of study. Potentially, a quality text mining application on relevant technical literature can:

(1) Enhance the retrieval of information from global databases.
(2) Identify the technology infrastructure (authors, journals, organizations) of a technical area.
(3) Discover new technical concepts or new technical relationships from related or disparate technical literatures.
(4) Identify and categorize the main technical themes and sub-themes in a large body of technical literature.
(5) Identify the relationships between technical themes and infrastructure components.
(6) Identify the time varying differences (or commonalities) between technical themes and infrastructure components.
(7) Provide projections of novel research directions and potential impacts.

The remainder of the paper is organized as follows. The following section provides the justification for the study. Section 2 presents a rather comprehensive description of text mining from a process perspective using a structured modeling method called IDEFØ. Section 3 introduces the case study. Section 4 discusses the findings, and Section 5 concludes the paper.

1.1. Reasoning and justification for this study

As mentioned previously, it is not difficult to justify the development of an approach that will apply automation to the task of organizing a growing body of knowledge in a given field of study. Here we have identified the following as important justifications.

1.1.1. Volume of relevant literature for any given topic

There is now a large and growing volume of published journals, conference proceedings, and practitioner literature available for consideration and study in practically any academic subject. Given such a large volume, the resource-limited researcher may tend to consider only a subset of the available literature, if only for utilitarian purposes and economy of effort. In so doing, however, it is entirely possible to completely overlook important work in "lesser" journals or from supposedly "non-related" fields of study.

The authors invite you to consider the following illustration. In preparing this paper, the authors conducted a typical search of electronic references from a well-known and established online reference library, ProQuest. For the search term "text mining", 31 scholarly journal references were identified in such diverse journals as Nature (Declan, 2004), Personnel Psychology (John, 2005), Analytical Chemistry (Ronald & Ronald, 2001), and Decision Support Systems (Thian-Huat, Hsinchun, Wai-ki, & Bin, 2005). Any or all of these 31 references could have important information for someone interested in the topic. The significance of the existing literature volume is even higher for more mature, literature-rich topic areas. For example, using the same approach just described using ProQuest, a search for the terms ["database management" OR "data base management"] yielded 1,943 scholarly journal articles from a wide variety of journals from various fields of study. Attempting to review even just the abstracts of that many journal articles would probably be impractical for most researchers, even when one truly desires a thorough review of the available literature.

What one could use in such a case is some assistance through automation to determine objectively which of the available references are likely related. One might then safely concentrate on a subset of the literature without feeling that some important reference might have been inadvertently overlooked or ignored. Text mining seems to be a viable technique from which to derive this kind of assistance.

1.1.2. Much of the research literature is now available in electronic form

As indicated in the previous section, much of the available literature for many fields, particularly the later editions of various journals, is now available in electronic form. The libraries of most research-oriented institutions probably have access to the wealth of available scholarly work that now exists in electronic form.

Electronic catalogs such as ABI/Inform, JStor, ProQuest, and Web of Science provide the researcher with direct access to at least titles and abstracts of many journal articles, and in many cases, the full text of articles is also available online. Of course, most of these catalogs are only available by paid subscription, and such subscriptions are often quite costly to maintain for the institution. However, as this paper is being written, the online service giant Google currently has released a beta version of a new product called Google Scholar.[1] This currently free resource claims to be working to make much of the available scholarly research available through a relatively simple online search:

"Google Scholar provides a simple way to broadly search for scholarly literature. From one place, you can search across many disciplines and sources: peer-reviewed papers, theses, books, abstracts and articles, from academic publishers, professional societies, preprint repositories, universities and other scholarly organizations. Google Scholar helps you identify the most relevant research across the world of scholarly research."

With such volumes of scholarly literature becoming increasingly and more easily available, it continues to be more important for the researcher to have the means at his disposal to intelligently and effectively sift through the piles of literature to identify and retrieve the gems of knowledge that pertain to his chosen subject. Text mining can be at least one of the tools that provide such means.

1.1.3. Black-box tools are currently available

Some so-called "black box" commercial tools that one might employ to accomplish some of the tasks of text mining are currently available. The reason we use the term "black box" to describe them is that while they are interesting and do seem to provide some valid information about groups of references that are apparently derived through text mining techniques, the actual algorithms and techniques used to accomplish the analysis are largely hidden from the user. In this paper, we have worked to describe how such analyses are typically carried out, and additionally provide the methodology and algorithms whereby researchers can verify for themselves whether they are appropriate for the task at hand.

1.2. Comparison to traditional manual analysis

Most readers of this paper are undoubtedly familiar with usual methods of reviewing literature in a given area of research. Normally, one starts by defining the topic of interest by developing a list of relevant key words that might have been used by other authors and editors of related research to describe individual pieces of research in the literature. One then goes to the library catalog, or more recently and more likely to the online search engines, and begins entering the selected key words as search topics. As lists of potential relevant articles appear on the screen, the researcher attempts to identify the more relevant ones by scanning the titles and abstracts, and perhaps opening and scanning the full text of some articles, if available. This process is repeated until the researcher has "enough" literature to support the current study.

This approach presents at least a few shortcomings:

- The key words the researcher chooses for the search may not precisely correlate to the key words designated by the authors and/or editors of the work being reviewed, even if they are close in meaning, so all potentially relevant references may not be retrieved.
- The researcher may be consciously or unconsciously optimizing the type or number of online libraries being accessed, and thereby potentially missing important existing works.
- For those potentially relevant works finally identified, the researcher may not fully grasp whether some works are actually relevant due to vague or even fanciful titles being used; the researcher may then neglect to even read the abstract.

Text mining can help to overcome these shortcomings by objectively analyzing all of the abstracts of all the literature, for even those stored in multiple online libraries. None are overlooked or marginalized due to concerns of economy of effort.

2. Overview of text mining

The information age that we are living in is characterized by rapid growth in the amount of data collected and made available in electronic media. In order to process this

[1] see http://scholar.google.com

flood of information, a new paradigm, commonly called Knowledge Discovery in Databases (or more commonly referred to as Data Mining) was coined. Data Mining is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) stored in structured databases, where the data are organized in records structured by categorical, ordinal and continuous variables. However, a vast majority of business data are stored in documents that are virtually unstructured. According to a recent study by Merrill Lynch and Gartner, 85–90% of all corporate data are stored in some sort of unstructured form (i.e., as text) (McKnight, 2005). This is where the text mining fits into the picture. Text mining is the process of discovering new, previously unknown, potentially useful information from a variety of unstructured data sources including business documents, customer comments, Web pages, and XML files.

