Overview and Semantic Issues of Text Mining

Anna Stavrianou
Université Lumière Lyon2, France
anna.stavrianou@univ-lyon2.fr

Periklis Andritsos
University of Trento, Italy
periklis@dit.unitn.it

Nicolas Nicoloyannis
Université Lumière Lyon2, France
nicolas.nicoloyannis@univ-lyon2.fr

ABSTRACT

Text mining refers to the discovery of previously unknown knowledge that can be found in text collections. In recent years, the text mining field has received great attention due to the abundance of textual data. A researcher in this area is requested to cope with issues originating from the natural language particularities. This survey discusses such semantic issues along with the approaches and methodologies proposed in the existing literature. It covers syntactic matters, tokenization concerns and it focuses on the different text representation techniques, categorisation tasks and similarity measures suggested.

1. INTRODUCTION

The field of text mining has received a lot of attention due to the always increasing need for managing the information that resides in the vast amount of available text documents. Text documents, as opposed to information stored in database systems, are characterized by their unstructured nature. Ever increasing sources of such unstructured information include the World Wide Web, governmental electronic repositories, biological databases, news articles, blog repositories, e-mails.

Text mining is the data analysis of text resources so that new, previously unknown knowledge is discovered [34]. It is an interdisciplinary field that borrows techniques from the general field of Data Mining and it, additionally, combines methodologies from various other areas such as Information Extraction (IE), Information Retrieval (IR), Computational Linguistics, Categorization, Topic Tracking and Concept Linkage [23; 53].

It is often ambiguous to distinguish between the field of IR and that of text mining. This happens because they both deal with text and its particularities, so they both have to face similar issues. IR has lent several algorithms and methods to text mining. The difference between these two fields is mainly their final goal. In IR, the objective is to retrieve documents that partially match a query and select from these documents some of the best matching ones [76]. Text mining is about discovering unknown facts and hidden truth that may exist in the lexical, semantic or even statistical relations of text collections.

Another field that has lent methodologies to text mining is Information Extraction (IE). IE differs from text mining because it regards the extraction of specific, structured data (e.g. names of people, cities, book titles) and prespecified relationships [71] rather than the discovery of new relations and general patterns. In Text Mining the information found is unsuspected and unexpected, though in IE it is predefined and it matches the interest specified by the user [48; 53; 71]. IE techniques may be part of the text mining task in order to facilitate the knowledge extraction.

The text mining process consists of a data analysis of a corpus or corpora and it is concisely illustrated in Figure 1. Taking a collection of text resources, a text mining tool would proceed with the data analysis. During this analysis many sub-processes could take place such as parsing, pattern recognition, syntactic and semantic analysis, clustering, tokenization and application of various other algorithms. Following the data analysis, the results are evaluated and the new, previous unknown knowledge may emerge. The retrieved text information can be used in various ways such as database population and reconciliation.

Text Mining associates text documents and database models. This association can be summarized in the following points:

- population of a database schema with data retrieved from web documents
- discovery of information existing in texts and storage to a relational or XML format
- integration and querying of text data after it has been stored in databases
- deduplication of a dataset by using standard data mining techniques, such as clustering.

A great deal of Statistics and Machine Learning techniques exist and contribute to the data analysis, and therefore the text mining task. However, during the text mining process, many issues arise because of the automatic natural language processing (NLP) limitations, which the aforementioned techniques do not always take into consideration. A researcher needs to have a thorough overview of the existing difficulties posed by text before deciding on how to cope with them. In this paper we concentrate on the

semantic issues present in text mining and we refer to some approaches that have attempted to handle these issues.

This paper is organized as follows. Section 2 discusses the reasons that make text mining significant and Section 3 refers to NLP issues. In Section 4 the focus is on the text representation techniques discussed in the existing literature, while Section 5 deals with text categorization and the similarity measures used. Section 6 refers briefly to ontologies and Section 7 concludes the paper.

Throughout the paper, "terms", "features" and "tokens" are used interchangeably according to context. The same stands for the words "text" and "document".

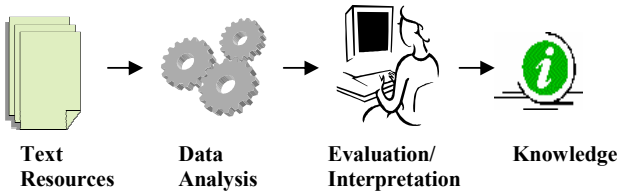

Figure 1. Text mining process

2. TEXT MINING MOTIVATION

The objective of text mining is the discovery of new knowledge within text collections. The magnitude of applications is significant.

In the biomedical field, most of the information is stored in text format so, association of terms and ideas is highly needed [2; 17; 35]. Swanson and Smalheiser [73; 74] were among the first to observe linkages between text collections, and conclude a medical cause and effect hypothesis that was not then known in the medical academia. This proves that the analysis of correlations of information across text collections is advantageous in the biomedical sector since unknown causes of diseases can be identified and as a result new medical treatments can be found. Of course, we should note that a lot of biomedical data is also stored in relational databases and the results of text mining can be used to facilitate further integration, update and querying of these sources.

Text mining tools and methodologies have a lot to offer to data integration tasks. They enable the identification of similarities between text attributes that originate from different sources, reducing in this way the uncertainty and improving the data integration accuracy. Similarity measures in text mining extend beyond string-based similarity metrics. They may take into account syntactic and semantic information and they may be applied to words, phrases or even bigger pieces of text. Selecting the most appropriate distance measure remains an important issue in the field. Since semantics is part of text mining, the semantic representation of text sources is more direct and the discovery of semantic mappings between the various sources and the mediated schema [31] is more straightforward. The benefits of text mining to data integration during the merging of two companies can also be seen in [23].

During data integration, issues such as record linkage and data cleaning are significant and they can also profit from the use of text mining approaches. Reducing redundant information and matching same entities across different sources and various representations, can be improved by using distance measures introduced in the text mining field. Semantics can help in dealing with incomplete information and erroneous data.

The applications of text mining can extend to any sector where text documents exist. For instance, history and sociology researchers can benefit from the discovery of repeated patterns and links between events, crime detection can profit by the identification of similarities between one crime and another [23], and unsuspected facts found in documents may be used in order to populate and update scientific databases.

