CONTEXT-DEPENDENT CONFLATION, TEXT FILTERING AND CLUSTERING

By

Ronald N. Kostoff, Ph. D.
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217
Phone: 703-696-4198
Fax: 703-696-4274
Internet: kostofr@onr.navy.mil

Joel A. Block, M.D.
Rush Medical College
Chicago, IL 60612

KEYWORDS: Text Filtering; Trivial Word Filtering; Stop Word Lists; Stemming; Conflation; Factor Matrix; Factor Loading; Factor Analysis; Context-Dependent Trivial Words; Concept Clustering; Document Clustering; Raynaud's Phenomenon; Raynaud's Syndrome; Raynaud's Disease; Fractals

ABSTRACT

The presence of trivial words in text databases can impact record or concept (words/ phrases) clustering adversely. Additionally, the determination of whether a word/ phrase is trivial is context-dependent. The objective of the present paper is to demonstrate a context-dependent trivial word filter to improve clustering quality. Factor analysis was used as a context-dependent trivial word filter for subsequent term clustering. Medline records for Raynaud's Phenomenon were used as the database, and words were extracted from the record Abstracts. A factor matrix of these words was generated, and the words that had low factor loadings across all factors were identified, and eliminated. The remaining words, which had high factor loading values for at least one factor and therefore were influential in determining the theme of that factor, were input to the clustering algorithm. Both quantitative and qualitative analyses were used to show that factor matrix filtering leads to higher quality clusters and subsequent taxonomies.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01 SEP 2004		**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
CONTEXT-DEPENDENT CONFLATION, TEXT FILTERING AND CLUSTERING	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
Ronald Kostoff; Joel Block	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Office of Naval Research,800 N. Quincy St.,Arlington,VA,22217	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The presence of trivial words in text databases can impact record or concept (words/ phrases) clustering adversely. Additionally, the determination of whether a word/ phrase is trivial is context-dependent. The objective of the present paper is to demonstrate a context-dependent trivial word filter to improve clustering quality. Factor analysis was used as a context-dependent trivial word filter for subsequent term clustering. Medline records for RaynaudŸs Phenomenon were used as the database, and words were extracted from the record Abstracts. A factor matrix of these words was generated, and the words that had low factor loadings across all factors were identified, and eliminated. The remaining words, which had high factor loading values for at least one factor and therefore were influential in determining the theme of that factor, were input to the clustering algorithm. Both quantitative and qualitative analyses were used to show that factor matrix filtering leads to higher quality clusters and subsequent taxonomies. Additionally, a fractals database obtained from the Science Citation Index was used to demonstrate the value of factor matrices to determine interchangeability of word variants, and show the context dependency requirements for conflation.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**48**	

Additionally, a fractals database obtained from the Science Citation Index was used to demonstrate the value of factor matrices to determine interchangeability of word variants, and show the context dependency requirements for conflation.

INTRODUCTION

Science and technology (S&T) form the core of modern economies and militaries. Global S&T expenditures are in the neighborhood of 500 billion dollars to a trillion dollars annually, depending on one's definition of S&T. No single organization, or even nation, can begin to cover the full spectrum of S&T development required for a modern competitive economy or military. Cooperative S&T development efforts, leveraging, exploiting, and awareness of external S&T efforts are required if an organization or nation is to remain competitive.

Governments and industrial organizations need ready access to the results of all global research performed in order to:

1) Track research impacts, to help identify benefits arising from sponsored research;
2) Evaluate science and technology programs;
3) Avoid research duplication;
4) Identify promising research directions and opportunities;
5) Perform myriad oversight tasks; and, in general,
6) Support every step of a strategic research planning/ selection/ management/ evaluation process that makes optimal use of S&T investment resources.

In addition, recent counter-terrorism concerns have highlighted the need for ready access to, and analysis of, databases that could link people with institutions and activities. In the S&T arena, this requires linking research performers with organizations, countries, and technical areas.

Complementing this massive global S&T expenditure is equally massive documentation that can be collectively called the global S&T literature. It consists of myriad S&T planning and vision/ requirements documents (S&T planning literature), S&T program descriptive documents (ongoing S&T program literature), evaluation documents of ongoing and completed S&T programs/ projects (S&T program assessment literature), and myriad S&T

output and product documents (S&T output literature, such as papers, patents, etc). In order to be able to extract useful information from this massive literature, semi-automated text analysis processes, known collectively as text mining, are required.