Text mining (also known as text data mining[2] and knowledge discovery in textual databases[3]) can be described as the process of deriving novel information from a collection of texts (also known as a corpus). By novel information, we mean associations, hypotheses or trends that are not explicitly present in the text sources being analyzed. Even though text mining is considered a part of the general field of data mining, it differs from regular data mining. The main difference is that in text mining, the patterns are extracted from natural language text rather than from structured databases of facts (Yang & Lee, 2005). Databases are designed for programs to process automatically; text is written for people to read. We do not have programs that can "read" and "understand" text (at least not in the manner human beings do). Furthermore, despite the phenomenal advances achieved in the field of natural language processing (Manning & Schutze, 2003), we will not have such programs for the foreseeable future. Many researchers think it will require a full simulation of how the mind works before we can write programs that read and understand the way people do (Hearst, 2003). So what does text mining do? On the most basic level, it numericizes the unstructured text document and then, using data mining tools and techniques, extracts patterns from them.

Benefits of text mining are obvious in the areas where a large number of textual data are collected from business transactions. For example, the free-form text of customer interactions allows trending during time in the areas of complaint (and praise), warranty claims and error tracking, all of which is clearly input to product development and service allocation. In a recent study, Zhang and Jiao (2007) showed the utility of text mining in customer personalization in B2C e-commerce applications. Likewise, market outreach programs and focus grouping are processes rich in generating data. By not restricting the feedback to a codified form, the subject can present, in her own words, what she thinks. Another area where the automated processing of unstructured text made a lot of sense is electronic communications and e-mails. Text mining not only can help classify and filter junk e-mails, but also used to automatically respond to e-mails (Weng & Liu, 2004). Text mining also allows for the sifting through of data in legal, healthcare and other industries traditionally rich in documents and contracts (McKnight, 2005).

Other applications of text mining include (McKnight, 2005; Weng & Lin, 2003):

- Information extraction (identifying key phrases and relationships within text by looking for predefined sequences in text via the process called pattern matching).
- Topic tracking (by keeping user profiles and, based on the documents the user views, predicts other documents of interest to the user).
- Summarization (possessing and summarizing the document to its essence to save time on the part of the reader).
- Categorization (identifying the main themes of a document and, in doing so, placing the document into a pre-defined set of topics categories).
- Clustering (grouping documents that are similar to each other without having a pre-defined set of categories).
- Concept linking (connecting related documents by identifying their commonly shared concepts and, by doing so, helping users to find information that they perhaps would not have found using traditional searching methods).
- Question answering (finding the best answer to a given question by knowledge-driven pattern matching).

In this paper, we provide a structured view and a detailed explanation of the process of text mining. We first outline the process using the IDEFØ activity modeling method, and then provide a case study where we apply text mining to a large collection of information systems literature to extract patterns. We conclude with remarks on the current state and the future directions of the promising field of text mining.

2.1. Text mining process

Fig. 1 depicts the context diagram of the text mining process. In IDEFØ, a context diagram presents the scope of the process along with its interfaces with other processes (i.e., its environment). In essence, it draws the boundaries around the specific process to explicitly identify what is to be included in and what is to be excluded from the representation of the process. A short description of the IDEFØ modeling method is given in Appendix A.

[2] M. Hearst. Untangling text data mining. In Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, 1999.

[3] Ronen Feldman and Ido Dagan. Knowledge discovery in textual databases (KDT). In Knowledge Discovery and Data Mining, pages 112–117, 1995.

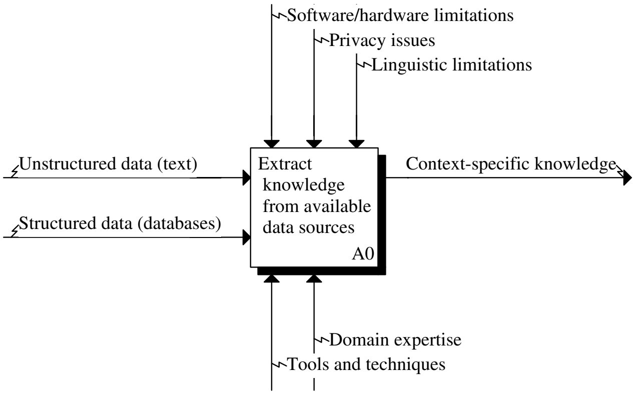
Software/hardware limitations
Privacy issues
Linguistic limitations
Unstructured data (text)
Structured data (databases)
Extract knowledge from available data sources
A0
Context-specific knowledge
Domain expertise
Tools and techniques

Fig. 1. Context diagram for the text mining process.

As the context diagram shows, the inputs to the knowledge discovery process are the unstructured and structured data collected, stored and made available to the process. The output of the process is the context specific knowledge that can be used for decision making. The controls (i.e., constraints) of the process include software and hardware limitations, privacy issues and the difficulties related to the processing of the text that is presented in the form of natural language. The primary purpose of text mining (within the context of knowledge discovery) is to process unstructured (textual) data (along with structured data that reside in databases) to extract meaningful numeric indices from the text. Thus, it makes the information contained in the text accessible to the various data mining (statistical and machine-learning) algorithms.

In Fig. 2, the decomposition of the context diagram is shown. The first activity is to collect all of the documents related to the context (domain of interest) being studied. This collection may include textual documents, XML files, e-mails, Web pages and short notes. Once collected, they are transformed and organized in a manner such that they all are in the same representational form (e.g., ASCII text files). This list of organized documents is then processed so that it can be converted into a term-by-document matrix where the relationships between the terms and documents are characterized by appropriate indices. The details of the document processing are given below. After the processing of documents, combined with the structured data (if any), these indices are used to extract patterns (desired knowledge) for managerial decision making.

The details of document processing are shown in Fig. 3. The main goal here is to convert the list of organized documents (also known as the corpus) into a term-by-document matrix where the cells are filled with the most appropriate indices. The assumption is that the essence of a document can be represented with a list and frequency of the terms used in that document. However, are all terms important in characterization of documents? The answer is obviously "no." Some of the terms (articles, auxiliary verbs, terms used in almost all of the documents in the corpus) have no differentiating power and, therefore, should be excluded from the indexing process. This list of terms, commonly called "stop terms," is specific to the domain of study and should be identified by the domain experts. On the other hand, one might choose a set of pre-determined terms under which the documents are to be indexed (this list of terms is conveniently called "include terms"). Additionally, one would provide the indexing process with synonyms (pairs of terms that are to be treated the same)