Text mining can definitely facilitate the work of researchers. It can allow them to find related research issues to the ones they are working on, retrieve references to past papers and articles which may have been forgotten and discover past methodologies that may add on the nowadays research. Text mining may also reveal whether links exist between two different research domains without requiring the effort to understand the documents in both domains.

Another research field that may benefit from text mining is that of Information Retrieval since it is often required to execute queries that need the identification of semantic relations between texts. The application of text mining to IR may also improve the precision of IR systems [84] and reduce the number of documents that a single query returns.

Various other tasks can profit from text mining techniques. Examples consist of updating automatically a calendar by extracting data from e-mails [27; 48; 78], identifying the original source of a news article [49], monitoring inconsistencies between databases and literature [54]. Finding out such inconsistencies requires the collaboration of database as well as text mining techniques. Missing database values could be filled in by data discovered and retrieved from relevant literature.

Text categorization techniques may also be part of text mining. They intend to organize a set of texts, identify the structure of a text collection and group documents according to their common features. In this way, unstructured repositories obtain some structure, the labeling, search and browsing of documents is enabled [68] and the data analysis becomes efficient and effective.

3. TEXT MINING AND NLP

The majority of concerns in text mining are posed by the particularities of natural language. In this section, we will refer to the components of a language and the associated issues. We will focus more on the semantic rather than the statistical techniques since it seems that the statistics alone are not sufficient for the mining of a text [83].

A language consists of an alphabet, a grammar and a set of rules that define the syntax. The alphabet is the set of symbols used by a language. According to [70], the letters and the sequences of letters have a statistical structure which means that they do not all appear with the same frequency. The grammar of a language is the set of rules that define how the symbols of the alphabet can interact with each other, while the syntax consists of the rules that capture the way the words can be united to form a sentence. According to Sapir [65], "all grammars leak" since people tend to use the language freely, without adhering to rules. This stands, for example, in e-mails and chat dialogues where ill-formed expressions are often used for the sake of simplicity.

Describing text by a grammar can lead to erroneous identifications of lexical tokens, inability to capture syntactic text errors or identify certain items such as names [78]. Basic syntactic rules can though capture key patterns in the language structure. The syntactic rules depend on the language of the text and it is better if they are defined by linguists [46]. The rules may contain some uncertainty as in the case of a Probabilistic Context Free Grammar (PCFG) whose rules have probabilities attached to them.

3.1 Text Mining Issues

Some of the natural language issues that should be considered during the text mining process are listed in Table 1 and they are discussed in this paper.

Table 1. Issues of text mining

Issue	Details
Stop list	Should we take into account stop words?
Stemming	Should we reduce the words to their stems?
Noisy data	Should the text be clear of noisy data?
Word Sense Disambiguation	Should we clarify the meaning of words in a text?
Tagging	What about data annotation and/or part of speech characteristics?
Collocations	What about compound or technical terms?
Grammar / Syntax	Should we make a syntactic or grammatical analysis? What about data dependency, anaphoric problems or scope ambiguity?
Tokenization	Should we tokenize by words or phrases and if so, how?
Text Representation	Which terms are important? Words or phrases? Nouns or adjectives? Which text model should we use? What about word order, context, and background knowledge?
Automated Learning	Should we use categorization? Which similarity measures should be applied?

One concern relates to the generation of a stop list. Having a stop list which usually contains high frequency words such as 'a', 'the' or 'of' that are to be ignored from a text, is an idea inherited by IR. In IR, it has been widely used due to performance improvement. In text mining, though, it is not as useful since common terms seem to provide information [62; 81]. Common stop words can even help in clarifying the semantics of a text segment. For instance, in the phrase "she was arrested", the words "she" and "was" are important. The first one identifies the person that received the action and the second one, although common as a stop word, is actually a keyword since the phrase without it -"she arrested" - has a totally different meaning.

Stemming or in other words lemmatization, on the other hand, does not seem to be dependent on the domain but on the language of the text. It reduces a word to its root e.g. it replaces 'reading' or 'reader' by 'read', so that similarity detection can be achieved. The task fulfilled by stemming is in a way analogous to number normalization so that comparisons are achieved. Even in the case of stemming, though, it can be argued that applying lemmatization techniques to a piece of text may affect the semantics.

Correcting spelling mistakes and replacing acronyms and abbreviations can also be part of the text mining process in order to eliminate noisy data before the main processing starts. During this text cleaning, the use of a dictionary or thesaurus may be useful. The text cleaning, here, differs from the data cleansing in the databases field in that it is mainly about misspellings rather than schema inconsistencies, integrity constraints or invalid data.

The automatic NLP also needs to deal with the ambiguity of the language. The word sense disambiguation (WSD) problem, which is about finding out the most probable meaning of a polysemous word, is one issue. One approach to solve this is by considering the context in which a particular word is found. This process may include obtaining the grammatical category of a word, for instance, detecting if the word 'play' is a noun or a verb in a specific

phrase. There are two types of disambiguation; the supervised and the unsupervised. The supervised one is often carried out with the help of a dictionary or a thesaurus. In the unsupervised disambiguation, the different senses of the word are not known. Yarowsky [82] has presented an unsupervised approach to the WSD problem with high accuracy results. WSD techniques may be applied to some reference reconciliation tasks in order to detect references of the same entity. This, though, assumes that the particular entities incorporate some kind of semantics.

Tagging concerns the application of part of speech (PoS) tags, XML or SGML mark-up to corpora. PoS tags capture certain syntactic categories such as nouns, verbs and adjectives, and they can be used for the identification of noun phrases or other parts of speech. In case unknown words exist in a text, there are ways to find the most probable tags since the possibility of some tags having unknown words is not the same for all of them [46]. The Brown corpus [11] and the Penn Treebank [58] are text collections that are tagged by grammatical tags.

Another issue is that of the collocations that may exist in a text. These are phrases, such as "radio therapy", that make sense only if considered as a whole. In collocations, the meaning of the whole is greater than the meaning of the sum of its parts. In other words, the semantics of a collocation are not equal to the semantics of its parts, so studying the properties of the single words does not convey the meaning of the collocation itself. A syntactic analysis may lead to collocation discovery in a text.

If a syntactic analysis takes place, the order in which the words appear in the text is an issue that should be considered. The parsing of a sentence could start either by the beginning or by the end of it and sometimes it could even start by the main verb since this usually directs the development of a sentence.