One analytical technique commonly used for achieving most of the objectives listed above is a component of text mining: clustering of related textual objects. Clustering is fundamentally a separation process, and is used in many different disciplines, such as isotope separation in chemistry and physics, and impurity separation in water purification. In text analysis, clustering is intrinsically more complicated than in the physical separation processes, because of the multiple meanings and contextual dependence of words, phrases, and word/ phrase patterns.

Additionally, in the S&T text, the high technical content phrases/ words are imbedded in a much larger sea of low technical content words/ phrases, known collectively as trivial words or stop words. If these context-dependent trivial words/ phrases are retained during the clustering process, they can then become the nucleation centers for clustering rather than the desired high technical content words/ phrases, thereby leading to diffuse and misleading clusters. Removing these context-dependent trivial words/ phrases prior to the clustering process would be a major contributor towards defining the clusters more sharply and accurately.

One candidate method for removing context-dependent trivial words prior to clustering is factor analysis (Kendall, 1956; Kaiser, 1960; Cattell, 1966; Cooper, 1983; Coovert and McNelis, 1988; McArdle, 1990; Jackson, 1991; Yalcin and Amemiya, 2001; Browne, 2001), commonly used to identify the pervasive themes in text databases based on correlations, and subsequently group these themes. Factor analysis tends to provide a better estimate of quantitative relationships among the theme components than cluster analysis, whereas cluster analysis tends to provide a better estimate of the structural relationships among the themes, especially hierarchical clustering methods.

Factor and cluster analysis algorithms have existed for decades, and are well validated. Understood less well are how factor and cluster analysis should operate synergistically, and how best to select the data required for input to these algorithms. This study proposes the use of factor analysis as a context-dependent word filter for cluster analysis.

The factor matrix filtering approach first identifies the high technical content words from raw text using factor analysis, and discards the remainder as trivial. It then uses the high technical content words as input to the clustering algorithm. The paper provides an estimate of the benefit of this approach, using a Raynaud's Phenomenon database at the test bed. This paper starts with the Background of the various study elements, describes the Approach used, presents the Results, and ends with Conclusions.

BACKGROUND

As summarized in the Introduction, this paper includes text mining, clustering, trivial word removal, factor analysis, and Raynaud's Phenomenon. The present section will provide sufficient background material on each of these topics to clarify the procedures described in the Approach section.

Text Mining

Text mining has been developed to extract useful information from the global S&T literature, in order to supplement conventional human-based approaches (Hearst, 1999; Trybula, 1999; Feldman, 1999; Lagus et al, 1999; Weiss et al, 1999; Losiewicz et al, 2000; Kuhnhold, 2000; Visa, 2001; Kostoff et al, 2001c; Zhu and Porter, 2002; Kogan et al, 2003; Perrin and Petry, 2003). Its component capabilities of *computational linguistics* and *bibliometrics* can be summarized as follows.

Science and technology *computational linguistics* [Kostoff, 2003a; Hearst, 1999; Zhu and Porter, 2002; Losiewicz et al, 2000] is a process that underlies the extraction of useful information from large volumes of technical text. It identifies pervasive technical themes in large databases from technical phrases that occur frequently. It also identifies relationships among these themes by grouping (clustering) these phrases (or their parent documents) on the basis of similarity. *Computational linguistics* can be used for:

- Enhancing information retrieval and increasing awareness of the global technical literature [Kostoff et al, 1997; Greengrass, 1997; TREC, 2003]

- Potential discovery and innovation based on merging common linkages among very disparate literatures [Swanson, 1986; Swanson and Smalheiser, 1997; Kostoff, 2003b; Gordon and Dumais, 1998]
- Uncovering unexpected asymmetries from the technical literature [Goldman et al, 1999; Kostoff, 2003c]
- Estimating global levels of effort in S&T sub-disciplines [Kostoff et al, 2000a, 2002, 2004a; Viator and Pestorius, 2001]
- Helping authors potentially increase their citation statistics by improving access to their published papers, and thereby potentially helping journals to increase their Impact Factors [Kostoff et al, 2004a, 2004b]
- Tracking myriad research impacts across time and applications areas [Davidse and VanRaan, 1997; Kostoff et al, 2001b].

