12-1-2007

Identifying Synonymous Terms in Preparation for Technology Mining

Cherie C. Trumbach
University of New Orleans, ctrumbac@uno.edu

Identifying Synonymous Concepts in Preparation for Technology Mining

Cherie Courseault Trumbach; Dinah Payne

Department of Management, University of New Orleans, New Orleans, LA 70148, USA

Abstract

In this research, the development of a *Concept-Clumping Algorithm* designed to improve the clustering of technical concepts is demonstrated. The algorithm developed first identifies a list of technically relevant noun phrases from a cleaned extracted list and then applies a rule-based algorithm for identifying synonymous terms based on shared words in each term. An assessment of the algorithm found that the algorithm has an 89% - 91% precision rate, was successful in moving technically important terms higher in the term frequency list, and improved the technical specificity of term clusters.

Keywords: Text Mining; Data Quality; Knowledge Discovery; Term Similarity; Text Cleaning

1. Introduction

Tech mining is the application of text mining tools to science and technology information, with a reliance on science and technology domain knowledge to inform its practice. Some of its uses include monitoring technologies, competitive technical intelligence, and developing technology policy. Tech mining is done by exploiting science and technology databases such as EI Compendex, Inspec, or Medline using a variety of analysis methods. Methods range from simple bibliometrics, or counting of bibliographic content, to text data mining using machine learning techniques. Bibliometrics has been used to develop indicators of innovation activities; it relies heavily on the structured fields in these databases. However, analyzing the free text found in the abstract field or in full documents would provide added power to analysts. While there are many methods for analyzing free text, these methods are often not well suited to the purposes of tech miners in analyzing technical concepts, particularly in the cleaning stage of text data mining [1].

There are approximately five major technique categories in the overall text data mining process: Document Retrieval, Processing, Cleaning, Mining, and Visualization. As part of the mining process, there are a number of technique categories that are subcategories of, or supplements to, these major categories, such as Clustering, Visualization or Summarization. This research focuses on the Cleaning process, arguably the most important step in the TDM process. In the text data mining process, significant Cleaning of extracted free text is typically required in order to accurately portray the prevalence of concepts in the corpus. Cleaning removes as much irrelevant material as possible and combines words that represent the same concept. This research particularly focuses on improving the conceptual representation of technical corpuses retrieved from databases of publication abstracts.

Text Data Mining applied to technical documents gives rise to issues that differ from general news corpus applications. Terms that are "uncommon," and therefore, interesting, in a news corpus may be considered "common," and therefore uninteresting, in technical applications. For example, words related to research studies, such as *study, research, results,* or *experiment,* are not "common" in news stories. However, almost all records in a technical publication database represent these concepts in some form. This paper demonstrates a *Concept-Clumping Algorithm* as an addition to, not replacement of, existing methods in the Text Data Mining Cleaning process. The algorithm first identifies a list of technically relevant noun phrases from an extracted list and then applies a rule-based algorithm for identifying synonymous terms based on shared words in each term extracted.

This research utilizes VantagePoint, a commercial text data mining tool designed to analyze text gathered from large technology publication databases. VantagePoint scans the records, identifies trends, profiles, maps, and decomposes technologies, meeting the technical intelligence needs of decision-makers. The text records, which serve as the focus of this demonstration, were taken from the Cleaned abstract phrases from samples of five technology record sets (remote sensing, fuel cells, geographic information systems, pollution monitoring, and magnetic storage) obtained from three separate databases, including Compendex, INSPEC, and Pollution Abstracts. Each sample consisted of between 176-263 records taken from one year out of the entire record set. These records are used to provide a demonstration of the benefits of a *Concept-Clumping Algorithm* designed to ultimately improve the conduct of free text analysis in technical databases in comparison to only using a Cleaning algorithm. While this project uses a list produced by VantagePoint, the algorithm itself is independent of any particular software package and can be used on any technical list. An assessment of the algorithm found that the algorithm has an 89% - 91% precision rate, was successful in moving technically important terms higher in the term frequency list, and improved the technical specificity of term clusters.

2. Background on Text Data Cleaning

In the text data mining process, the development of an appropriate list of terms[1] from which to conduct analysis requires significant effort. Processing (term extraction) and Cleaning are the two primary processes involved in the list development. Processing entails parsing terms from the text and using a Parts-of-Speech tagger to distinguish nouns, verbs, etc. In mining technical concepts, nouns are of primary interest because it is nouns that capture domain specific concepts [2]. The first step in Processing is the defining of a word/phrase. For

---

[1] Note that for this research, a "word" is a string set apart by spaces, a "phrase" is one or more words, and a "term" is a phrase that is identified as a unique phrase from the abstract of a scientific/technical journal article. A "phrase" consists of one or more *words* and every *phrase* belongs to a set of *phrases* that is a subset of *words* in a *term*. Each line in a VantagePoint abstract phrases list is considered a "term." For example, a term might be "general engineering science." It consists of three words: general, engineering, and science. There are six phrases. First, each of the single words just mentioned are considered single-word phrases. The two-word phrases are "general engineering" and "engineering science." Finally, "general engineering science" is a three-word phrase.

instance, terms can be determined by every space between each word, in which case all terms would be single words. Terms can also be determined by Natural Language Processing algorithms, including NP-Chunking, to identify actual phrases (i.e. "Information Retrieval") [3; 4]. Another approach is simply to use windows of adjacent words. Parts-of-Speech taggers then distinguish nouns, verbs, etc. Some extraction techniques are capable of identifying specific entity types, such as whether a noun is a person, organization, phone number, date, address, or geographical location [5; 6]. Since the analysis of technical records only requires capturing domain specific concepts, the exact entity type is not important [7; 5]. After an initial list of extracted terms is developed, Cleaning is required to permit effective analysis of the record set. Cleaning impacts the quality of other text mining techniques and determines the quality of the information fed into the actual mining algorithms.

The two main issues in cleaning text are related to the *selection* and *compression* of the terms. Selection is the way terms from text are determined to be candidate keywords for analysis. It involves narrowing the number of terms for analysis once they have been identified. Selection issues relate to identifying a term as a potential keyword for analysis and determining the significance of that word in the document. Many tools simply remove a small set of common words such as "the" and "of" or only use terms that meet a minimum frequency for clustering. One method breaks terms into sequences, and only use maximal frequent sequences, which are sequences of words that are frequent in the document collection and are not contained in any other longer frequent sequence. A frequency threshold is defined for the document set [8]. Kostoff and Block propose a method that uses factor analysis to determine which terms are high loading on the factors. These terms tend to have high technical content. The other terms are discarded as trivial [9]. Wilbur and Young present another such method. They offer a term strength concept based on "how strongly the term's occurrences correlate with the subjects of the documents in the database." Term strength is then fed into an algorithm for determining stop words, or terms to exclude. [10] Feldman et al offers three different approaches to statistically select terms [11].