Tokenization is an issue that regards the splitting of a text into units and it may take place during the data analysis. A text can be tokenized in paragraphs, sentences, phrases of any length and single words. The delimiters used vary. A common delimiter is the space or the tab between words. Punctuation marks can be used as well, such as full stops, exclamation marks or commas. Particularities of the delimiters may need to be considered. For example, the full stop is used in abbreviations so apparently it does not always mark a sentence ending. Also, considering the space as a tokenization symbol will keep the compound phrases apart.

Common stop words such as 'and', 'the' or 'a' can be considered as delimiters [6] or even specific domain stop words (e.g. technical terms) dependent on the domain the text belongs to. The terminology is a sensitive issue whose extraction has been dealt with in some papers [10; 20]. Bourigault [10] defines the technical terms as noun phrases which have a meaning even if they exist outside a text.

Tokenization can be done in paragraphs or sections. This is often referred to as discourse segmentation. In [43] text segments are found by calculating the lexical cohesion between word lists. Changes in the lexical cohesion can be considered as segment boundaries. Another example is the TextTiling algorithm [33] which partitions a text into subtopics. The algorithm splits the text in phrases of certain length, it checks the term repetition and the lexical similarity between these phrases, and it defines the thematic boundaries wherever the similarities change dramatically. The evaluation of this algorithm shows that human judgment is reflected in the way the segmentation is done.

4. TEXT REPRESENTATION

Text representation depends on the task in hand and it allows for easier and more efficient data manipulation. Some examples of tasks and how they have been modelled are discussed in this section. The reason why text representation is dealt with on a separate section is because there has been a lot of discussion in the current literature and many models have been proposed.

Similarly to database models, text models intend to capture the relationships between data. Text models, though, describe free text and not structured data. The relationships may be derived by statistical ways and not necessarily through logical associations. Moreover, the operations of a text model are usually between vectors and the data do not comply with a logical schema.

Text representation may serve as an intermediate step between raw text data and database models. For example, organizing data found in documents into relational tables requires some text and semantic analysis that is applied on text models. Database models are used for data storage and curation, while text representation models permit the discovery of similarities among texts, topic identification and text linkages that may not be obvious.

The most widely used representation is the Vector Space Model (VSM) [64]. According to this, the text is described by a vector whose dimension is the number of text features and its content consists of a function of the frequencies with which these features appear in the corpus or corpora. This model is also referred to as the bag-of-words model because the order and the relations between the words are ignored.

The majority of representations proposed are an extension of the VSM model. There are some representations that focus on phrases instead of single words [6; 15; 51], some that give importance to the semantics of words or the relations between them [16; 42; 59] and others that take advantage of the hierarchical structure of the text [4]. These different approaches are discussed in the following sections.

4.1 Feature Extraction

A lot of discussion dating back to IR concerns whether frequent or rare terms are more suitable to represent a text and whether single words or phrases are better terms.

The frequency with which a term appears in a corpus or corpora can clarify the significance of this term in a specific document. A frequency measure can be binary to underline absence or presence, it can vary from 0 to 1 or it can be given by a mathematical function. Normalization is usually needed so that the length of the document and the number of unique terms is taken into account. For instance, in a very small text that contains only 10 unique terms, all the terms are important regardless of their frequency.

An excellent example of a statistical index that gives a quantitative answer as to whether a term, being frequent in one document, is really worth being extracted when it is also frequent in a collection of documents is the well-known *tf-idf* index. This index promotes terms that appear many times in a single document but very few times in a collection of them [67].

Statistical information can be gathered either for distinct words or phrases. Lewis [45] supports that words provide better statistical quality. This is because the words which constitute a phrase may appear multiple times in a document while the phrase itself may be present only once and as a result the frequencies can be misleading.

On the other hand, phrases provide more semantic information than the single words because they give an idea of the context. A word is characterized by the company it keeps [24] and since words may have multiple meanings, we do need to know at least the phrase that contains the word in question, so as to approach the semantics with higher certainty. The experiments of Blake and Pratt [6] demonstrate the benefit of using special phrases and concepts over words for the representation of medical texts.

The interest in collecting statistical and semantic information has led to the issue of choosing between statistical and syntactic phrases [15; 51; 67]. A statistical phrase is a phrase that appears in a statistical way inside a text, while a syntactic one is a phrase whose grammar and syntax rules reveal some semantics. A statistical phrase is retrieved by statistical methods while a syntactic phrase can be extracted using linguistic methods.

Salton [63] combines statistical and syntactic phrases for book indexing. He carries out a syntactic analysis of the sentences of a document and then he extracts from the syntactic tree some of the existing noun phrases. He gives importance to the frequency of terms within a document and within a collection of documents and he marks the noun phrases of the document title.

In Table 2, the advantages and disadvantages of considering words or phrases as terms are shown.

Table 2. Advantages and disadvantages of words and phrases

	ADVANTAGES	DISADVANTAGES
WORDS	• good statistics • synonyms • existence of tools / algorithms (e.g. WordNet [79], WSD algorithms)	• no context information • problem with collocations
PHRASES	• context information • semantic quality • collocations can be captured	• average statistical quality

When we have to make a decision between using words or phrases, the important is not which kind of phrases is better but whether they have to offer something more than the single terms [28; 51]. As it can be seen from Table 2, phrases fill in the gaps that words cannot cover and vice versa. Phrases inform about the context, while words provide higher statistical quality. Therefore, it seems that a combination of both is the best way to capture text features.

4.2 Representation Models

The VSM model can only capture information related to the frequencies of text features. Alternative models have been proposed in the existing literature covering special cases and various tasks.

A structured text having sections, paragraphs and sentences is better than a totally unstructured set of words [43]. Therefore, considering text properties such as the location of a word in a text can lead to a better representation. The words present in the title of a document have usually higher significance. It can also be considered that the first paragraph of a document is often an introduction while the last one is usually a conclusion.

The context of a term is also a useful piece of semantic information. Rajman and Besançon [59] have represented the context as a vector that contains the co-occurrence frequencies between a term and a predefined set of indexing features. Nenadic and Ananiadou [54] use context patterns in biomedical documents. These patterns are in the form of regular expressions and they contain PoS tags and ontology information.