Evaluative *bibliometrics* [Narin, 1976; Garfield, 1985; Schubert et al, 1987] uses counts of publications, patents, citations and other potentially informative items to develop science and technology performance indicators. Its validity is based on the premises that 1) counts of patents and papers provide valid indicators of R&D activity in the subject areas of those patents or papers, 2) the number of times those patents or papers are cited in subsequent patents or papers provides valid indicators of the impact or importance of the cited patents and papers, and 3) the citations from papers to papers, from patents to patents and from patents to papers provide indicators of intellectual linkages between the organizations that are producing the patents and papers, and knowledge linkage between their subject areas [Narin et al, 1994]. Evaluative *bibliometrics* can be used to:

- Identify the infrastructure (authors, journals, institutions) of a technical domain (Kostoff et al, 2000b, 2004b),
- Identify experts for innovation-enhancing technical workshops and review panels,
- Develop site visitation strategies for assessment of prolific organizations globally,
- Identify impacts (literature citations) of individuals, research units, organizations, and countries (Kostoff et al, 2001b, 2004c)

Evaluative bibliometrics can also used to help generate extensive background material for research papers, and for comprehensive literature reviews and surveys. The documents most cited (relative to their contemporaries) by a retrieved topical literature of interest can be considered

to be seminal, and form the core of the background material. Other relevant documents can be added to enhance the background material and eliminate gaps in the narration. Another advantage of this citation-assisted background (CAB) approach (Kostoff, 2004d) over traditional literature reviews is that the core seminal papers identified are based on the larger technical community's consensus (highest citations), rather than solely based on the author(s) personal experiences and biases.

A typical text mining study of the published literature develops a query for comprehensive information retrieval, processes the retrieved database using computational linguistics and bibliometrics, and integrates the processed information.

Clustering

Clustering is grouping by attributes. Since technical text can contain many attributes, many taxonomies can be generated from a body of text. Two types of clustering that have been used extensively by the first author are concept clustering and document clustering, Concept clustering is the grouping of related words or phrases to identify technical themes in the text database. It has been used to generate literature taxonomies (Kostoff and DeMarco, 2001c), to facilitate Web searching (Khare, 2003), to summarize text (Ko et al, 2003), to generate hypotheses and discovery (Stegmann and Grohmann, (2003), and to generate thesauri (Hodge and Austin, 2002). Document clustering (Cutting et al, 1992a; Guha et al, 1998; Hearst et al, 1998; Karypis et al, 1999; Rasmussen, 1992; Steinbach et al, 2000; Willet, 1988; Zamir and Etzioni, 1998; Karypis, 2002; Guerrero-Bote et al, 2003; Casillas et al, 2003; Schenker et al, 2003) is the grouping of related documents by theme.

Clusters can be aggregated into a hierarchical structure, to provide a taxonomy or classification scheme of the discipline(s) being studied. The quality of the final clusters and taxonomy is strongly dependent on the quality of the words selected for input to the factor and cluster analyses. If important high technical content words and phrases are omitted from the input, the themes derived from these words will be lost to the final results. If too many non-technical words are selected for the input, then artificial clusters will be generated based on overlap of non-technical words, and/ or words/ phrases will be re-assigned among clusters due to non-technical linkages. A misleading taxonomy will result.

One of the reviewers of this paper asked how the proposed improved clustering approach could lead to improved information retrieval from Medline (the source of the Raynaud's records used as the database). The response depends on how clustering is employed in the information retrieval process.

In the first author's Simulated Nucleation information retrieval approach (Kostoff et al, 1997), clustering (including the factor matrix clustering approach presented in the present paper) is used to reduce time spent on query development, as well as to expand the query. This iterative relevance feedback query expansion technique starts with generation of a test query. Records are then retrieved from a source database (e.g., Medline) using this query. The next series of steps involves separating the relevant from non-relevant records, and identifying text patterns characteristic of relevant records but essentially non-existent in non-relevant records, and vice versa. The query is then modified with these patterns so that it will retrieve more relevant records, and will filter out more non-relevant records.

Before clustering was used by the first author to supplement the information retrieval process, the separation of relevant from non-relevant records was performed manually. Many records (Abstracts, if records are journal articles) were read, and the relevant and non-relevant were separated. Computational linguistics were performed on each group (relevant, non-relevant) of records, to identify the text patterns characteristic of each group.

For the past few years, the first author has used clustering for the relevant/ non-relevant separation. Records are retrieved, then immediately clustered. In almost all cases, the relevant records will cluster together, as will the non-relevant records. Substantial time is saved by substituting computer-based clustering for manual separation. Additionally, when selecting additional words for the query, words can be selected from all the clusters to insure that all the main concepts have at least some representation in the query. Factor matrix filtering provides sharper clusters, insuring that query terms selected will contribute to improving recall and precision.