In this research, a method based on the Zipf distribution was utilized. The Zipf distribution takes as a premise the idea that the log of the rank versus the log of the frequency of a term is linear. The method used finds that line and the terms with the highest and lowest rank that fall below the line are eliminated [12]. In order to bolster the frequency or strength of terms in abstracts or full text documents, compression is used. Compression is grouping together synonymous terms. Stemming is the most basic type of compression. Porter introduced stemming with a rule-based algorithm for combining words that share a common stem such as "computer" and "computers [13]." Recent improvements on the basic stemming algorithm include the creation of stemming algorithms in other languages such as Arabic or Spanish, improving the performance of the stemming algorithm, and utilizing stemming in Retrieval functions [14;15 16]. Another method proposed by Wilbur and Kim uses the tri-grams found in the words that form a phrase with similarity measures typically used for documents in order to determine the level of similarity between phrases. While this method only compares two words and does attempt to group multiple words together and is typically has been used for spell-checking endeavours, it has potential for other text mining compression applications [17].

VantagePoint's List Cleanup function uses a stemming algorithm and shared words in reverse order to improve the compression. In this case, words such as "technology manager," "managing technology" and "technology management" are combined. However, terms such as "engineering science and "general engineering science" or "internet commerce" and "web commerce" would still not be identified as a single concept. The compression of synonymous terms based on context is a more sophisticated level of compression. Ahonen-

Myka use the concept of equivalence class, defined as sets of phrases that occur together in the same documents frequently enough, to combine synonymous concepts [8]. Phrases belonging to some equivalence class are replaced by the name of the class. However, this approach may combine as one, words that are not actually synonymous, but are simply related concepts. The problem is that in using these false synonyms to identify conceptual relationships, in future text mining steps, second-order relationships will be identified as first-order. It is essentially clustering twice. Another approach, which is fairly manual, identifies synonymous terms using natural language dictionaries [18]. In all of these approaches, terms are compressed across multiple documents. Many text mining software products currently on the market, however, limit the Cleaning of nouns to a task within a document as a component of entity extraction. Some packages link a last name listed in a document with a full name in the same document. The same is true for company acronyms and company full names. However, if the acronym or last name is in a different document, then the association is missed. On the other hand, methods that actually attempt to identify synonymous terms often require some type of coding for domain knowledge [19]. However, if a purpose in analyzing technology abstracts is to identify unknown relationships or emerging technologies, then, an unsupervised statistical approach to Cleaning that does not require training is necessary.

On the flip side of identifying synonymous terms, is word sense disambiguation (WSD). WSD typically involves distinguishing the correct sense of polysems. The algorithm presented in this paper uses ideas from word sense disambiguation, particularly from the *topical context* area. This area relies on the "repeated use of words which are semantically related throughout a text" and large window sizes , are shown to successfully disambiguate noun phrases.[20; 21; 22] Though we are not distinguishing individual occurrences of polysems, we are similarly forcing terms to choose between one "sense" and another, based on the term's context in an technical abstract. Terms must be determined to be more of a synonym to one set of terms or another. The problem with WSD approaches for technology analysis is that even unsupervised methods, such as Naïve Bayes and Exemplar approaches, require training. The three main lines of WSD research focus on efficiency in sampling, use of lexicons such as Wordnet, and using the Internet to collect word sense samples [23]. However, similarity measures, typically utilized to determine similarity between documents which do not require training seem better suited to analyzing fast-changing, technically specific sources. For the same reasons, lexicon-based approaches are not ideal either.

In such research, more accurate concept representations, combining as many actual synonyms as possible, can mean more accurate end-results. The discussion that follows highlights the need for a *concept-clumping algorithm* when working with the free text found in technology abstract.

Table 1. List of Keywords and Abstract Phrases

List of Keywords	List of Abstract Phrases
• Pollution control • Sonochemistry • Mass Transfer • Ultrasonic applications • Reaction Kinetics • Sonochemical Reacting Systems	• Environmental Sonochemistry • Environmental remediation • Ultrasonic waves • Kinetic analysis • Sonochemical engineering • Chemical analysis

	• Mass transfer • Aqueous solutions • Chemical processing • Cheaper reagents • **Novel means** • Shorter reaction cycles • Smaller plants • Large-scale applications • Growing area • Existing knowledge • Outline directions • Exciting field

Source: "Sonochemistry: Environmental Science and Engineering Applications." It demonstrates the difference in terms listed in the keywords list versus those listed in the abstract phrases.

3. The Need for Concept-Clumping

In attempting to identify the underlying structure of a technology, technology analysts have frequently used keywords over abstract phrases, due to the many challenges inherent in free-text. While keywords are technologically sound, they are more general, may be chosen by the database administrator, or limited to choices provided by a particular journal. Emerging ideas may be masked under a broader category until there is sufficient publishing to warrant creating a new topic category. On the other hand, problems with free text are numerous. A primary problem is the variation in the words that are used. The level of specificity may result in a large number of missed relationships. Another problem is that a document may contain words that provide no conceptual insight into the content of the document, such as "novel means" shown in the sonochemistry abstract record example in Table 1. Table 1 provides a comparison of the keywords and phrases extracted from an abstract in an example technical record. Additionally, as mentioned previously, there are occasions where the same concepts may be discussed in a variety of ways, even within the same abstract. Therefore, in order to effectively analyze the information, the data should be Cleaned and Clumped to accurately portray the prevalence of the concepts in the dataset. As mentioned above, the idea is to remove as much irrelevant material as possible and to combine terms that are synonymous. The data clumping algorithm developed for this research first identifies a list of relevant noun phrases and then applies a rule-based algorithm for identifying synonymous terms based on shared words. The value in this approach is that it attempts to compress terms that are true synonyms, and not just closely related concepts. The algorithm does not claim to be generalizable to all text sets, but is intended for use with technical periodical abstracts. Further research will be necessary to determine the generalizability of results to other types of text document sets.

4. Description of the Concept-Clumping Algorithm

In performing additional term clumping, the intention is to increase the analytical validity of using abstract phrases to perform analysis. The basic outline of the algorithm is as follows:

Remove hyphens, numbers, and punctuation

1. Remove common words
2. Clump phrases with four or more words in common into a new phrase
3. Name the new phrase the shortest phrase name
4. Calculate the prevalence of the remaining words
5. Clump phrases with three words in common into a new phrase
6. When a conflict arises, use a similarity measure to determine with which group of phrases that the conflicted phrase will clump
7. Name the new phrase the phrase name with the highest prevalence score
8. Repeat steps 5) – 7) for two word matches
9. An in-depth description of the above steps follows below.