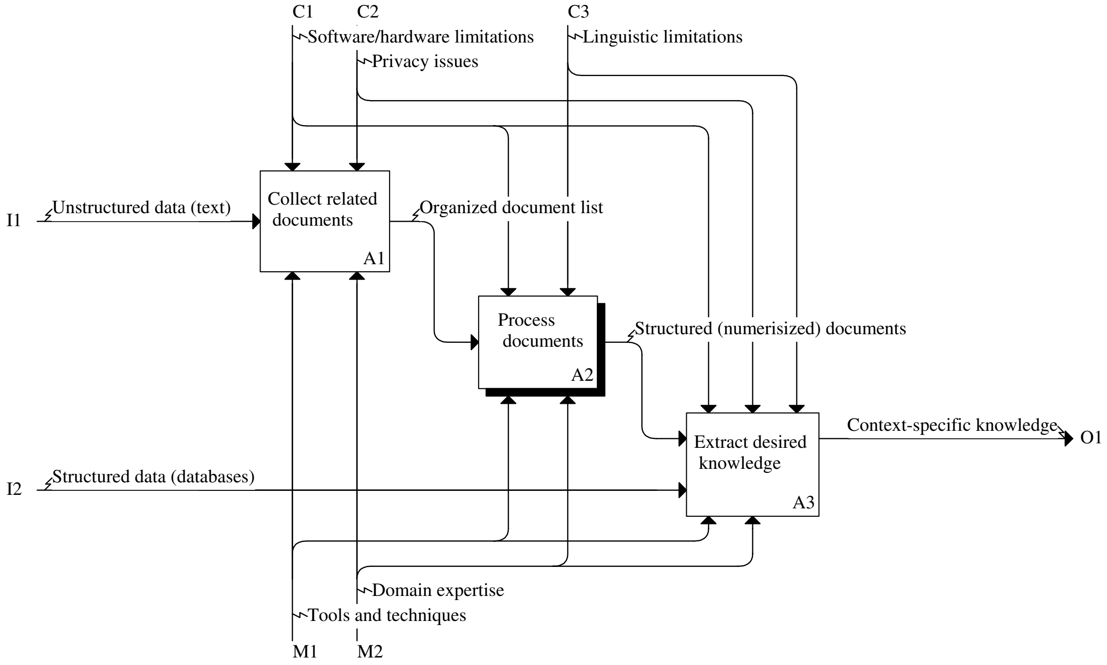
C1
C2
C3
Software/hardware limitations
Privacy issues
Linguistic limitations
I1
Unstructured data (text)
Collect related documents
A1
Organized document list
Process documents
A2
Structured (numerisized) documents
Extract desired knowledge
A3
Context-specific knowledge
O1
I2
Structured data (databases)
Domain expertise
Tools and techniques
M1
M2

Fig. 2. Decomposition of the context diagram.

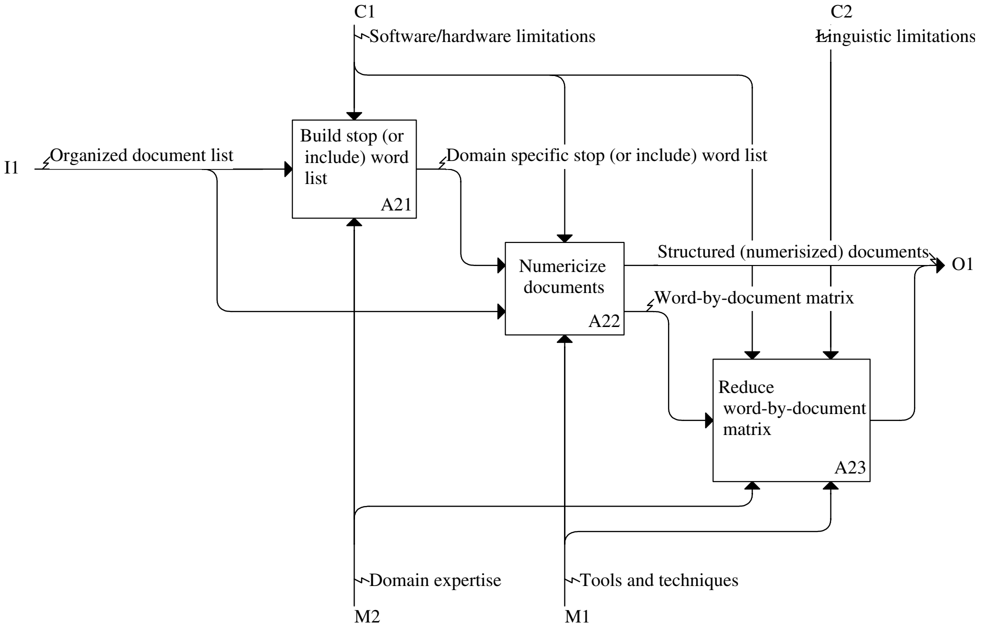
C1
Software/hardware limitations
C2
Linguistic limitations
I1
Organized document list
Build stop (or include) word list
A21
Domain specific stop (or include) word list
Numericize documents
A22
Structured (numerisized) documents
O1
Word-by-document matrix
Reduce word-by-document matrix
A23
Domain expertise
Tools and techniques
M2
M1

Fig. 3. Decomposition of "process documents" activity.

and specific phrases (e.g., "Eiffel Tower") so that the entries in the index would be accurate.

Another filtration that should take place to accurately create the indices is *stemming*, which refers to the reduction of words to their roots so that, for example, different grammatical forms or declinations of verbs are identified and indexed as the same word. For example, stemming will ensure that both "model" and "modeled" will be recognized as the same word.

The first form of the term-by-document matrix includes: all of the unique terms identified in the corpus (as its columns), excluding the ones in the stop term list; all of the documents (as its rows); and the occurrence count of each term for each document (as its cells). If, as is commonly the case, the corpus includes a rather large number of documents, then there is a very good chance that the term-by-document matrix will have a very large number of terms. Processing such a large matrix might be time consuming and, more importantly, might lead to inaccurate patterns.

At this point, one has to decide:

(1) What is the best representation of the indices?
(2) How can we reduce the dimensionality of this matrix to a manageable size?

2.1.1. What is the best representation of the indices?

Once the input documents have been indexed and the initial word frequencies (by document) computed, a number of additional transformations can be performed to summarize and aggregate the information that was extracted.

1. *Log frequencies*. First, various transformations of the frequency counts can be performed. The raw word or term frequencies generally reflect on how salient or important a word is in each document. Specifically, words that occur with greater frequency in a document are better descriptors of the contents of that document. However, it is not reasonable to assume that the word counts themselves are proportional to their importance as descriptors of the documents. For example, if a word occurs one time in document A, but three times in document B, then it is not necessarily reasonable to conclude that this word is three times as important a descriptor of document B as compared to document A. Thus, a common transformation of the raw word frequency counts (wf) is to compute:

$$f(\mathrm{wf}) = 1 + \log(\mathrm{wf}) \quad \text{for wf} > 0.$$

This transformation will "dampen" the raw frequencies and how they will affect the results of subsequent computations.
2. *Binary frequencies*. Likewise, an even simpler transformation can be used that enumerates whether a term is used in a document; i.e.:

$$f(\mathrm{wf}) = 1 \quad \text{for wf} > 0.$$

The resulting documents-by-words matrix will contain only 1s and 0s to indicate the presence or absence of the respective words. Again, this transformation will dampen the effect of the raw frequency counts on subsequent computations and analyses.
3. *Inverse document frequencies*. Another issue that one may want to consider more carefully and reflect in the indices used in further analyses is the relative document frequencies (df) of different words. For example, a term

such as "guess" may occur frequently in all documents, while another term such as "software" may only occur in a few. The reason is that one might make "guesses" in various contexts, regardless of the specific topic, while "software" is a more semantically focused term that is only likely to occur in documents that deal with computer software. A common and very useful transformation that reflects both the specificity of words (document frequencies) as well as the overall frequencies of their occurrences (word frequencies) is the so-called inverse document frequency (for the *i*'th word and *j*'th document):

$$i\mathrm{df}(i,j) = \begin{cases} 0 & \text{if } \mathrm{wf}_{ij} = 0 \\ (1 + \log(\mathrm{wf}_{ij})) \log \frac{N}{\mathrm{df}_i} & \text{if } \mathrm{wf}_{ij} \geqslant 1 \end{cases}$$