Trivial Word Removal

An intrinsic problem in any text grouping procedure, or indeed in any text feature extraction procedure, is dilution of the results by the inclusion of

trivial words/ phrases. In technical literature, trivial words/ phrases are terms with little technical information content. The most common trivial words are generic connectors such as ‘the’, ‘of’, ‘and’, etc. They tend to have low technical content regardless of context. Many other words become trivial in selected contexts.

The presence of trivial words can alter the results of clustering substantially. Experiments by the first author on the details of technical text clustering in project narratives and journal article Abstracts showed that, on average, there were very few non-trivial words in each document that were important for determining the main theme of the article. In an ideal world, these few words would serve as the anchors for clustering, and produce sharp well-defined clusters. Unfortunately, the presence of the trivial words/ phrases provides competing anchors for the clusters, and generates clusters and taxonomy structures that are less-well defined. As stated in (Wang et al, 2003), “In many document data sets, only a relatively small number of the total features may be useful in classifying documents, and using all the features may adversely affect performance". Any techniques that would pre-process data prior to clustering, and eliminate any words/ phrases that are trivial in the context of the application, would help sharpen the resultant clusters.

Most, if not all, text mining approaches start this word selection/ filtering process by attempting to reduce the dimensionality of the system (reducing the number of words or phrases to be manipulated). After the raw text of the literature to be analyzed is converted to words or phrases, all text mining approaches separate out the ‘trivial’ terms from the high technical content terms. Almost all the approaches reported in the literature assume the ‘trivial’ terms are context-independent, and these approaches use a pre-determined trivial word or stop list to remove such words from the initial pool of words.

One innovative approach for identifying context-dependent trivial words selected articles about genes from diverse classes, then eliminated the high frequency words in the total dataset by assuming these high frequency words were generic rather than gene specific, and would serve to diffuse the clustering process (Kankar et al, 2002).

Some approaches use statistical techniques to eliminate trivial words/ phrases. A key concept is that words/ phrases that appear uniformly over the collection are trivial from the viewpoint of theme discrimination among documents/ concepts. Weighting terms based on maximal tf*idf (term frequency-inverse document frequency) is a popularly-used filter (Salton and Buckley, 1998). Feldman et al (1998) use the standard deviation of the relative frequency of a term over all the documents of the collection as a weighting. Stensmo's probabilistic information retrieval system (Stensmo, 2002) bypasses the need for stop-word lists by removing unneeded parameters dynamically based on a local mutual information measure.

Perhaps the most extensive work in this area is attributable to Bookstein (Bookstein et al, 1995, 1998, 2003) and Wilbur (Wilbur and Sorotkin, 1992; Kim and Wilbur, 2001; Wilbur and Yang, 1996). Bookstein asserts that the greater the deviation of a term's distribution from a Poisson distribution, the more likely that the term is a useful one. His approach examines term clumping tendencies in full text, and is not very relevant to Abstracts or titles. Wilbur examines a variety of statistical measures to estimate term importance, including term strength (how strongly the term's occurrences correlate with the subjects of documents in the database).

Other techniques use noun phrase extraction based on natural language processing, such as part-of-speech tagging and filtering (Brill, 1993, 1994; Cutting, 1992b). Identification of parts of speech may not be easily made for literature that describes the latest science and technology, with its continual infusion of new terminology. Lexicons, dictionaries, and thesauri have an intrinsic lag time, and consequently new terms may appear incomprehensible to any computer-based tagger. Additionally, there are both semantic and syntactic components to identifying high quality phrases. While tagging can address the syntactic issue, it cannot eliminate the semantic problem. As stated in the above papers, all the statistical feature selection approaches and part-of-speech tagging approaches leave much to be desired, mainly because of the influence of context on feature desirability.

After the words/ phrases have been selected initially, almost all techniques use some further processing of the words/ phrases before input to the factor or clustering algorithms. Many techniques use stemming (e.g., Porter, 1980; Kostoff, 2003d, see also Appendix 3 of present document), where words are grouped according to common roots, and conflated (singulars are combined with their plurals, full spellings are combined with their acronyms, and

different tenses are combined). A few techniques select synonyms from a controlled vocabulary to reduce the number of words while retaining the concepts (see Weeber et al (2001), which used a medical thesaurus for this purpose in literature-based discovery). Finally, for those techniques that use both factor analysis and clustering, each approach is pursued independently.