The basic starting point for the algorithm is a Cleaned list of simple abstract noun phrases as determined by the Natural Language Processing and fuzzy-matching algorithms contained in the VantagePoint software package. The NLP algorithm in VantagePoint separates noun phrases connected by conjunctions. Non-alphanumeric characters are then removed, combining terms such as "high-density" and "high density." Then, the algorithm removes non-technical, common single words from the list published by White [24]. Finally, with only multiword noun phrases and uncommon single word nouns remaining, the list is ready for clumping.

The basis of the remaining portion of the algorithm is the existence of shared words. Shared words are the words that exist together in more than one term. For example, "engineering science" and "general engineering science" share two words: "engineering" and "science." Identifying equivalent concepts is a difficult process; by starting with shared words, a high level of precision can be achieved and the number terms compared to one another is limited, thereby reducing the processing time to a reasonable level.

The algorithm first searches for terms with four words in common. If terms have four words in common, these terms are combined together and named for the shortest term. In the rare occasion that a conflict arises, the algorithm chooses the first grouping that occurs in the thesaurus. This approach appears somewhat random; however, initial analysis revealed that these terms are likely *all* conceptually the same and would be grouped together in the three-shared words step in the algorithm anyway.

Secondly, terms sharing three words in common are each given a prevalence rating. The formula for the prevalence rating is:

$$P(b) = \sum_{\substack{\forall \text{Docs where} \\ (b) \in D(i)}} \frac{\text{Instances of (b) in D(i)}}{\text{\# of relevant terms in Doc (i)}}$$

where:

P(b) = prevalence rating for term (b)

(b) = a term in the abstract phrase list

D(i) = the set of terms contained in Document (i) in the record set

This method is used because it gives a higher rating both to terms that appear in many documents and to terms that appear more frequently in one document. Words are also given a higher prevalence if they appear in shorter abstracts.

Once the prevalence rating is determined, the algorithm searches for groups of terms that share a three-word phrase. These terms are clumped into one term. If a term shares phrases with multiple groups, a similarity measure will determine the group to which the term belongs. The basis of the similarity measure is a standard approach to similarity used in Information Retrieval where similarity of terms has been researched most frequently. The premise is that two terms are semantically similar if they occur in the same context [25]. Typically, similarity is used to determine the similarity between documents. Similarity may be used to cluster similar documents, expand queries, identify duplicate documents, or identify plagiarized documents [26; 27; 28; 29]. In this case, the similarity relationship of interest is among terms and not documents. Other approaches to similarity are taxonomy-based. The similarity between two items depends on the relationship or distance of the terms in a hierarchically structured lexical resource, such as WordNet [30]. Taxonomy-based approaches would require incorporating a lexical resource such as WordNet into Vantage Point. A problem with such an approach, for the purposes of this research, is that the terms that are most likely represented differently in the record sets occur in newer technical areas. These areas would less likely appear in a lexical resource. Therefore, a contextual similarity approach is more suitable for technical publications. The similarity measure used, from Cutting et al [31], asserts that a term is most similar to the term group that co-occurs with terms most similar to the original term's co-occurring terms. This measure is calculated from the term-document matrix.

Therefore, for each document α in a corpus C, let $c(\alpha)$ be each word in the document and its frequency. Let V be the set of unique terms occurring in C. Then $c(\alpha)$ can be represented a vector of length $|V|$;

$$c(\alpha) = \{f(w_i, \alpha)\}_{i=1}^{|V|}$$

w_i = ith word in V

$f(w_i,\alpha)$ = the frequency of w_i in α.

Using the cosine between monotone element-wise functions of $c(\alpha)$ and $c(\beta)$, the similarity measure between two documents can be determined by

$$s(\alpha,\beta) = \frac{(g(c(\alpha)), g(c(\beta)))}{\| g(c(\alpha)) \| \| g(c(\beta)) \|}$$

where g is a monotone damping function using a component-wise square-root, "(,)" denotes inner product, and "|| ||" denotes vector norm. The aforementioned equation can be applied to determine the similarity between the group of documents in which the group of terms that share a phrase appear (Γ) and the documents in which the term that shares phrases with multiple groups appears (x) [31].

Once all of the three common phrase matches have been made, the term chosen to represent the group is the term with the highest prevalence rating. The two-shared-words clumping process then begins. The same process utilized in matching terms that share three common words is utilized to match terms that share two common words. Note that this research stops at two shared-words in common. Future research may look at improving the algorithm to effectively handle terms that only share one word in common. The assumption is that as the number of shared words decreases, the less likely it is that the shared words indicate a similarity and, therefore, different approaches may be necessary.

"Precision" tests the ability of the algorithm to accurately identify that two words are synonymous. The overall precision was evaluated by running the algorithm against an abstract record corpus. Each term was manually compared to the term that the algorithm named the group for determination as to whether it is actually similar in concept. The naming algorithm is important because it ultimately determines the term that is chosen to represent all of the terms in the group.

4.1. Revision to the Algorithm

After initial testing, one important adjustment was made to the algorithm. In some cases, because the algorithm forces the term to choose between groupings starting at the level of the greatest number of shared words, the multiword search terms create some inaccurate groupings, if that term appears in numerous separate concepts. The reason is that the different variations in spelling of the search term would be considered at the same time as different categories of the search term. "Carbonate fuel cell systems" has as many shared words with "solid oxide fuel cell" as it does with "carbonate fuel cells." The algorithm ran at sufficient accuracy for the "geographic information system" and the "pollution monitoring" record sets. However, the problem became evident after running the algorithm on the "remote sensing" and "fuel cell" record sets. At the two-shared-words iteration, "carbonate fuel cell system" would have to choose between "solid oxide fuel cell" and "carbonate fuel cell." Since the terms cell(s) very rarely appear without fuel, ignoring "cell(s)" improves the accuracy of the algorithm. "Carbonate fuel cell system" would not have to consider "solid oxide fuel cell" as a partner. In the remaining record sets, the noun part of the search term which may appear in a variety of forms was ignored, meaning "sensing," "sensor," "cell," and "cells," by the algorithm. Ignoring the search term word that rarely appears without the other is a way of forcing additional strength between concepts that contain the search term. It requires an additional shared word, allowing different categories of the search term to be considered before variations in spelling of the search term itself.

As revised, the algorithm macro now gives the user the option of ignoring a string or set of strings from consideration. In the future, something like "sub" might be ignored. "Sub" is used in abstracts to indicate a subscript. So, in scientific abstracts "O_2" would be written as O(sub)2. Further research will be required to determine what terms should be added to a list of terms to ignore. If there are terms that should be ignored across all record sets, the algorithm should be programmed to read these words from a stopwords list. The goal is to create a list that is not domain specific.