N-grams can also be used to discover the context of a word. Caropreso et al. [15] have used n-grams in order to represent and categorize text. They replace some unigrams with bigrams and they use functions such as document frequency and information gain in order to score the n-

grams extracted from the text. Their results are better when bigrams are used over unigrams. Similar results have been shown in [52].

Cimiano et al. [16] model the context of a term as a vector of syntactic dependencies found in a text corpus. They extract a concept hierarchy by applying a method based on the formal concept analysis. A linguistic parser extracts the syntactic dependencies. Then, they assign weights to these dependencies and they create a lattice of formal concepts. The problem is that the size of this lattice increases according to the number of concepts.

Kehagias et al. [42] have experimented by using sense-based representations where the features chosen are not single words but the meanings of them. The results of the research have not shown improvement in the accuracy of text classification compared to the accuracy achieved by the word-based representations.

Carenini et al. [14] propose a hierarchy of extracted features. They attempt to map texts that describe product reviews to a UDF (user-defined features) hierarchy. The advantage of using such a taxonomy, as it is reported in the paper, is adding background user knowledge to the model and reducing the redundancy. The disadvantage is that for every (sub-) domain a UDF hierarchy has to be created.

Similarly to Carenini et al. [14], Bloehdorn et al. [8; 9] match the syntax of sentences found in a text against a library that contains regular expressions patterns. The concepts found are added to the bag-of-words model creating in this way a "hybrid feature vector".

Recently, matrix space models (MSM) have been proposed for text representation [4]. This representation is based on the idea that a document is a hierarchy of document extracts e.g. sections, paragraphs and sentences and as a result term-by-section, term-by-paragraph and term-by-sentence matrices can be respectively created. In [4] they deal with term-by-sentence matrices. Their experiments regard query evaluation for IR and the results are close to the ones achieved by Latent Semantic Indexing (LSI) with low computational cost. Accuracy is said to be high for multi-topic documents. The advantage of this kind of matrix representation over the VSM and the LSI model is that it "remembers" the intermediate steps of the construction of the final matrix.

In Table 3, we present some of the approaches covered by the existing literature together with the text units they focus on, the representation types they use and the task they are dealing with.

Table 3. Text representation approaches

Approach	Terms	Representation Type	Objective
Antonellis and Gallopoulos [4]	Sentences	term-by-sentence matrices	Text mining
Blake and Pratt [6]	words, phrases, concepts	association rules	Representation of medical texts
Bloehdorn et al [8; 9]	words and concepts	combination of bag-of-words and concept hierarchy	Text clustering and classification
Carenini et al [14]	concepts	Hierarchy	Feature extraction
Caropreso et al [15]	phrases	n-grams	Text categorization
Cimiano et al [16]	concepts	concept hierarchy	Automatic acquisition of a taxonomy
Kehagias et al [42]	word senses	sense-based vector	Text categorization
Mladenic and Grobelnik [52]	phrases	n-grams	Text learning
Rajman and Besançon [59]	words and compounds	Vector	IR
Salton [63]	noun phrases	Tree	Book indexing
VSM [64]	words	Vector	IR
Varelas et al [77]	words	Tree	Semantic similarity for IR

5. CATEGORIZATION

The data analysis of corpora often involves the identification of the inherent structure of the document collection, the labeling of documents and text segments and the generation of clusters according to a similarity measure. The task that deals with the organization of an unstructured collection of documents to a structured repository is called text categorization and it aims at facilitating storage, search and browsing [68].

Text mining tools and algorithms can benefit from the organization of documents into categories because it is simpler to analyze structured texts. This means that categorization can be an intermediate step of the text

mining process and it may enable the discovery of links and patterns not easily noticeable between the documents.

The categorization task can be supervised or unsupervised, dependent on whether the groups or categories are known from the beginning or not.

During a supervised classification process, the first step is to define the documents that will be used. There are three sets of documents; the training set with annotated documents, the development set used to test the classifier before it is completed, and finally, the test set that comprises the documents which will evaluate the performance of the classifier. The intersection of these three sets should be the empty set. Subsequently, the representation of these documents and categories is decided. The training of the model begins, the parameters are tuned and the model is applied to the test documents. The computational cost of text annotation and the difficulty in obtaining training data, has led the researchers to alternatives such as semi-supervised techniques [3; 18; 55] that use a small set of labeled data.

In the unsupervised case which is called clustering, there are no labeled documents. A similarity measure is defined and the documents are compared with each other in order to be divided into clusters. The objective is to achieve a low inter-cluster and a high intra-cluster similarity.

The text categorization algorithms can be applied in many cases. The thematic labeling of a document collection, the classification of movie text reviews into positive and negative ones, the distinction of spam e-mails from the rest and the automatic organization of Web pages are examples of categorization.

In this section, the word 'categorization' is used to refer to both supervised and unsupervised cases.

5.1 Categorization Tasks

The categorization task may vary according to the intra-document or inter-document associations that need to be captured. Thus, the categorization goal should be clear before deciding which algorithm to apply. The goal can be the identification of the documents that deal with the same topic, the semantic orientation of a review, the selection of the articles written by the same author, the disambiguation of the meaning of a polysemous word in a text or even the distinction between interesting and not interesting texts based on the preferences of a person. In the existing literature, various categorization cases have been considered. Here we briefly discuss some of them.

In the case of thematic categorization, the focus is usually on noun terms that may characterize a topic. Automated learning has been the machine learning approach to this categorization type. Several learning algorithms have been applied. Yang and Liu [80] have presented a comparison of some of them, stating that SVM, k-nearest-neighbor and LLSF perform better than neural networks and naïve Bayes. Dumais et al. [21] show that SVM are better than naïve Bayes and decision trees. Accuracy is reported to be even better in the experiments of Apte et al. [5] who have used multiple decision trees produced by the boosting or the bagging approach. Sebastiani [67] has attempted to draw a conclusion as to which classifier is the best by taking into account the experiments of various authors as well as the differences during these experiments in steps like pre-processing or parameter tuning. His conclusion is that boosting-based [66], example-based (e.g. k-NN), based on regression methods (e.g. LLSF) classifiers and SVM are regarded as top classifiers. Neural networks and online linear classifiers (e.g. perceptron, WIDROW-HOFF) follow the aforementioned top ones and they are considered to be very good. Recently, the Latent Dirichlet Allocation [7] model has been proposed in order to point out which topics are discussed in a document collection.