In this formula (see also formula 15.5 in Manning & Schutze, 2003), N is the total number of documents, and df_i is the document frequency for the *i*th word (the number of documents that include this word). Hence, it can be seen that this formula includes both the dampening of the simple word frequencies via the log function (described above) and a weighting factor that evaluates to 0 if the word occurs in all documents ($\log(N/N = 1) = 0$), and to the maximum value when a word only occurs in a single document ($\log(N/1) = \log(N)$). It can easily be seen how this transformation will create indices that reflect both the relative frequencies of occurrences of words, as well as their semantic specificities over the documents included in the analysis.

How do we reduce the dimensionality of this matrix to a manageable size? There are several options to pursue (any of which can be used in combination) for managing matrix size:

1. have a domain expert go though the list of terms and eliminate the ones that do not make much sense for the context of the study,
2. eliminate the ones with very few occurrences in very few documents, and
3. transform the matrix into a reduced dimensionality using singular value decomposition.

Singular value decomposition (SVD). The purpose of this technique, which is closely related to principal components analysis, is to reduce the overall dimensionality of the input matrix (number of input documents by number of extracted terms) to a lower dimensional space, where each consecutive dimension represents the largest degree of variability (between words and documents) possible (Manning & Schutze, 2003). Ideally, you might identify the two or three most salient dimensions, accounting for most of the variability (differences) between the words and documents and, hence, identify the latent semantic space that organizes the words and documents in the analysis. In some way, once such dimensions can be identified, you have extracted the underlying "meaning" of what is contained (discussed, described) in the documents. To be more specific, assume matrix A represents an $m \times n$ word occurrence matrix where m is the number of input documents and n is the number of terms selected for analysis. SVD computes the $m \times r$ orthogonal matrix U, $n \times r$ orthogonal matrix V, and $r \times r$ matrix D, so that $A = \mathrm{UDV}'$, and so that r is the number of eigen values of $A'A$.

3. Methodology

3.1. Data gathering and preparation

As a pilot test of using text mining to survey and analyze existing literature, the authors decided to use the articles published in the three major journals in the field of management information systems: MIS Quarterly (MISQ), Information Systems Research (ISR) and the Journal of Management Information Systems (JMIS). Using the standard digital libraries and the online publication search engines, the authors downloaded and collected all of the available articles. In order to maintain the same time interval for all three journals (for potential comparative longitudinal studies), the journal that has the most recent starting date for its digital publication availability is used as the start time for this study (i.e., JMIS articles digitally available since 1994). The publication start times for the remaining two journals were adjusted accordingly; that is, the articles dating prior to 1994 for MISQ and ISR were removed from the dataset. For each article, we extracted the title, abstract, author list, published key words, volume, issue number, and year of publication. We then loaded them into a simple database file. Microsoft Access was used as the container for the resulting library of articles. Also included in the combined dataset was a field that designates the journal type for projected discriminatory studies. To include only the relevant articles, we removed any editorial notes, research notes, and executive overviews from the collection. Table 1 shows the nature of the data collected in a tabular format, while Table 2 shows the basic descriptive statistics of the article collection.

In our analysis, we chose to use only the abstract of an article as the source of information extraction. We have not included the key words listed with the publications for two main reasons: (1) under normal circumstances, the abstract would already include the listed key words, and therefore inclusion of the listed key words for the analysis would mean repeating the same information and potentially giving them unmerited weight, and (2) the listed key words may be terms that authors would LIKE their article to be associated with (as opposed to what is really contained in the article), therefore potentially introducing unquantifiable bias to the analysis of the content.

Table 1
A tabular representation of the fields included in the combined dataset

Journal	Year	Author(s)	Title	Vol/ No.	Pages	Keywords	Abstract
MISQ	2005	A. Malhotra, S. Gosain and O. A. El Sawy	Absorptive capacity configurations in supply chains: Gearing for partnerenabled market knowledge creation	29/1	145–187	Knowledge management supply chain absorptive capacity interorganizational information systems configuration approaches	The need for continual value innovation is driving supply chains to evolve from a pure transactional focus to leveraging interorganizational partner ships for sharing
ISR	1999	D. Robey and M. C. Boudreau	Accounting for the contradictory organizational consequences of information technology: Theoretical directions and methodological implications	2-Oct	167–185	Organizational transformation impacts of technology organization theory research methodology intraorganizational power electronic communication mis implementation culture systems	Although much contemporary thought considers advanced information technologies as either determinants or enablers of radical organizational change, empirical studies have revealed inconsistent findings to support the deterministic logic implicit in such arguments. This paper reviews the contradictory
JMIS	2001	R. Aron and E. K. Clemons	Achieving the optimal balance between investment in quality and investment in selfpromotion for information products	18/2	65–88	Information products internet advertising product positioning signaling signaling games	When producers of goods (or services) are confronted by a situation in which their offerings no longer perfectly match consumer preferences, they must determine the extent to which the advertised features of

Table 2
Descriptive statistics for the article collection

Journal name	Number of articles	Dates of publication	Volumes/ numbers included
MIS Quarterly (MISQ)	246	1994–2005	18/1–29/3
Information Systems Research (ISR)	253	1994–2005	5/1–16/3
Journal of MIS (JMIS)	402	1994–2005	10/4–22/1
Total	901	12 years	

4. Results

The first exploratory study was to look at the longitudinal perspective of the three journals, (i.e., evolutions of research topics over time). In order to conduct a longitudinal study, we divided the 12-year period (from 1994 to 2005) into four 3-year periods for each of the three journals. This framework led to 12 text-mining experiments with 12 mutually exclusive datasets. At this point, for each of the 12 datasets, we used text mining to extract the most descriptive terms from these collections of articles represented by their abstracts. The results are tabulated in Tables 3–5. Table 3 shows the results for ISR for each of the four periods. Each of the four major columns are representative of the four periods, within which the top 15 descriptive terms are listed and ranked based on their frequency. Tables 4 and 5 show the similar results for JMIS and MISQ, respectively.

Results presented in Tables 3–5 may be used to comment on the evolution of the descriptive terms (i.e., research topics) published in the three journals. The interpretation of these results is given in Section 5.