5. Algorithm Results and Impact

In this demonstration, the completed Algorithm was programmed into VantagePoint and was run on the Cleaned Abstract Phrases from samples of the five record sets from the selected topic areas. For demonstration purposes, each sample consists of between 176-263 records taken from one year out of the entire record set.

The output produced is a set of VantagePoint thesaurus files, which combined together provide the entire clumped group and the term that is ultimately chosen as the representative term for the group of terms deemed similar. For example, the output file contained the following segment:

**hard disk drives

hard disk drives

double prime hard disk drives

hard drives

The "**" indicates the name that the terms in the lines below it will be given.

5.1. Precision Results

Each term was evaluated to determine if the representative term provides an accurate portrayal of the term under consideration. The file was opened as an Excel Spreadsheet and each term in the group was evaluated to determine if "hard disk drives" is a conceptually accurate representation of the term. For this segment, all of the terms are "Good Matches." Therefore, the spreadsheet was marked as in Table 2.

Table 2. Hard Disk Drive Matches

Bad Matches	Good Matches	**hard disk drives
	1	hard disk drives
	1	double prime hard disk drives
	1	hard drives

The column totals were tabulated in order to determine the precision of the algorithm in that record set. Only output combining terms are considered. So, consider the following output in Table 3.

Table 3. High Density Recording Matches

Bad Matches	Good Matches	Terms
		**high density television
1		high density
1		high bit density

1		high density partial response channels
	1	high density television
1		high superficial density
		**magnetic property
		magnetic property
		**thin film head elements
	1	thin film
	1	polished thin film disk
	1	thin film head on disk wear tests
	1	thin film rigid disk
	1	thin film disks
	1	isotropic longitudinal CoCrTa Cr thin film head
	1	thin film head elements
	1	Co Pt thin film patterns
	1	conventional thin film head sliders
	1	thin film corrosion
	1	thin film corrosion model
	1	thin film discs
	1	thin film magnetism
	1	thin film optics
	1	thin film type recording head
		**magnetic heads
	1	magnetic heads
	1	small magnetic heads
		**thin films heads
	1	thin film inductive heads
	1	conventional thin film inductive heads
	1	inductive thin film magnetic recording heads

	1	thin film inductive recording heads
	1	thin film magnetic recording heads
	1	thin film recording heads
	1	CoTaZr amorphous thin film disk heads
	1	thin film inductive disk drive heads
	1	thin film magnetic heads
	1	thin film read write magnetic heads
	1	conventional thin film heads
	1	modified thin film heads
	1	similar thin film heads
	1	thin film heads TFHs
	1	thin films heads

The "B" column is a marker for "Bad Matches" and the "G" column is a marker for "Good Matches." Notice that the group member "amorphous magnetic film" does not have a "1" in either column. This term is the only term in its group and, therefore, was not included in the calculation. There are 33 terms that are considered Good Matches and 4 that are considered "Bad Matches." In some cases, judgments were made by reviewing individual abstracts to determine the context of the term in the record set.

Where precision = (Good Matches)/ (Good Matches + Bad Matches), the above sample had a precision of 33/37 or 89.2%. Moreover, the precision of the algorithm on the samples were above 89% for all five record sets (Table 4).

Table 4. Technology Cases: Clumping Algorithm Precision Calculations

File	**# Records**	**Precision**
Fuel Cells (1995)	197	91.1%
Remote Sensing (2002)	263	89.7%
Magnetic Storage (1992)	220	91.7%
GIS (1992)	176	90.7%
Pollution Monitoring (2003)	181	91.4%

5.2. The Effect of Clumping on Frequency Lists

Technology mining can be broken down into four levels: lists, matrices, maps, and trends. The foundation is the list. Experts and institutional players as well as indicators of technology activity are identified first by the lists and the additional analysis based on the lists. The analyses seek to answer questions such as

What research is taking place in the technology domain?

Who is conducting that research? What is their expertise?

How is the research focus changing over time?

Hence, the importance of starting with a list that accurately portrays the research domain.

The effect of the algorithm is apparent in the "Top 20" term list for each of the example record sets. The Clumped Abstract Phrases list is shown alongside the Cleaned Abstract Phrases list and the Cleaned Abstract Phrases list with the common words removed. Individual points of interest are discussed below each Top 20 list (Tables 5 to 9).

Consider the lists in Table 5. The Cleaned Abstract Phrases list only contains two multiword phrases containing "fuel cells" (the search term itself) and "solid oxide fuel cells." However, clumping allows for many of the multiword concepts to increase in prominence on the list. In comparison to the original list, four additional terms containing the phrase "fuel cells" are now on the list and an additional two terms in comparison to the list without stop words. Additionally, the concept "solid oxide fuel cells" increases from 11 records to 30 records. The combined "solid oxide fuel cells" entry consists of the following original terms:

solid oxide fuel cells

solid oxide fuel cells SOFCs

reduced temperature solid oxide fuel cells SOFCs

novel solid oxide fuel cell SOFC system

SOFC Solid Oxide Fuel Cells interconnector material

solid oxide fuel cell SOFC cells

solid oxide fuel cell SOFC performance

chemical cogenerative solid oxide fuel cell

solid oxide fuel cell electrolytes

solid oxide fuel cell systems

Table 5. Fuel Cell Top 20 Abstract Phrases

	# Recs	Abstract Phrases Cleaned	# Recs	Abstract Phrases Cleaned (stop word removed)	# Recs	Abstract Phrases Clumped
1	50	fuels cells	50	Fuels cells	50	fuels cells

2	33	Cs	33	Cs	33	Cs
3	24	developments	24	Developments	31	deg
4	24	results	14	Temperatures	30	solid oxide fuel cells SOFCs
5	20	effects	12	Electrodes	24	developments
6	14	study	12	Electrolytic	15	direct methanol polymer electrolyte membrane fuel cells
7	14	temperatures	12	Hydrogenation	15	molten carbonate fuel cells
8	14	uses	12	Increasing	14	temperatures
9	13	operator	11	Applications	12	current density
10	12	cells	11	solid-oxide fuel cells	12	electrodes
11	12	electrodes	9	cathodically	12	electrolytic
12	12	electrolytic	9	solid-oxide fuel cells SOFCs	12	hydrogenation
13	12	hydrogenation	8	COS	12	increasing
14	12	increasing	8	potentials	12	oxygen
15	12	oxygen	7	thicknesses	12	yttria stabilized zirconia YSZ
16	12	systems	6	characteristics	11	applications
17	11	applications	6	conductivity	10	high efficiency
18	11	solid-oxide fuel cells	6	electrical power	9	cathodically
19	10	activity	6	molten-carbonate fuel cells	9	phosphoric acid fuel cells
20	10	catalysts	6	pressurization	9	proton exchange membrane fuel cells

The simple ability to combine “solid oxide fuel cells” and “solid oxide fuel cells SOFCs” would increase the representation of the this type of fuel cell from 11 records to 18 records. Some other important terms not on the list originally were: direct methanol polymer electrolyte membrane fuel cells, molten carbonate fuel cells, phosphoric acid fuel cells, yttria stabilized zirconia YSZ, and proton exchange membrane fuel cells.