A sentiment classification task deals with the classification of a document according to the subjective opinion of the author [37]. In this case, the focus is on finding the semantic orientation of a word, namely its positive or negative attitude. Hatzivassiloglou and McKeown [32] focus on adjectives and they study phrases where adjectives are connected with conjunction words such as 'and' or 'but'. They use a log-linear regression model so as to clarify whether two adjectives have the same orientation and then they divide the adjectives into two subsets considering the subset with higher frequency to be the "positive" one. Turney and Littman [75] highlight the importance of context, since a positive word may have a negative meaning in a metaphorical or ironic context. In order to discover the semantic orientation of words, they use an LSA-based measure to find out the statistical relation of a specific word towards a set of positive or negative words.

Kamps et al. [39] use WordNet [79], a lexical database, to detect the semantic orientation of adjectives and they calculate the semantic distance as the path length between two graph nodes which contain words. Pang et al. [57] deal as well with sentiment classification for movie reviews. Their experiments show that algorithms such as SVMs, naïve Bayes and Maximum Entropy, that give good results in thematic categorization, do not perform as well during sentiment classification. Additionally, they point out that the presence or absence of a word seems to be more indicative of the content of a review rather than the frequency with which a word appears in a text.

Sentiment classification seems to be more difficult than the topic-based one and it cannot be based on just observing the presence of single words. In [75] it is mentioned that sarcasm may be an obstacle for the clarification of the

semantic orientation of a text. More sophisticated methods need to be employed so as to differentiate between the subjective and objective opinion of a reviewer or between the objective description of a movie and references to other people's comments. An initial step in recognizing subjective and objective statements is presented in [37] where they focus on identifying comparative sentences.

5.2 Measuring Similarity

Another part of a categorization task is the selection of a similarity measure in order to identify the mutual characteristics of various documents. Dissimilarity measures, which focus on how dissimilar two concepts are, may also exist. Any dissimilarity function can be transformed into a similarity one but the opposite does not always stand [76]. The similarity measures proposed in the existing literature can be divided into two categories; the statistical and the semantic ones.

From the statistical point of view, measuring the term frequency and the co-occurrence frequency has been widely used. According to Resnik [60], the co-occurrence frequency is a proof of relatedness. Hoskinson [36] uses a combination of document co-occurrence and term frequency measures in order to classify concepts which are defined as the most frequent terms. Among the most popular statistical measures are the cosine coefficient, the Euclidean distance and the chi-square which are used by text classifiers in order to compare two vectors.

The semantic-based similarity measures the distance between the meanings of two terms. WordNet [79] is often used in order to find out word senses or semantic relations between wording features. It is an electronic database of the English language that consists of words organized into subsets according to their meaning. These subsets are synonym sets called synsets, and they are linked by relations such as inheritance or part-whole relationships. For languages other than English, there are some projects found in the Global WordNet Association web site [29] such as EuroWordNet [22].

Varelas et al. [77] have used the WordNet XML Web-Service to create XML tree structures for terms that exist in documents or queries, with the intention of measuring the semantic similarity between them. They calculate the information content of each term and then they measure the similarity between two terms with the help of WordNet.

Measuring the similarity between two nodes in WordNet or a similar hierarchy can be done in many ways. The edge-counting method measures the path length from one node to another. To avoid problems that appear by not taking into account the density of the hierarchy, an information content measure has been used [61; 69] in some cases, showing improvement in the results. The information content measures the amount of information that can be given by a concept or a term. The more abstract a concept is in a hierarchy, the higher it is and the less information it contains. As a result its information content has a low value. Additionally, the more information is shared between two words or two concepts, the more similar they are. Budanitsky and Hirst [12] have compared some similarity WordNet-based measures concluding again that using the information content is better than just counting the path length. According to Resnik [61], even in the case of an information content measure, word senses have to be considered since two words from the slang vocabulary can be wrongly considered similar.

Similarity measures have also been explored between phrases or blocks of phrases. Hearst [33] identifies lexical cohesion relations between pseudo-sentences of certain length by using a cosine measure and taking into account the frequency of terms in each block of sentences. Metzler et al. [49] have explored sentence-to-sentence similarity in an attempt to discover the original source of a document. They define five similarity levels; "unrelated", "on the general topic", "on the specific topic", "same facts" and "copied" and they apply similarity measures such as word overlap, frequency measures and probabilistic ones. In their initial experiments the word-overlap seems to outperform.

The aforementioned similarity measure types, as well as the units to which they can apply are summarized in Figure 2.

For the purpose of evaluating the similarity measures proposed, most researchers compare their similarity scores with the human judgement scores. The closer the scores are to the human results, the better the measure is. Varelas et al. [77], as well as, Seco et al [69] use the human scores gathered by the experiment of Miller and Charles [50].

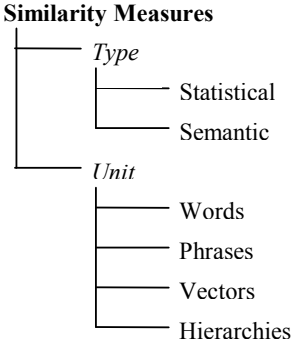

Figure 2. Similarity measures

Resnik [61] replicates the experiment of Miller and Charles using the same nouns they had used. Budanitsky and Hirst [12] agree that comparing against human answers is the best way but they point out that the human judgements consist of

a small set of answers that reflect the tendency of the users to give the most dominant sense to a word.

6. ONTOLOGIES AND TEXT MINING

Ontologies have been proposed for handling semantic heterogeneity when extracting information from various text sources such as the Internet [1]. Their importance lies in the fact that they represent a schema for a particular domain, clarifying in this way technical terms that appear in a text or specifying the relationship between certain domain concepts. During a text mining process, ontologies can be used to provide expert, background knowledge about a domain. In [71] the use of ontologies for text mining tasks is discussed.

An ontology consists of concepts, concept-relations, axioms and instances. The selection of concepts depends on the task and the domain information that needs to be captured. As a result, before defining the concepts, it is important to know to which questions a response will be wanted once the ontology is built.

Ontologies differ from database or XML schemas. They mainly represent a domain and the technical terms that surround it and they can be used at any time a semantic analysis is needed. There are neither data types involved nor integrity constraints. The semantics are defined based on the specific domain concepts and they are not dependent on a particular application as in databases. Contrary to XML, the semantics in ontologies are not user-defined but they follow the rules imposed by the relevant domain.