After the longitudinal exploration, we used the complete dataset (including all three journals and all four periods) to conduct a clustering analysis. Clustering is arguably the most commonly used analysis technique in text-mining studies. The idea of using clustering in this study was to identify the natural groupings of the articles (by putting them into separate clusters) and then list the most descriptive terms that characterize those clusters. After the identification of the most descriptive terms as per the process and guidelines explained in the Text Mining Process section, singular value decomposition was used to reduce the dimensionality of the term-by-document matrix, and then an expectation-maximization algorithm is used to create the clusters. Several experimental runs were conducted to identify the “optimal” number of clusters. The results of the clustering experiment are illustrated in Table 6. The first column in Table 6 shows the cluster number, the second column shows the top 10 most descriptive terms for each of the nine clusters, the third and the fourth columns show the frequency of the articles (i.e., the size of

Table 3
Descriptive terms for each of the four periods for ISR

Information systems research (ISR)							
1994–1996		1997–1999		2000–2002		2003–2005	
Key words	Fr.	Key words	Fr.	Key words	Fr.	Key words	Fr.
Model	29	Model	23	Model	38	Model	28
Process	26	Process	21	Process	29	Process	16
User	18	Empirical	18	Empirical	23	Decision	15
Decision	16	Framework	16	Business	23	User	13
Time	14	Electronic	16	Construct	20	Organizational	12
Structure	14	Development	16	User	18	Optimal	12
Performance	13	Structure	15	Factor	17	Environment	12
Organizational	13	Performance	15	Framework	16	Development	12
Development	12	Case	15	Development	13	Change	12
Business	12	Business	15	Benefit	13	Value	11
Perspective	11	Set	13	Set	12	Performance	11
Construct	11	Method	13	Performance	12	Time	10
Change	11	Organizational	12	Organizational	12	Empirical	10
Method	10	Industry	12	Metric	12	Cost	10
Implementation	10	Environment	12	Individual	12	Task	9

Table 4
Descriptive terms for each of the four periods for JMIS

The Journal of MIS (JMIS)							
1994–1996		1997–1999		2000–2002		2003–2005	
Key words	Fr.	Key words	Fr.	Key words	Fr.	Key words	Fr.
Model	34	Process	40	Model	47	Model	52
Process	33	Model	39	Process	45	Process	33
Business	27	Organizational	28	Management	33	Value	24
Development	23	Development	27	Market	26	Business	23
User	22	Decision	24	Organizational	25	Performance	22
Management	21	Support	24	Empirical	23	Factor	22
Factor	21	Management	23	Performance	22	Development	22
Implementation	18	User	22	Practice	21	Quality	20
Performance	17	Business	22	Development	20	Empirical	19
Change	17	Case	20	Time	19	Product	18
Structure	16	Structure	18	Factor	19	Organizational	16
Electronic	16	Framework	18	Business	19	Framework	15
Cost	16	Communication	17	Structure	18	User	14
Benefit	16	Change	17	Cost	18	Project	14
Strategic	15	Variable	16	Perspective	17	Tool	13

Table 5
Descriptive terms for each of the four periods for MISQ

MIS Quarterly (MISQ)							
1994–1996		1997–1999		2000–2002		2003–2005	
Key words	Fr.	Key words	Fr.	Key words	Fr.	Key words	Fr.
Development	23	Model	24	Model	26	Model	34
Process	19	Organizational	19	Organizational	22	Organizational	24
Model	19	Case	18	Management	17	Management	23
Business	19	Process	17	Process	14	Process	22
Change	18	Management	17	Manager	14	Practice	18
Manager	16	Manager	16	Factor	14	User	15
Management	16	Empirical	16	Behavior	14	Behavior	15
Factor	16	User	14	Software	12	Performance	14
Benefit	14	Performance	14	Development	12	Individual	14
Value	13	Decision	14	Construct	12	Action	14
User	13	Change	14	Case	12	Resource	13
Practice	13	Perspective	13	Time	11	Factor	13
Case	13	Managerial	13	Performance	11	Value	12
Organizational	12	Business	13	User	10	Empirical	12
Electronic	12	Time	12	Empirical	10	Business	12

Table 6
Clustering results

Cluster number	Descriptive terms (top 10 key words)	Article frequency	
		(Count)	(Percent)
1	Error, discipline, mis, major, methodology, field, value, time, future, set	54	0.06
2	Network, policy, cost, market, team, industry, resource, manager, product, structure	89	0.10
3	Executive, financial, competitive, industry, advantage, investment, market, management, business, performance	87	0.10
4	Consumer, price, product, customer, online, service, site, quality, market, cost	58	0.06
5	User, database, instrument, model, validation, field, construct, development, computer, individual	121	0.13
6	Strategic, innovation, nature, survey, success, management, investment, practice, business, empirical	56	0.06
7	Medium, computer-mediated, gss, communication, task, face-to-face, laboratory, interaction, idea, social	60	0.07
8	Process, change, organizational, practice, case, management, method, tool, framework, base	181	0.20
9	Risk, project, construct, software, investment, factor, manager, behavior, development, value	195	0.22
	Total	901	100

the cluster) included in those clusters in terms of count and percentage.

After the exploration of the nine clusters for the complete dataset, we analyzed the content of those clusters from (1) representation of the journal types, and (2) representation of the time. The idea was to explore the potential differences and/or commonalities among the three journals and changes during time; that is, to answer the questions such as:

"Are there clusters (which are the representation of different research themes) specific to a single journal?"
"Is there a time-varying (or stable during time) characterization of those clusters?"

Fig. 4 illustrates the representation of the three journals (in terms of number of articles included) for each of the nine clusters, whereas Fig. 5 illustrates the distribution of