Using the concept-clumping algorithm, “yttria stabilized zirconia YSZ” is counted in 12 records. Without the algorithm, the most frequent variation of this term only appears in 2 records. Therefore, without the algorithm it would not be used in the mapping function at all. Phosphoric acid fuel cells is another term that makes the Top 20 list only after clumping. It consists of the following terms.

four phosphoric acid fuel cell monocells

kilowatt phosphoric acid fuel cell

phosphoric acid fuel cell cathodes

phosphoric acid fuel cell technology

phosphoric acid fuel cells

pressurized phosphoric acid fuel cell

phosphoric acid electrolyte

platinum bearing phosphoric acid

pyro phosphoric acid

Two phosphoric acid fuel cell terms that are not included in this grouping are "phosphoric acid fuel cell power plants" and "PAFC power plants," which the algorithm determined were more similar to a fuel cell power plants grouping.

After numerical and punctuation characters are removed from the list, common words with up to ten letters are removed. Notice the impact that this has on the Abstract Phrase list for Remote Sensing (Table 6). The five most frequent terms (results, data, study, methods, used) are removed from the list. Terms are removed that would be included in a wide array of records but do not uniquely distinguish the scientific concepts in the record.

Table 6. Remote Sensing Top 20 Abstract Phrases

	# Recs	**Abstract Phrases Cleaned**	**# Recs**	**Abstract Phrases Cleaned (stop words removed)**	**# Recs**	**Abstract Phrases Clumped**
1	72	Results	26	Applications	79	remote sensing
2	40	Data	25	Remote sensing	26	applications
3	35	Study	24	Estimators	24	estimators
4	34	Methods	22	Development	22	development
5	32	Used	19	Approaches	19	approaches
6	26	applications	14	Techniques	15	Synthetic Aperture Radar SAR images
7	26	Presented	12	Atmosphere	14	experimental results
8	25	remote sensing	12	Experimental results	14	techniques

9	24	Effects	12	Information	12	Atmosphere
10	24	Estimators	12	Potentiality	12	information
11	22	Accuracy	11	Relationships	12	potentiality
12	22	Analysis	10	Classifications	11	land cover classification
13	22	development	10	Combinations	11	ms
14	21	Surfacing	10	Vegetation	11	relationships
15	21	Systems	8	Correlators	10	classifications
16	20	Measures	8	Distribution	10	combinations
17	19	Approaches	8	Remote sensing applications	10	km
18	18	Problems	8	Sensitivity	10	vegetation
19	17	Images	8	Study cases	9	conditions
20	16	Regions	8	utilization	9	Gaussian maximum likelihood GML classification

Notice the Magnetic Storage Cleaned Abstract Phrases contain a number of generic single terms (Table 6). In the Clumped Abstract Phrases list, there are a few "thin film" entries, such as "thin film heads," that were not in either "Top 20 Cleaned Abstract Phrases" list. The output file looks as follows:

**thin films heads

thin film inductive heads

conventional thin film inductive heads

inductive thin film magnetic recording heads

thin film inductive recording heads

thin film magnetic recording heads

thin film recording heads

CoTaZr amorphous thin film disk heads

thin film inductive disk drive heads

thin film magnetic heads

thin film read write magnetic heads

conventional thin film heads

modified thin film heads

similar thin film heads

thin film heads TFHs

thin films heads

Table 7. Magnetic Storage Top 20 Abstract Phrases

	# Recs	Abstract Phrases Cleaned	# Recs	Abstract Phrases Cleaned (common words removed)	# Recs	Abstract Phrases Clumped
1	34	Results	20	Ms	32	Mu
2	29	Heads	16	Development	20	High density recording
3	28	Uses	15	Techniques	20	Ms
4	27	Effects	14	Magnetic property	20	Thin film recording media
5	21	Presents	11	Applications	17	Thin film heads
6	20	Ms	10	Experimental results	16	Developments
7	19	Disks	9	Directions	16	Thin film magnetic recording disks
8	18	Measures	9	Distributions	15	Techniques
9	17	Methods	9	Increasing	15	Thin film head elements
10	16	Described	9	Recording heads	14	Magnetic property
11	16	Developments	7	Improvements	12	Deg
12	15	Techniques	7	Influences	11	Applications
13	14	Magnetic property	7	Magnetic heads	11	Experimental results
14	13	Functions	7	Thicknesses	11	MIG heads
15	13	Systems	6	Air-bearing surfaces	11	Recording heads
16	12	Magnets	6	Calculations	10	Finite element method FIM
17	12	Taping	6	Hard-disk drives	10	Intermittent head disk contacts
18	11	Applications	6	High-density recording	9	Air bearing surfaces
19	11	C	6	Mechanisms	9	Directions
20	11	Problems	6	Reductions	9	Disk drives

Table 8. GIS Top 20 Abstract Phrases

	# Recs	Abstract Phrases Cleaned	# Recs	Abstract Phrases Cleaned (common words removed)	# Recs	Abstract Phrases Clumped
1	54	GIS-Geographic Information System	54	GIS-Geographic Information System	83	GIS Geographic Information System
2	43	GIS	43	GIS	63	geographical information systems
3	36	Data	32	geographical information systems	43	GIS
4	32	geographical information systems	24	Applications	24	applications
5	32	Results	24	Developments	24	developments
6	31	Systems	17	Management	21	spatial data
7	30	Uses	15	Spatial data	13	U S
8	24	Applications	12	Researches	12	multiple remote sensing images
9	24	Developments	11	Relationships	12	researches
10	20	Analysis	10	Processing	11	land use category
11	20	Informing	7	Approaches	11	relationships
12	20	Study	7	Potentials	10	ground water
13	18	Maps	7	Wide variety	10	processing
14	16	Timing	6	Attribution	10	remotely sensed
15	15	spatial data	6	Collective	9	data sets
16	14	Areas	6	Environments	9	land uses
17	14	Numbers	6	Users interface	8	United States
18	14	Plans	5	Characteristics	8	water resources
19	14	Tools	5	Classifications	7	approaches
20	14	Users	5	Data sets	7	Extensive water quality data

The GIS list reveals the limitation of the clumping algorithm. The first three terms on the list are "GIS Geographic Information System," "Geographical Information Systems" and "GIS." These terms are clearly the same concept, but share at most only one word in common. The algorithm only reviews terms that share at least two words in common. This GIS case reveals a drawback to the two-shared word limit. However, if only one-shared word were necessary every term containing the word "information" would have to be compared against each other. Reapplying the concepts of ignoring common words, stemming, and similarity could result in a more powerful algorithm that could address these issues.