It is the contention of the present paper that the words to be selected as input to the cluster analyses should be context-dependent. 'High-technical content' has different meanings for different literatures and applications. 'Trivial' has different levels of context dependency. Stemming and conflation should be dependent on context as well. This paper will show how context-dependency can be used in the word or phrase selection process through factor matrix filtering. A companion paper shows how context-dependency can be incorporated in the conflation process through factor matrix filtering (Kostoff, 2003d, see also Appendix 3 of present document).

Factor Analysis

Factor analysis of a text database aims to reduce the number of words/ phrases (variables) in a system, and to detect structure in the relationships among words/ phrases. Word/ phrase correlations are computed, and highly correlated groups (factors) are identified. The relationships of these words/ phrases to the resultant factors are displayed clearly in the factor matrix, whose rows are words/ phrases and columns are factors. In the factor matrix, the matrix elements Mij are the factor loadings, or the contribution of word/ phrase i (in row i) to the theme of factor j (in column j). The theme of each factor is determined by those words/ phrases that have the largest values of factor loading. Each factor has a positive value tail and negative value tail. For each factor, one of the tails dominates in terms of absolute value magnitude. This dominant tail is used to determine the central theme of each factor.

One of the key challenges in factor analysis is defining the number of factors to select. Different approaches have been suggested in the literature, but the two most widely used are the Kaiser criterion (Kaiser, 1960; Jackson, 1991), and the Scree test (Cattell, 1966). The Kaiser criterion states that only factors with eigenvalues greater than unity should be retained, essentially requiring that a factor extracts at least as much variance as the equivalent of one original variable. The Scree test plots factor eigenvalue (variance) vs factor number, and recommends that only those factors that extract

substantive variance be retained. Operationally, the factor selection termination point becomes the 'elbow' of the Scree plot, the point where the slope changes from large to small.

As one of the reviewers of this paper correctly noted, the "interpretation of the Scree Plot is partly subjective", and "consistency among different interpreters is low". Appendix 1 discusses this interpretation problem in more detail, and shows that part of the inconsistency may stem from the fractal-like nature of the Scree Plots. Appendix 1 also discusses the consequences of selecting factor matrices with different numbers of factors.

In most previous studies performed by the first author, the Kaiser criterion has been used to select the number of factors for the factor matrix. These previous studies have used an Excel add-in to generate the factor matrices, and, due to Excel's limitations on columns, have been limited approximately to 250 x 250 correlation matrices, or 250 words. The Kaiser criterion has yielded factor numbers in the range of 20-45, considered a reasonable number for analysis. However, in the present Raynaud's Phenomenon study, another software package that did not require Excel was used (TechOasis), and 659 words were used for the correlation matrix. The Kaiser criterion yielded 224 factors, a number far too large for detailed factor analysis, and of questionable utility, since many of the eigenvalues were not too different from unity. It was decided to examine the Scree Plot for factor number determination.

Raynaud's Phenomenon

Both authors are conducting a study of Raynaud's Phenomenon (a peripheral circulatory disorder) using text mining, to extend the literature-based discovery techniques from Swanson's classical paper (Swanson, 1986). Of central interest are the pervasive medical themes of the Raynaud's Phenomenon literature, identified by factor analysis and cluster analysis.

Since each factor from the factor analysis, or cluster from the cluster analysis, addresses some aspect of Raynaud's Phenomenon, an overview of Raynaud's Phenomenon will be presented before discussing the factor and cluster results. Because the main Raynaud's terminology used in the literature is not consistent (in many cases, Raynaud's Disease is used interchangeably with Raynaud's Phenomenon or Raynaud's Syndrome), the overview will include the distinction among these Raynaud variants.

Raynaud's Phenomenon is a condition in which small arteries and arterioles, most commonly in the fingers and toes, go into spasm (contract) and cause the skin to turn pale (blanching) or a patchy red (rubor) to blue (cyanosis). While this sequence is normally precipitated by exposure to cold, and subsequent re-warming, it can also be induced by anxiety or stress. Blanching represents the ischemic (lack of adequate blood flow) phase, caused by digital artery vasospasm. Cyanosis results from de-oxygenated blood in capillaries and venules (small veins). Upon re-warming, a hyperemic phase ensues, causing the digits to appear red.