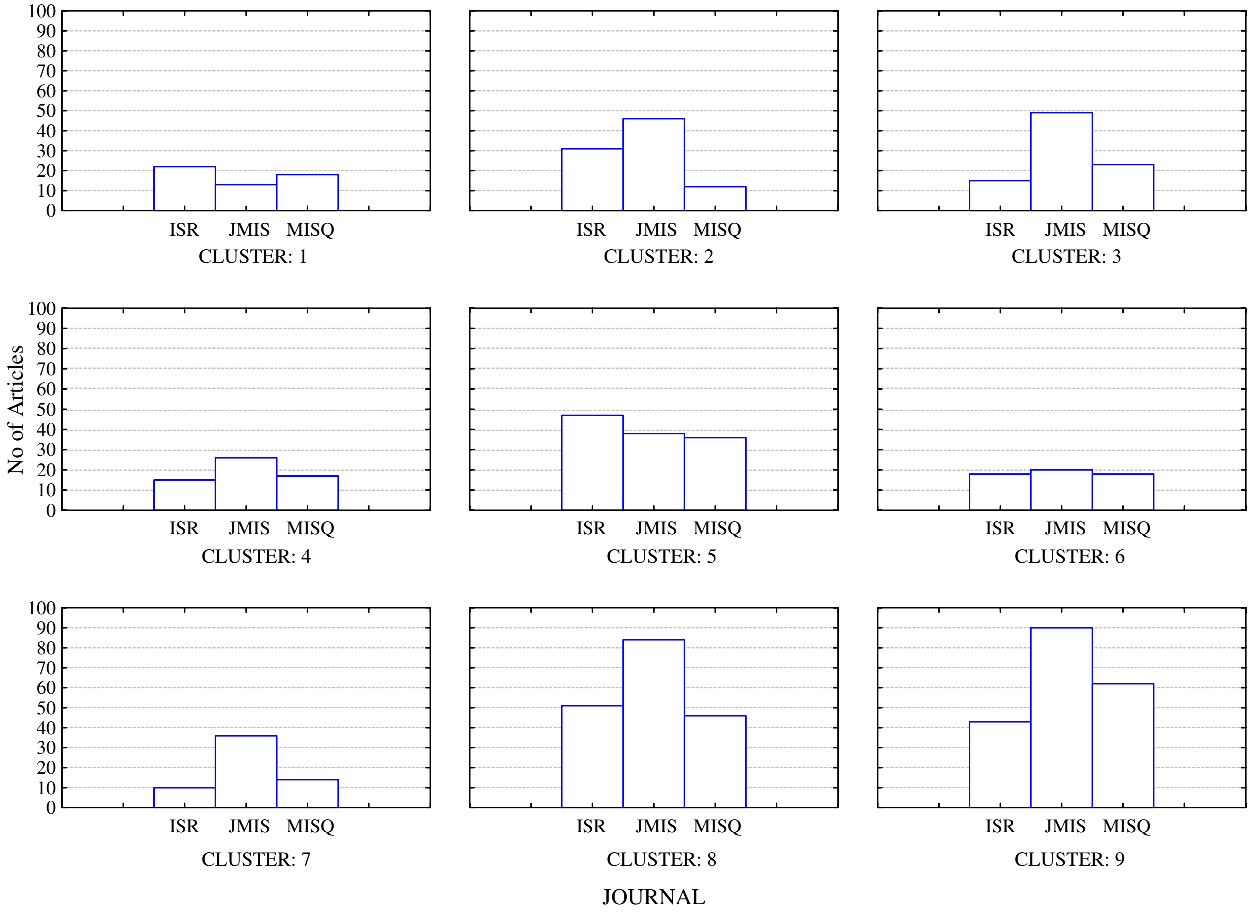

Fig. 4. Representation of the three journals in the nine clusters.

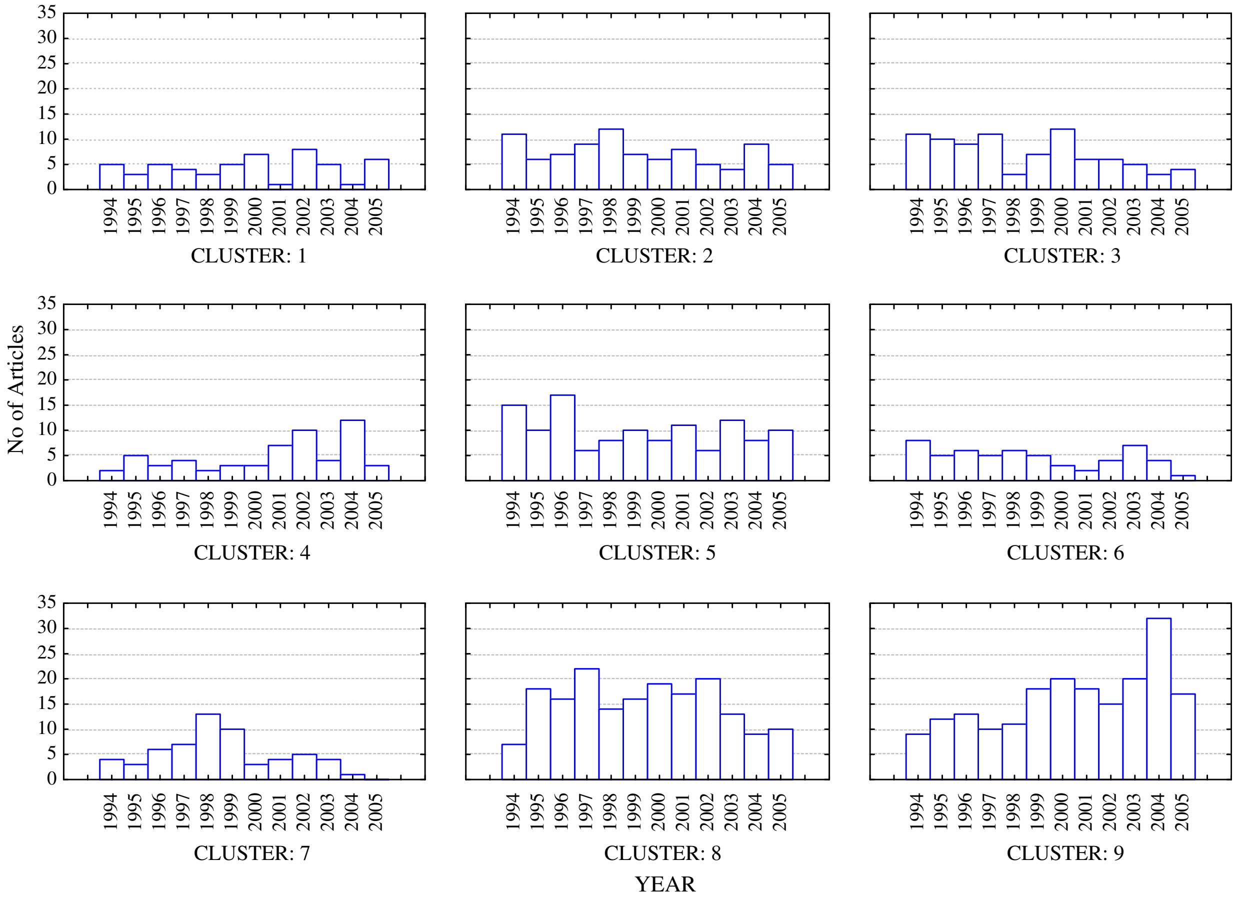

Fig. 5. Distribution of articles over time for each of the nine clusters.

the articles during a 12-year period for each of the nine clusters.

5. Discussion

In this section, we discuss and elaborate on the findings of this exploratory text-mining study, as they are presented in Section 4. Though somewhat comprehensive, it would be immature to claim that this section has a complete coverage of all that could be discussed about the presented results; rather, it intends to cover only the highlights of the study. The discussion starts with commenting on the keyword-driven longitudinal study for each of the three journals, followed by commenting on both the discriminatory and longitudinal characteristics of the clustering exercise.