The existence of generic ontologies is limited. Their purpose is to be reusable but they are not so useful [30]. SUMO (Suggested Upper Merged Ontology) [72] is one such upper ontology. It is a foundation ontology that consists of abstract and general concepts independent of any domain. Based on its structure, domain-specific ontologies can be built. Niles and Pease [56] have attempted to map SUMO concepts to WordNet synsets.

Sebastiani [68] claims that until now ontologies do not seem to benefit the text categorization, although at the same time there has not been an exhaustive research on this matter. Looking at this issue, from a different point of view, it seems that text mining can offer more to the generation and update of ontologies rather than the ontologies to text mining.

Statistical or Machine Learning techniques have been dealing with the problem of extracting ontologies from text [13]. The biggest challenge, however, that becomes eminent both while processing the document and the elicited ontology is the true semantics of the content and the result. In many cases, authors rely on simple relationships among the members of extracted ontologies and the overall results need still rely on advance natural language techniques and human judgment [25; 26]. The aforementioned work has been realized in the Ontogen tool, which enables the construction of an ontology based on machine learning and text mining techniques.

Apart from ontologies, conditional random fields (CRFs) have been proposed for providing background knowledge to a system [44], segmenting and labeling sequential data [47]. A CRF specifies the probabilities of possible label sequences given an observation sequence, and it can be used when patterns may not always stand [44]. The probabilities may depend on current, past and future observations. In [19], CRFs are presented as graphical models which enable the extraction of patterns of association from a text. They are used in two ways; initially they are applied so that family relationships are extracted from biographical texts to form a graph and then this graph is fed into the CRF again in order to re-extract associations. In this case, the graph plays the role of an ontology that has been generated by the data itself.

Semi-CRFs extend CRFs by using multi-word instead of single-word segments. In [47], semi-CRFs are used in order to extract entities from unstructured data and integrate them into a relational database while taking into account the key constraints.

7. CONCLUSION

The continuous expansion of textual data has led to the need for text mining techniques and methodologies in order to better study and exploit the content-oriented relations between text documents. Text mining is an open research area where the issues discussed in this paper are still not finalized. For the purpose of approaching these issues, it is better to clarify the mining objective before the data analysis starts, since each task has different requirements.

Taking into account the language a text is written in is important since the language highlights the morphological or syntactic analysis needed. Moreover, the domain of a text collection underlines what technical terms may be present in the text or which words are redundant. Certain decisions and approaches may not be suitable for every type of text [38] due to the fact that term distribution varies between abstracts, articles, and collections of articles.

NLP interacts with text mining. Measurable results, though, are needed so as to find out which NLP techniques can be applied to what text mining applications [40; 41]. In general, we should think carefully before reducing the feature list, removing stop words or applying lemmatization techniques to the texts. Noisy data may also prevent some techniques from working efficiently, so they should be corrected before the processing starts.

The ambiguity is a characteristic of free text. As a result, word sense disambiguation will need to take place during

the processing of certain phrases or words that are considered important for the text semantics. Identifying collocations can also help in disambiguating the meaning of some phrases.

The representation of a text is a crucial issue. Most of the researchers agree that an extension of the bag-of-words model is essential but there is still no agreement as to which kind of text properties and features should be taken into account. The attributes of the representation model depend on what kind of information we want to capture. Background knowledge, word context, and word or phrase location can be some desired properties. The text features selected can be identified with the help of tokenization and dimension reduction techniques. It is important, though, to consider where features will be looked for since certain document sections, such as the "References", should better be avoided [83]. Using a combination of words and phrases is recommended. Concepts can be part of the representation as well, but more research is required on this matter.

Classifying a text collection into categories may enable the text processing. The similarity measures chosen for the categorization depend on which type of semantic or statistic distance between documents needs to be captured. The measures can apply to words, phrases, vectors or hierarchies. A combination of both syntactic and semantic measures may be considered.

New, previously unknown knowledge can also be identified by studying the semantic relations between the information stored in databases and the existing literature. This is an open issue that can be explored with the help of text mining and database methodologies.

8. REFERENCES

[1] Abadi, D., Marcus, A., Madden, S., and Hollenbach K. 2007. Scalable Semantic Web Data Management Using Vertical Partitioning. In *Proc. of the 33rd VLDB*, Austria, pp. 411-422.

[2] Ananiadou, S., Chruszcz, J., Keane, J., Mcnaught, J., and Watry, P. 2005. The national centre for text mining: aims and objectives. In *Ariadne 42*, Jan. 2005.

[3] Ando, R.K., and Zhang, T. 2005. A high-performance semi-supervised learning method for text chunking. In *Proc. of the 43rd ACL*, Ann Arbor, pp 1-9.

[4] Antonellis, I., and Gallopoulos, E. 2006. Exploring term-document matrices from matrix models in text mining. In *Proc. of the SIAM Text Mining Workshop 2006, 6th SIAM SDM Conference*, Maryland.

[5] Apte, C., Damerau, F., and Weiss, S. 1998. Text mining with decision rules and decision trees. In *Conference on Automated Learning and Discovery*, Carnegie-Mellon University.

[6] Blake, C., and Pratt, W. 2001. Better rules, fewer features: a semantic approach to selecting features from text. In *Proc. of IEEE DM Conference (IEEE DM)*, San Jose, CA, pp. 59-66.

[7] Blei, D., Ng, A., and Jordan, M. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research 3*, pp. 993–1022.

[8] Bloehdorn, S., Cimiano, P., and Hotho, A. 2005. Learning ontologies to improve text clustering and classification. In *Proc. of the 29th Annual Conference of the German Classification Society (GfKl),* Magdeburg, Germany, pp. 334-341.

[9] Bloehdorn, S., and Hotho, A. 2004. Text classification by boosting weak learners based on terms and concepts. In *Proc. of the 4th ICDM*, Brighton, UK, pp. 331-334.

[10] Bourigault D. 1992. Surface grammatical analysis for the extraction of terminological noun phrases. In *Proc. of the 14th COLING-92*, Nantes, pp. 977-981.

[11] Brown Corpus. http://helmer.aksis.uib.no/icame/brown/bcm.html

[12] Budanitsky, A., and Hirst, G. 2001. Semantic distance in WordNet: an experimental, application-oriented evaluation of five measures. *Workshop on WordNet and Other Lexical Resources, 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, PA.