Table 9. Pollution Monitoring Top 20 Abstract Phrases

	# Recs	Abstract Phrases Cleaned	# Recs	Abstract Phrases Cleaned (common words removed)	# Recs	Abstract Phrases Clumped
1	61	Results	42	Concentrations	42	concentrations
2	51	Study	21	Zn	21	Zn
3	42	Concentrations	19	Contamination	19	contamination
4	36	Data	19	Pb	19	Pb
5	29	Sites	17	Cu	17	Cu
6	21	Zn	16	Cd	16	Cd
7	20	Effects	14	Sub(2	13	heavy metals
8	19	Contamination	13	Heavy metals	13	pollutants
9	19	Pb	13	Pollutants	12	air pollution
10	18	Soils	12	CO	12	air quality
11	18	Used	11	Contributions	12	Co
12	17	Cu	11	Distributions	11	contributions
13	16	Cd	11	Ni	11	distributions
14	15	Impacts	10	Study area	11	environmental heavy metal ions
15	15	Low	9	Determined	11	Ni
16	15	Sampling	9	Indicators	10	PM sub
17	14	Analysis	8	Air	10	study area
18	14	Area	8	Correlations	9	high concentrations
19	14	Increases	8	Depositions	9	indicators
20	14	Sediments	8	Fe	9	polycyclic aromatic hydrocarbons

In the case of Pollution Monitoring, some terms rose in prominence on the list, while terms such as "heavy metals," " environmental heavy metal ions," and "polycyclic aromatic hydrocarbons" were included on the list. The group for "polycyclic aromatic hydrocarbons" consists of the following terms:

**polycyclic aromatic hydrocarbons

polycyclic aromatic hydrocarbons PAHs

polycyclic aromatic hydrocarbons

low molecular weight polycyclic aromatic hydrocarbons PAH

particle bound polycyclic aromatic hydrocarbons

polycyclic aromatic hydrocarbon PAH exposure

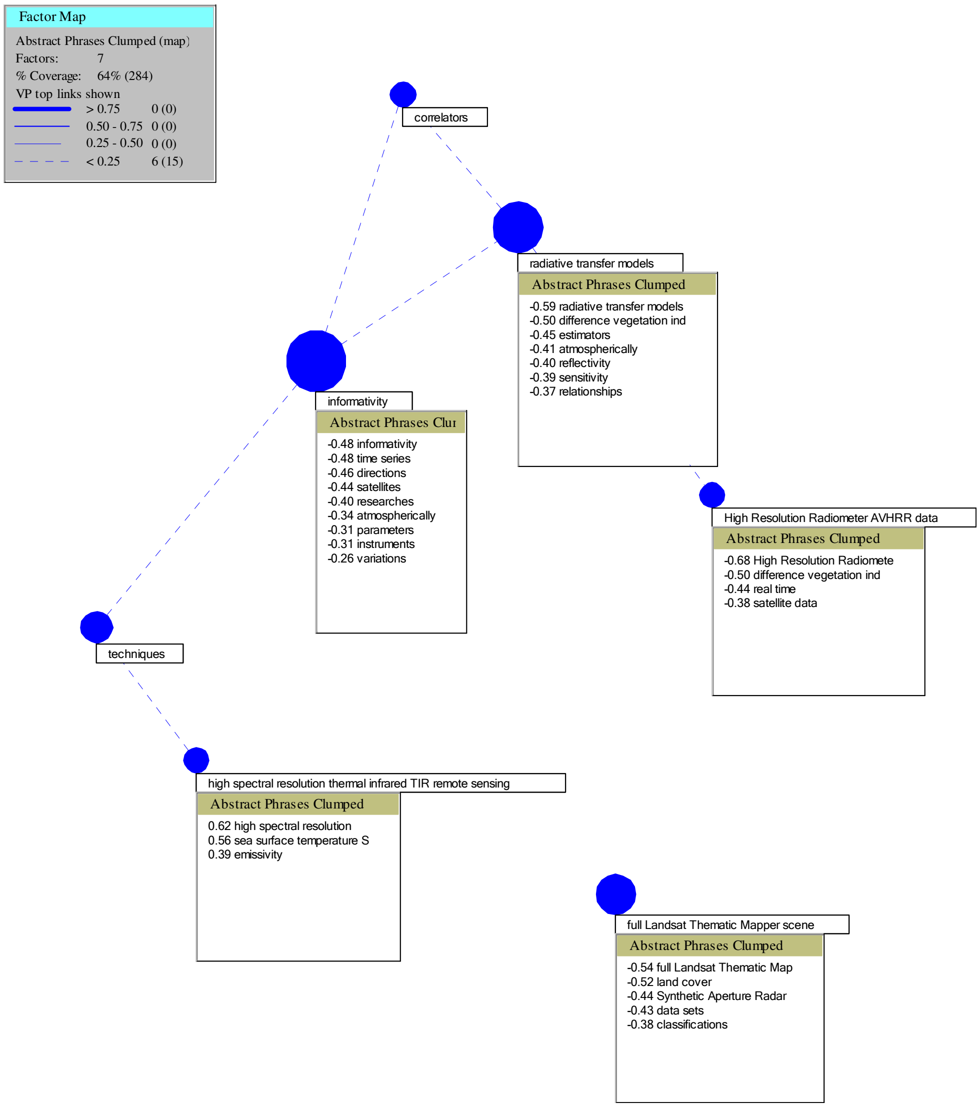
The most frequent occurrence of any one of these terms is the title term, which appears in two records. A term that clearly conceptually belongs with this group is “PAHs,” which occurs in 7 records. An improvement in the algorithm should attempt to match such a term with like concepts.

Improvements in the accuracy of the “Top Twenty List” are important by itself. The list provides valuable insight into the important topics discussed in the domain. However, lists are only the starting point for analysis. Cluster maps are used to identify related research topics that may not be identified in a simple document search. Clusters of related terms are identified as are links between clusters. A more accurate and technically focused list can greatly affect both the accuracy and richness of the clusters utilized by the technology analyst.