In Table 3, a list of the most frequently used key words is listed for each of the four time periods for the journal of Information Systems Research. As can be seen, the top two key words, namely model and process, ranked (as one and two, respectively) the same throughout the covered time period of 12 years. An interesting key word that is of interest is "empirical". It looks like the empirical research articles started do be a part of the mix in the second time period, increased in importance in the third period and somehow lost their popularity in the last time period. If you look at the last time period only, key words such as decision, optimal, environmental, change, and value make their way into the top key words list while they were absent from the previous three time periods. The time-varying key word list for the Journal of MIS is presented in Table 4. As can be seen, the first two most commonly used key words for each of the four periods are the same as the ones observed for the journal of Information Systems Research (namely, model and process). In contrast to these two journals, the MIS Quarterly has "organization" as the second most commonly observed key word for the last three periods, replacing the key word "process". An overall look at the three journals, as they are represented with the most frequently observed 15 key words, would reveal that ISR is more likely to publish technical and algorithmic research papers (representative key words include optimal, decision, task, time, etc.) while MISQ is more dominant in behavioral research (representative key words include behavior, action, individual, practice, etc.). JMIS seems to be between the two; closer, however, to MISQ.

Such evolution of key words may indicate what have been and what are currently the popular research topics published in a journal or in a group of journals, kind of being the indicator of research trends during time.

In cluster analysis, we used the complete set of all papers for all three journals as the input data. The initial purpose was to identify the natural groupings of the articles based on the composition of the descriptive terms (most discriminating key words). Since we did not know how many clusters the data should be categorized into and since there is not an optimal way of determining the number of clusters, we followed an experimental procedure where we started with specifying a small number of clusters (say three) and then, for each run, gradually increased the number of clusters. At a point when clusters began to have fewer than the desired number of papers, we stopped the experimentation process and moved forward with this intuitively justifiable number of clusters. In this study, this number was nine. These nine clusters and their descriptive terms, along with the number and percent frequency of the articles, are shown in Table 6. By looking at these key word lists, we can name those clusters with descriptive labels. For instance, the most densely populated cluster (cluster number 9) was characterized as a collection of articles that deal with measurement of value and risk aspects of software projects. Similarly, the second most densely populated cluster (cluster number 8) was characterized as a collection of articles that deal with analysis of change management. The next most populated cluster (cluster number 5) seemed to compile studies that deal with database modeling. Similar labels were derived for the rest of the clusters.

To further explore the nature of these clusters, we went back and created a dataset where, in addition to the articles and the cluster numbers, we added the journal in which they are published and the year of publication. The idea was to, first see whether there is a significant domination of a journal type for any of the nine clusters and, second, whether the publication year is a significant characteristic of any of the nine clusters. Fig. 4 shows the distribution of all three journals (as per the number of articles from each of those three journals represented) for each of the nine clusters. A quick glance at these bar charts may reveal that some of the clusters (cluster numbers 1, 4, 5, and 6) have roughly an equal number of articles from the three journals, while some of them are dominated by one or two of the three journals. Because there is not a clear domination of an individual journal for those nine clusters, we may infer that, based on the kinds of articles that they publish, there is not a significant difference among the journals.

Fig. 5 is an illustration of the nine clusters on time (x-axis) and frequency (y-axis) dimensions using simple bar charts. As can be seen, the number of articles (from all three journals) represented in cluster numbers 3, 5 and 6 gradually decreased over time (maybe indicating that what this cluster is representing is becoming less interesting), while the number of articles represented in clusters 9 and 4 increased over time. Cluster number 7 seems to have spiked in popularity in 1998 and 1999 and then gradually disappeared. This type of time-based view to a specific literature may help identify the future directions of certain topics and topic categories, which may be helpful for a junior faculty or a Ph.D. student in selecting a topic that had a potential to be popular for a while.

This section provides a sample of the kind of general insights one could draw from the results of a text-mining study on a collection of scholarly articles. The depth and nature of the insights depend largely on the main purpose of conducting such a study, as well as the expertise level one has in the specific application domain.

6. Conclusion

The amount of unstructured data are increasing at a higher rate than any traditional (manual) method can keep up with. As the digitized data (either structured or unstructured) becomes more widely available and accessible, tools that allow us to extract information and knowledge from this mountain of data with ease (e.g. data mining and text mining) are likely to become more valuable. Even though text mining is a relatively new technology, its applications and benefits have already been realized in such fields as medicine, homeland security, law, education and customer relationship management.

This study shows that it is relatively straightforward to apply text-mining techniques to a readily available set of data in the form of downloaded journal article abstracts. A total of 1123 articles from three major MIS journals were downloaded and analyzed using text mining, with the objective of identifying major themes of research and how the major themes may have varied during time both within individual journals and within the entire set.

The field of text mining is presently in a growth phase in the research literature. Researchers may apply it to a wide spectrum of unstructured, textual data, from genetic sequence analyses to business e-mail. It makes sense to find ways to use this powerful tool to help analyze the status and direction of the research itself.

6.1. Potential future directions

Most of the current text-analysis algorithms are concerned with counting frequencies of occurrence of individual words, simple variations of those words (e.g. plurals, verb tenses, etc.), and possibly simple phrases. A potentially richer analysis would be to consider natural language processing, including more context-sensitive uses of words and phrases. Such an approach might consider common synonyms and other such contextual similarities to derive notions of what pieces of literature seem to "hang together" in terms of content.

Another future trend for text-mining application is the integration of data mining and text mining into a single framework where both structured and unstructured data of a specific domain may be processed for a more complete knowledge extraction. Even though such an integrated framework sounds like common sense, the current applications are far from realizing it.

Other interesting and somewhat unexplored uses of potential applications of text mining in the field of academic research are as follows:

- Analyzing surveys that include open-ended textual responses. The usual structured analysis of such survey results may miss some of the subtleties or non-standard language used by respondents. This is especially true in the cases where the size of the survey pool is rather large for manual processing. Text mining might be able to identify these previously overlooked pieces of valid and valuable information from large sets of data.
- Automatic processing of technical messages on discussion boards, e-mails, and in instant messaging logs. Text mining can be used to extract the general themes and key words from such a diverse collection of document excerpts. Methods of classification in conjunction with text mining can be used for ranking, "junk mail" filtering, and automatic classification.
- Analyzing secondary data textual reports such as warranty or insurance claims, diagnostic interviews, legal depositions, patents, etc. Much of the information collected for these scenarios tends to be in open-ended, unstructured textual form. Automated text mining of this information might, for example, discover a clustering of repair problems of a certain type or for a particular model or part.
- Automating the environmental scanning by analyzing an industry or competitors by directed "crawling" of Web sites and performing text mining on the resulting set of textual information. Product information, company activities, and other potentially valuable information could be automatically retrieved, compiled, and analyzed using text mining.

There are numerous other potentially fertile grounds for applied text mining research. With this paper, we hope to spark some interest in the topic and to provide an interesting and relevant example of how to apply it.