[13] Buitelaar, P., Cimiano, P., and Magnini, B. 2005. Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, USA.

[14] Carenini, G., Ng, R.T., and Zwart, E. 2005. Extracting knowledge from evaluative text. In the *3rd KCAP*, Banff, Alberta, Canada, pp. 11-18.

[15] Caropreso, M.F., Matwin, S., and Sebastiani, F. 2001. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. In *Text Databases and Document Management: Theory and Practice*, AMITA G. CHIN, Ed. Idea Group Publishing, Hershey, PA, 78-102.

[16] Cimiano, P., Hotho, A., and Staab, S. 2005. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, 24, pp. 305-339.

[17] Cohen, K.B., and Hunter, L. 2004. Natural language processing and systems biology. In *Artificial Intelligence methods and tools for systems biology,* Dubitzky and Pereira, Springer Verlag.

[18] Cong, G., Lee, W., Wu, H., and Liu, B. 2004. Semi-supervised text classification using partitioned EM. In *9th DASFAA,* Jesu Island, Korea, pp., 482-493.

[19] Culotta, A., Mccallum, A., and Betz, J. 2006. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Human Language Technology - North American Chapter of the Association for Computational Linguistics Annual Meeting*, NY, 296-303.

[20] Daille, B., Gaussier, E., and Langé, JM. 1994. Towards automatic extraction of monolingual and bilingual terminology. In *Proc. of the 15th International Conference on Computational Linguistics,* 515-521.

[21] Dumais, S., Platt, J., Heckerman, D., and Sahami, M. 1998. Inductive learning algorithms and representations for text categorization. In *Proc. of the 7th CIKM*, Bethesda, MD, 148–155.

[22] *EuroWordNet*. http://www.illc.uva.nl/EuroWordNet

[23] Fan, W., Wallace, L., Rich, S. and Zhang, Z. 2006. Tapping the power of text mining. In *Communications of the ACM 49(9)*, pp. 76-82.

[24] Firth, J.R. 1957. A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, Philological Society, Oxford, 1-32. Reprinted in *Selected papers of J.R.Firth 1952-1959,* Longman, London.

[25] Fortuna, B., Grobelnik M., and Mladenic D. 2006. Background Knowledge for Ontology Construction. In *Proc. of the 15th International Conference on WWW*, Edinburgh, Scotland, UK, pp. 949-950.

[26] Fortuna, B., Mladenic, D., and Grobelnik, M. 2005. Semi-automatic Construction of Topic Ontologies. In *Joint International Workshops, EWMF 2005 and KDO 2005, on Semantics, Web and Mining*, Porto, Portugal, pp. 121-131.

[27] Freitag. D. 1998. *Machine Learning for Information Extraction in Informal Domains.* Ph.D. thesis, Carnegie Mellon University.

[28] Furnkranz, J., Mitchell, T., and Riloff, E. 1998. A case study in using linguistic phrases for text categorization on the WWW. *Working Notes of the AAAI / ICML, Workshop on Learning for Text Categorization,* Madison, WI, pp. 5-12.

[29] *Global WordNet Assoc*. http://www.globalwordnet.org/

[30] Gomez-Perez, A., and Benjamins, V.R. 1999. Overview of knowledge sharing and reuse components : ontologies and problem-solving methods. In *Proc. of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods*, Stockholm, Sweden.

[31] Halevy, A., Rajaraman, A., and Ordille, J. 2006. Data Integration: The teenage years. In *Proc. of the 32nd VLDB*, Korea, pp. 9-16.

[32] Hatzivassiloglou, V., and Mckeown, K.R. 1997. Predicting the semantic orientation of adjectives. In *Proc. of the 35th ACL and the 8th Conference of the European chapter of the ACL*, New Brunswick, NJ, pp. 174-181.

[33] Hearst, M.A. 1994. Multi-paragraph segmentation of expository text. In *Proc. of the 32nd ACL*, Las Cruces, NM, pp. 9-16.

[34] Hearst, M.A. 1999. Untangling text data mining. In *Proc. of the 37th ACL*, College Park, MD, pp. 3-10.

[35] Hirschman, L., Park, J.C., Tsujii, J., Wong, L., and Wu, C. 2002. Accomplishments and challenges in literature data mining for biology. In *BioInformatics*, *18(12)*, pp. 1553-1561.

[36] Hoskinson, A. 2005. Creating the ultimate research assistant. *IEEE Computer, 38(11)*, pp. 97-99.

[37] Jindal, N., and Bing, L. 2006. Identifying comparative sentences in text documents. In *Proc. of the 29th SIGIR*, Seattle, USA, pp. 244-251.

[38] Kageura, K., and Umino, B. 1996. Methods of automatic term recognition. *Technology Journal, 3(2)*, pp. 259-289.

[39] Kamps, J., Marx, M., Mokken, R.J., and Maarten De Rijke 2004. Using WordNet to measure semantic orientations of adjectives. In *Proc. of the 4th LREC*, vol. IV, European Language Resources Association, Paris, 2004, pp. 1115-1118.

[40] Kao, A., and Poteet, S. 2004. Report on KDD conference 2004 panel discussion - can natural language processing help text mining? *SIGKDD Explorations 6(2)*, Dec. 2004, pp. 132-133.

[41] Kao, A., and Poteet S. 2006. Text mining and natural language processing – Introduction for the special issue. *SIGKDD Explorations 7(1)*, June 2006, pp. 1-2.

[42] Kehagias, A., Petridis, V., Kaburlasos, V.G., and Fragkou, P. 2001. A comparison of word- and sense-based text categorization using several classification algorithms. *Journal of Intelligent Information Systems*, *21(3),* pp. 227-247.

[43] Kozima, H. 1993. Text segmentation based on similarity between words. *In Proc. of the 31st ACL*, Columbus, Ohio, USA, pp. 286-288.

[44] Lafferty, J., Mccallum, A., and Pereira, F. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proc. of the 18th ICML*, Williamstown, MA, pp. 282-289.

[45] Lewis, D.D. 1992. An evaluation of phrasal and clustered representations on a text categorization task. In *Proc. of SIGIR,* Copenhagen, Denmark, pp. 37–50.