5.3. *Impact on Clusters*

The value in the clumping algorithm rests in creating a more accurate dataset to input into analysis methods. In this paper, the results from applying clustering to abstract phrases that have been clumped in comparison to abstract phrases that have only been cleaned have been described. The first step in creating a cluster map based on Principal Component Analysis (PCA), the clustering method used in the VantagePoint software program, is determining which terms will be included in the clustering. There are a number of ways to make this determination; however, regardless of methodology, the term must occur in at least two documents in order for any co-occurrence based method to work. Using all terms with at least two occurrences is one method and another is to take a percentage of the terms. However, as discussed in the background, there are more sophisticated approaches such as the Zipf’s distribution approach. After the terms for the cluster map were determined, maps were created for a random sample of each of the five full datasets, a Cleaned Abstract Phrases map, and a Clumped Abstract Phrases map. These sample sizes ranged from 434 – 880 records. The Remote Sensing Clumped Abstract Phrases Map shown in Figure 1 is an example of one of the maps.

Figure 1. Remote Sensing Clumped Abstract Phrases Map



While there are a number of methods to evaluate clusters such as entropy and cohesion, those methods are better suited to evaluate clustering methods applied to a crafted dataset. In this case, the same clustering method was used with altered inputs. A simple t-test in SPSS was used to compare the cleaned and clumped means for each of the metrics. The results are listed in Table 10.

Table 10. Cluster Quantitative Measure Comparison of Means

	(A)	**(B)**	**(C)**	**(D)**	**(E)**	**(F)**	**(G)**	**(H)**	**(I)**	**(J)b**	**(K)**
Clnd Mean	21	594	11909	145	1.2	9	10	4.73	45	45	63
Std. Dev.	5	190	1620	100	0.7	2	2	0.86	10	29	11
ClmpMean	15	594	8265	134	1.6	9	11	5.10	53	47	59
Std. Dev.	4	190	1019	76	1.0	3	3	0.19	18	17	10

A) terms per document
B) Number of documents
C) Total number of terms
D) Number of terms used in clustering
E) Percentage of terms considered for clustering
F) Number of links on cluster map
G) Number of clusters on cluster map
H) Average Number of terms per cluster
I) Number of terms assigned to a cluster
J) Percentage of terms assigned to a cluster
K) Percentage of documents covered by the clusters

The only significant difference was in the total number of terms, which is reduced by 30%, a figure that is an expected result from removing some terms and combining others. These numbers are not necessarily a surprise since the same clustering algorithm was used on all the datasets. However, there are a couple of notable points. First, while the total number of terms shows that clumping results in significantly less number of terms, the number of terms chosen for clustering does not, indicating that a higher percentage of clumped terms are considered *impactful*. Secondly, the precision and impact of the clumping algorithm reveal that clumping conceptually represents the dataset well. The more important evaluation of the value of clumping in clustering is revealed in the actual clusters themselves.

The biggest difference between the two types of Abstract Phrase maps is the technical specificity of the terms included. Cleaned Abstract Phrases are dominated by the common generic terms. This circumstance exists for two reasons: the most common words are not removed and the more technical terms are included in phrases that are not gathered together as in the Clumped Phrases. For example, in the Magnetic Storage record set, the "friction" cluster in Cleaned Abstract Phrases includes the terms: "friction," "surfaces," "lubrication," "coefficients," "wearing," and "tribology." A similar cluster in the Clumped Phrase map contains phrases like "head disk interface," "surface roughness," "slider disk spacing," "Contact Start Stop durability" and "stiction." Cleaned Abstract Phrases contains more clusters that have little meaning because of the broad terminology included. Clusters such as these appear in the Remote Sensing dataset:

Accounts: used, limits, accounts, interpreting, selection, important

Presents: presents, ones, techniques, atmospherically, described, viewing, experimental results, improvements

In contrast, some of the Clumped Abstract Phrases clusters are:

AVHRR data: difference vegetation index NDVI, real time, satellite data, High Resolution Radiometer AVHRR data

TIR remote sensing: high spectral resolution thermal infrared TIR remote sensing, sea surface temperature SST, emissivity

Clearly, clumping provides richer details in the clusters.

6. Summary and Conclusions

The precision and impact of the clumping algorithm reveal that clumping conceptually represents the dataset well. Identifying terms that are synonymous is important to improving accuracy when mining free text. An algorithm was developed that has delivered at least an 89% precision rate in making such identifications. While this is a high level of precision, it does result in approximately 11% missed assignments. However, the algorithm can be implemented in such a way that the user can easily remove unsatisfactory groupings. This level of precision was achieved across five different technology areas (pollution monitoring, remote sensing, magnetic storage, fuel cells, and geographic information systems) and was used in three different databases (Compendex, Inspec, and Pollution Abstracts), all with about the same level of precision. These results indicate that the algorithm may be used with other types of technical free text such as Patents and the Internet. However, further research would be necessary due to the difference in writing styles. The impact of this algorithm can be seen in Top 20 lists in Tables 5 to 9. Terms that are conceptually important to the dataset (solid oxide fuel cells) have replaced very generic common words (study, results) at the top of the term list. Also, the viability of using Abstract Phrases with additional analysis methods such as clustering improves because the concept-clumping algorithm reduces the number of terms to consider for clustering by 30%. The terms left are the more technical terms. The result is the ability to use abstract phrases in analysis, in place of the structured, yet broad, keywords which have typically been used in analyzing publication records, which allows the more detailed nature of abstracts to be captured with the mining techniques. Clumped Abstract Phrases capture the broad relationships as well. However, from the Top 20 lists, terms that have the same meaning that are still not identified as being conceptually the same are also seen. Therefore, additional work will be needed to improve the recall of the algorithm without reducing the precision. The lists also reveal additional opportunities for improvement. If VantagePoint is to be used on files with the chemical elements discussed, a thesaurus for the elements in the periodic table may be useful.

7. Acknowledgements

Thanks to Doug Porter for his assistance in translating the algorithm into VantagePoint software code and everyone at Search Technology for their assistance in implementing this algorithm.

8. References

[1] A.L. Porter and S.W. Cunningham, *Tech Mining: Exploiting New Technologies for Competitive Advantage* (Wiley-Interscience, Hoboken, 2005).

[2] K. Chen and H.-H. Chen, Extracting noun phrases from large-scale texts: a hybrid approach and its automatic evaluation. In: *Annual Meeting of the ACL 1994: Proceedings of the 32nd Annual Meeting of the ACL*, (Las Cruces, 1994).

[3] J. Li-Ping, H. Hou-Kuan, and S. Hong-Bo, Improved feature selection approach TFIDF in text mining. In: International Conference on Machine Learning and Cybernetics 2002: Proceedings of the International

Conference on Machine Learning and Cybernetics, (Institute of Electrical and Electronics Engineers Inc., Beijing, 2002).