Appendix A. A brief description of IDEFØ activity modeling method

IDEFØ is a method designed to model the decisions, actions, and activities of an organization or a system. Its foundation dates back to the 1970s, when the US Air Force program for integrated computer integrated manufacturing (ICAM) sought to increase manufacturing productivity through systematic application of computer modeling technologies. The ICAM program identified the need for better analysis and communication techniques for people involved in improving manufacturing productivity. As a result, the ICAM program developed a series of techniques known as IDEF (ICAM Definition) techniques that included IDEFØ function modeling method, IDEF1 information modeling technique (evolved into IDEF1X data modeling method for entity relationship modeling) and IDEF2 dynamic modeling (IDEF, 1994). The most commonly known ones are IDEFØ and IDEF1X. Since its inception in 1983, the IDEF family has evolved into a comprehensive set of standardized methods and methodologies (IDEF, 2005).

IDEFØ is useful in establishing the scope of an analysis, especially for a functional analysis. As a powerful analysis tool, IDEFØ assists the modelers (as well as domain experts) in identifying what functions are performed, how those functions are related to one another, what is needed to perform those functions, what is produced and consumed by those functions, what limits and governs the executions of those functions, and who/what makes those functions happen. Because of its unmatched ability to capture the complete functional perspectives of the systems, IDEFØ models are often created as one of the first tasks of business system analysis efforts.

Fig. 6 shows the basic syntax for an IDEFØ model. The "box and arrow" graphics of an IDEFØ diagram show the function as a box and the interfaces to or from the function as arrows entering or leaving the box (these interfaces are also called "concepts"). To express functions, boxes operate simultaneously with other boxes, with the interface arrows "constraining" when and how operations are triggered and controlled. Specifically, a function represents an activity, a process or a transformation, and is identified by a verb or a verb phrase that describes what must be accomplished. The inputs (connected to the function-box from the left side as incoming flows) represent the data or objects that are transformed (or partially consumed) by the function to produce the output(s). The outputs (connected to the function-box from the right side as outgoing flows) represent the data or objects produced by the function. The controls (connected to the function-box from the top side as incoming flows) represent the conditions and constraints that govern the execution of the function and the production of correct outputs. The mechanisms (connected to the function-box from bottom side as incoming flows) represent the enablers (e.g., who, what, etc) needed to perform the function. The mechanisms can also include a call arrow that enables the sharing (and reuse) of details between models (via linking them) or within a model. The interfaces (i.e., inputs, controls, outputs and mechanisms) are to be described using nouns or noun phrases. Additionally, the description of the activities of a system can be

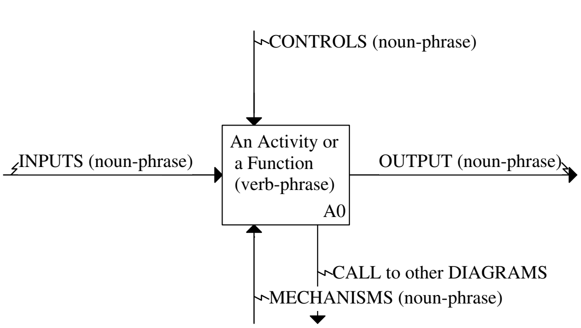

Fig. 6. IDEFØ function box and interface arrows.

easily refined into greater levels of detail until the model is as descriptive as necessary for the specific purpose of the modeling effort using what is called "decompositions". A typical IDEFØ model starts with a context diagram where the whole system is represented with a single function (a single box) and its interfaces (inputs, controls, outputs and mechanisms). This function that represents the system as a single module is then detailed on another diagram with sub-functions connected by interfaces. Each of these sub-functions may then be similarly decomposed to expose more details about the system. A diagram's external interfaces must be limited to the interfaces propagated (i.e., inherited) from the parent function.

As a structured modeling language, IDEFØ has the following characteristics that make it stand out:

1. It is comprehensive and expressive, capable of graphically representing a wide variety of enterprise operations to any level of detail.
2. It is a coherent and simple language, providing for rigorous and precise expression, and hence promoting consistency of usage and interpretation.
3. It enhances communication between modeler creators and model users through ease of learning and interpretation and its emphasis on hierarchical exposition of detail.
4. It is well-tested and proven, through many years of use in the US Air Force and other government agencies, and by private industry.
5. It can be generated by a variety of computer graphics tools; numerous commercial products specifically support development and analysis of IDEFØ diagrams and models.

In addition to the definition of the IDEFØ language, the IDEFØ methodology also describes procedures and techniques for developing and interpreting models, including ones for data/information gathering, diagram construction, review cycle and documentation. For details of the IDEFØ methodology, other than the language specification, the reader is referred to the IDEFØ method report (IDEF, 1994). The next subsection briefly summarizes the language specifications of the IDEFØ method.

References

Banker, R. D., & Kauffman, R. J. (2004). The evolution of research on information systems: a 50-year survey of the literature in management science. *Management Science, 3*(50), 281–298.

Declan, B. (2004). Frenchman is most thanked computer scientist. *Nature, 7019*(432), 790.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: an overview. In *Advances in knowledge discovery and data mining* (pp. 1–34). Cambridge, MA: AAAI/MIT Press.

Hearst, M. (2003). *What is text mining?* UC Berkeley: SIMS.

IDEF (2005). IDEF family of methods. <http://www.idef.com>.

John, M. F. (2005). Data and text mining: a business applications approach. *Personnel Psychology, 1*(58), 267.

Manning, D. M., & Schutze, H. (2003). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

McKnight, W. (2005). Building business intelligence: text data mining in business intelligence. In DM Review (pp. 21–22).

Miller, T. W. (2005). *Data and text mining: a business applications approach*. New Jersey: Pearson/Prentice Hall.

Romero, C., & Ventura, S. (2007). Educational data mining: a survey from 1995 to 2005. *Expert Systems with Applications, 33*(1), 135–146.

Ronald, N. K., & Ronald, A. D. (2001). Extracting information from the literature by text mining. *Analytical Chemistry, 13*(73), A370.

Thian-Huat, O., Hsinchun, C., Wai-ki, S., & Bin, Z. (2005). Newsmap: a knowledge map for online news. *Decision Support Systems, 4*(39), 583.

Yang, H.-C., & Lee, C.-H. (2005). A text mining approach for automatic construction of hypertexts. *Expert Systems with Applications, 29*(4), 723–734.

Weng, S.-S., & Lin, Y.-J. (2003). A study on searching for similar documents based on multiple concepts and distribution of concepts. *Expert Systems with Applications, 25*(3), 355–368.

Weng, S.-S., & Liu, C.-K. (2004). Using text classification and multiple concepts to answer e-mails. *Expert Systems with Applications, 26*(4), 529–543.

Zhang, Y., & Jiao, J. (2007). An associative classification-based recommendation system for personalization in B2C e-commerce applications. *Expert Systems with Applications, 33*(2), 357–367.