[46] Manning, C., and Schutze, H. 1999. *Foundations of Statistical Natural Language Processing.* The MIT Press, Cambridge, Massachusetts.

[47] Mansuri, I.R, and Sarawagi, S. 2006. Integrating unstructured data into relational databases. In *Proc. of the 22nd ICDE*, 29.

[48] McCallum, A. 2005. Information Extraction: Distilling Structured Data from Unstructured Text. *ACM Queue, 3(9), November 2005*.

[49] Metzler, D., Bernstein, Y., Croft, W.B., Moffat, A., and Zobel, J. 2005. Similarity measures for tracking information flow. In *Proc. of CIKM*, Bremen, Germany, pp. 517-524.

[50] Miller, G.A. and Charles, W.G. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes 6 (1),* pp. 1-28.

[51] Mitra, M., Buckley, C., Singhal, A., and Cardie, C. 1997. An analysis of statistical and syntactic phrases. In *Proc. of the 5th International Conference "Recherche d' Information Assistee par Ordinateur" (RIAO)*, Montreal, CA, pp. 200-214.

[52] Mladenic, D., and Grobelnik, M. 1998. Word sequences as features in text-learning. In *Proc. of the 7th Electrotechnical and Computer Science Conference,* Ljubljana, Slovenia, pp. 145-148.

[53] Mooney, R.J., and Bunescu, R. 2005. Mining knowledge from text using information extraction. *ACM SIGKDD Explorations 7(1)*, June 2006, pp. 3-10.

[54] Nenadic, G., and Ananiadou, S. 2006. Mining semantically related terms from biomedical literature. In *ACM TALIP Special Issue on Text Mining and Management in Biomedicine, 5(1)*, pp. 22-43.

[55] Nigam, K., and Ghani, R. 2000. Analyzing the effectiveness and applicability of co-training. In the *8th CIKM,* Kansas City, MI, pp. 86-93.

[56] Niles, I., and Pease, A. 2003. Linking lexicons and ontologies: mapping WordNet to the suggested upper merged ontology. In *Proc. of the 2003 International Conference on IKE*, Las Vegas, Nevada, pp. 412-416.

[57] Pang, B., Lee, L., and Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proc. of the 2002 EMNLP,* pp. 79-86.

[58] *Penn Treebank*. http://www.cis.upenn.edu/~treebank/home.html

[59] Rajman, M., and Besançon, R. 1999. Stochastic distributional models for textual information retrieval. In *Proc. of 9th ASMDA*, Lisbon, Portugal, pp. 80-85.

[60] Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of the 14th IJCAI-95,* Montreal, QC, Canada ,pp. 448-453.

[61] Resnik, P. 1999. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research, 11,* pp. 95-130.

[62] Riloff, E. 1995. Little words can make a big difference for text classification. In *Proc. of the 18th SIGIR,* Seattle, WA, pp. 130-136.

[63] Salton, G. 1988. Syntactic approaches to automatic book indexing. In *Proc. of the 26th ACL*, NY, 120-138.

[64] Salton, G., Wong, A., and Yang, C.S. 1975. A vector space model for automatic indexing. In *Communications of the ACM 18(11)*, pp. 613-620.

[65] Sapir, E. 1921. Language: an introduction to the study of speech. HARCOURT BRACE & CO., New York.

[66] Schapire, R.E. 1999. A brief introduction to boosting. *In Proc. of the 16th IJCAI,* Stockholm, pp. 1401-1405.

[67] Sebastiani, F. 2002. Machine learning in automated text categorization. In *ACM Computing Surveys, 34(1)*, pp. 1–47.

[68] Sebastiani, F. 2006. Classification of text, automatic. In *The Encyclopedia of Language and Linguistics 14*, 2nd ed., Elsevier Science Pub., pp. 457-462.

[69] Seco, N., Veale, T., and Hayes, J. 2004. An intrinsic information content metric for semantic similarity in WordNet. In *Proc. of the 16th ECAI*, Valencia, Spain, pp. 1089-1090.

[70] Shannon, C.E. 1948. A mathematical theory of communication. *Bell System Technical Journal, 27*, pp. 379-423.

[71] Spasic, I., Ananiadou, S., Mcnaught, J., and Kumar, A. 2005. Text mining and ontologies in biomedicine: making sense of raw text. *Briefings in Bioinformatics 6(3),* pp. 239-251.

[72] *SUMO*. http://ontology.teknowledge.com/

[73] Swanson, D.R., and Smalheiser, N.R. 1994. Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease. *Neuroscience Research Communications 15(1)*, pp. 1-9.

[74] Swanson, D.R., and Smalheiser, N.R. 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence 91*, pp. 183-203.

[75] Turney, P.D., and Littman, M.L. 2003. Measuring praise and criticism: inference of semantic orientation from association. *ACM TOIS 21(4)*, pp. 315-346.

[76] van Rijsbergen, C.J. 1979. *Information Retrieval.* 2nd edition, Butterworths, London.

[77] Varelas, G., VoutsakiS, E., Raftopoulou, P., Petrakis, E., and Milios, E.E. 2005. Semantic similarity methods in WordNet and their application to information retrieval on the web. In *Proc. of the 7th WIDM*, Bremen, Germany, pp. 10-16.

[78] Witten, I.H., Bray, Z., Mahoui, M., and Teahan, B. 1999. Text mining: a new frontier for lossless compression. In *Proc. of DCC*, Snowbird, Utah, pp. 198-207.

[79] *WordNet*. http://wordnet.princeton.edu/

[80] Yang, Y., and Liu, X. 1999. A re-examination of text categorization methods. In *Proc. of SIGIR,* Berkeley, CA, pp. 42–49.

[81] Yang, Y., and Pedersen, J. 1997. A comparative study on feature selection in text categorization. In *Proc. of the 14th ICML,* Nashville, TN, pp. 412–420.

[82] Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. of the 33rd ACL,* Cambridge, MA, pp. 189-196.

[83] Yeh, A.S., Hirschman, L., and Morgan, A.A. 2003. Evaluation of text data mining for database curation: lessons learned from the KDD challenge cup. *Bioinformatics 19 (Suppl. 1),* pp. i331-i339.

[84] Zaïane, O.R. 1998. From resource discovery to knowledge discovery on the internet. *Technical Report TR 1998-13,* Simon Fraser University, August, 1998.