[4] J.I. Serrano and L. Araujo, Evolutionary Algorithm for Noun Phrase Detection in Natural Language Processing, *The 2005 IEEE Congress on Evolutionary Computation: Proceedings of the 2005 IEEE Congress on Evolutionary Computing*, (IEEE Computer Society, Edinburgh, 2005)

[5] O. Kimball, R. Iyer, H. Gish, S. Miller, and F. Richardson, Extracting descriptive noun phrases from conversational speech. In: *International Conference on Acoustics, Speech, and Signal Processing, 2002: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Institute of Electrical and Electronics Engineers Inc., Orlando, 2002).

[6] I. H. Witten, Z. Bray, M. Mahoui, and B. Teahan, Text mining: a new frontier for lossless compression. In: *Data Compression Conference 1999: Proceedings of the Data Compression Conference*, (Institute of Electrical and Electronics Engineers Inc., Snowbird, UT , 1999).

[7] H. Kaji, Y. Morimoto, T. Aizono, N. and Yamasaki, Corpus-dependent Association Thesauri for Information Retrieval. In: International Conference on Computational Linguistics 2000: Proceedings of the 18th International Conference on Computational Linguistics, (Saarbrücken, 2000).

[8] H. Ahonen-Myka, O. Hienonen, and M. Klemettinen, Finding Co-occurring Text Phrases by Combining Sequence and Frequent Set Discovery. In: *Text Mining Workshop 1999: Proceedings of the Text Mining Workshop at IJCAI'99*, (Stockholm, 1999).

[9] R. N. Kostoff, J. A. Block, Factor Matrix Text Filtering and Clustering. *Journal of the American Society for Information Science and Technology* 56(9) (2005) 946-968.

[10] W. J. Wilbur, Y. Yang, An Analysis of Statistical Term Strength and its Use in the Indexing and Retrieval of Molecular Biology Texts, *Computers in Biology and Medicine* 26(3) (1996) 209-222.

[11] R. Feldman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler, and O. Zamir. *Text Mining at the Term Level* In: Second European Symposium on Principles of Data Mining and Knowledge Discovery 1998: Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery, (Springer-Verlag, London, 1998)

[12] R. J. Watts, A.L. Porter, D.Z. Zhu, *Factor Analysis Optimization: Applied on Natural Language Knowledge Discovery* In: Committee on Data for Science and Technology 2002: Frontiers of Scientific and Technical Data: Proceedings of the 18th International Conference CODATA 2002, (Montreal, 2002)

[13] M. F. Porter, An Algorithm for Suffix Stripping, *Program* 14 (1989) 130-137.

[14] M. Agosti, M. Bacchin, N. Ferro, and M. Melucci, Improving the automatic retrieval of text documents. In: Advances in Cross-Language Information Retrieval 2003: Proceedings of the Third Workshop of the Cross-Language Evaluation Forum, Revised Papers, (Springer, Rome, 2003).

[15] I. Diaz, J. Morato, and J. Llorens, An algorithm for term conflation based on tree structures, *Journal of the American Society for Information Science and Technology* 53(3) (2002) 199-208.

[16] T. Kurz and K. Stoffel, Going beyond stemming: creating concept signatures of complex medical terms, *Knowledge-Based Systems* 15(5-6) (2002) 309-313.

[17] W. J. Wilbur and W. Kim, *Flexible Phrase Based Query Handling Algorithms*, In: Aversa E, Manley C, Eds: Proceedings of the ASIST 2001 Annual Meeting 2001 (Information Today, Inc, Medford, 2001)

[18] Y. Kadoya, M. Fuketa, E. S. Atlam, K. Morita, T. Sumitomo, and J. Aoe, A compression algorithm using integrated record information for translation dictionaries, *Information Sciences* 165(3-4) (2004) 171-186.

[19] M. Palakal, M. Stephens, S. Mukhopadhyay, R. Raje, and S. Rhodes, *A multi-level text mining method to extract biological relationships*. In: IEEE Computing Society Bioinformatics Conference 2002: Proceedings of the IEEE Computing Society Bioinformatics Conference, (IEEE Computer Society, Stanford, 2002).

[20] N. Ide and J. Veronis, Word Sense Disambiguation: The State of the Art, *Computational Linguistics* 24(1) (1998) 1-40

[21] W. Gale, K. Church, and D. Yarowsky A Method for Disambiguating Word Senses in a Large Corpus, *Computers and the Humanities* 26 (1992) 415-439

[22] D. Yarowsky, *Unsupervised Word Sense Disambiguation Rivaling Supervised Methods*, In: Annual Meeting of the Association for Computational Linguistics 1995: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (Morgan Kaufmann Publishers, Cambridge, 1995)

[23] G. Escudero, L. Marquez, and G. Rigau, Naïve Bayes and Exemplar-Based Approaches to Word Sense Disambiguation Revisited, In European Conference on Artificial Intelligence ECAI 2000: Proceedings of the 14th European Conference on Artificial Intelligence (IOS Press, Amsterdam, 2000)

[24] J. White, *Word List* (2004). Available at: **http://calendarhome.com/wordlist.html** (12 March 2004).

[25] F. Crestani, Exploiting the Similarity of Non-Matching Terms at Retrieval Time, *Information Retrieval* 2(1) (2000) pp. 25-45.

[26] A. Chowdhury, O. Frieder, D. Grossman, and M. C. McCabe, Collection statistics for fast duplicate document detection, *ACM Transactions on Information Systems* 20(2) (2002) 171-191.

[27] L. Egghe and C. Michel, Strong similarity measures for ordered sets of documents in information retrieval, *Information Processing & Management* 38(6) (2002) 823-848.

[28] B. Jun-Peng, S. Jun-Yi, L. Xiao-Dong, and S. Qin-Bao, *A new text feature extraction model and its application in document copy detection*. In: International Conference on Machine Learning and Cybernetics 2003: Proceedings of the 2003 International Conference on Machine Learning and Cybernetics, ed 1, (Institute of Electrical and Electronics Engineers Inc., Xi'an, 2003).

[29] H. Rezankova, D. Husek, J. Smid, and V. Snasel, *Clustering of documents via similarity measures*. In: International Conference on Communications in Computing 2003: Proceedings of the International Conference on Communications in Computing, (CSREA Press, Las Vegas, 2003).

[30] J. Basu, R. Mooney, K. V. Pasupuleti, and J. Ghosh, *Evaluating the novelty of text-mined rules using lexical knowledge*. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2001: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, San Francisco 2001).

[31] D. R. Cutting, D. R. Karger, J. O. Pederson, and J. W. Tukey, *Scatter-Gather: A Cluster-based Approach to Browsing Large Document Collections*. In: International ACM SIGIR Conference on Research and Development in Information Retrieval 1992: Proceedings of the 15th Annual International ACM SIGIR

Conference on Research and Development in Information Retrieval, (Association for Computing Machinery, Copenhagen, 1992).