Raynaud's Phenomenon can be a primary or secondary disorder. When the signs of Raynaud's Phenomenon appear alone without any apparent underlying medical condition, it is called Primary Raynaud's, or formerly, Raynaud's Disease. In this condition, the blood vessels return to normal after each episode. Conversely, when Raynaud's Phenomenon occurs in association with an underlying condition or is due to an identifiable cause, then it is referred to as Secondary Raynaud's, or formerly, as Raynaud's Syndrome. The most common underlying disorders associated with Secondary Raynaud's are the auto-immune disorders, or conditions in which a person produces antibodies against his or her own tissues. In contrast to Primary Raynaud's, where the blood vessels remain anatomically normal after each episode, in Secondary Raynaud's there may be scarring and long-term damage to the blood vessels; thus Secondary Raynaud's is potentially a more serious disorder than Primary. Certain repetitive activities may result in a predisposition to Raynaud's Phenomenon. These cases of so-called "Occupational Raynaud's" typically result from the chronic use of vibrating hand tools.

Thus, while Raynaud's Phenomenon is a direct consequence of reduced blood flow due to reversible blood vessel constriction, it may be a function of many variables that can impact blood flow. These include:

*Inflammation from the auto-immune disorders that can cause swelling and thereby constrict blood vessels;
*Increased sympathetic nervous system activity, that can affect the timing and duration of the blood vessel muscular contractions that cause constriction;
*Heightened digital vascular reactivity to vaso-constrictive stimuli, that cause the blood vessels to over-react and over-contract;

*Deposits along the blood vessel walls that can reduce blood flow and increase the flow sensitivity to contraction stimuli;
*Blood rheological properties that offer additional resistance to blood flow, and magnify the impact of blood vessel constriction;
*Blood constituents and hormones that can act as vaso-constrictors or vaso-dilators.

APPROACH

In the first part of this study, 930 Medline Abstract-containing records related to Raynaud's Phenomenon, and published in the 1975-1985 time period (to approximate Swanson's database), were retrieved with a Raynaud's-specific query. These Abstracts were subjected to factor analysis and clustering, as part of the analysis.

All the single words were automatically extracted from the database of 930 Abstracts, subject to elimination of very trivial words (removing words like 'of', 'the', 'and', 'if', etc). Non-trivial single words (659) were then manually extracted (by experts) from the database of Abstracts, along with the number of documents in which each word appeared (document frequency). For this database, the 659 extracted words were near the limit that allowed a factor matrix to be computed. The co-occurrence of word pairs in the same document (word co-occurrence frequency) was computed, and a correlation matrix (659 x 659) of word pairs was generated. The variables were factorized, and a factor matrix was generated.

Factor Matrix Generation

Once the desired number of factors has been determined from the Scree Plot 'elbow', and the appropriate factor matrix has been generated, the factor matrix can then be used as a filter to identify the significant technical words for further analysis. Specifically, the factor matrix can complement a basic trivial word list (i.e., a list containing words that are trivial in almost all contexts, such as 'a', 'the', 'of', 'and', 'or', etc) to select context-dependent high technical content words for input to a clustering algorithm. The factor matrix pre-filtering will improve the cohesiveness of clustering by eliminating those words that are trivial words operationally in the application context.

The variance accounted for by each underlying factor (eigenvalue) was generated by Principal Components Analysis. Figure 1 shows the factor eigenvalue-factor number plot (Scree Plot) for the 659 un-rotated factors on a linear scale. The 'elbow', or break point, of the curve appears to be about fourteen factors.

INSERT FIGURE 1

Factor Matrix Filtering

The fourteen factor matrix determined by the Scree Plot of Figure 1 was examined in detail, and is presented in the Results section. To diversify the factor loading patterns, and simplify interpretation of each factor, varimax orthogonal rotation was used.

Factor Matrix Word Filtering and Selection

After the factor matrix has been generated, its highest technical content words are input to the clustering algorithm. In the present experiment, the 659 words in the factor matrix would have to be culled to the ~250 allowed by the Excel-based clustering package, WINSTAT. The ~250 word limit is an artifact of Excel. Other software packages may allow more or less words to be used for clustering, but all approaches perform culling to reduce dimensionality. The filtering process presented here is applicable to any level of filtered words desired.

Another caveat. A trivial word list of the type described previously (words that are trivial in almost all contexts) was used to arrive at the 659 words used for the factor matrix input. This was not necessary. The raw words from the word generator could be used as input, and would be subject to the same filtering process. To allow more important words to be used in this demonstration, the very trivial words were removed.

The factor loadings in the factor matrix were converted to absolute values. Then, a simple algorithm was used to automatically extract those high factor loading words at the dominant tail of each factor. The highest absolute value of factor loading for each word/ phrase was identified from the total factor matrix. The words/ phrases were ranked in inverse order of highest absolute value of factor loading. If word variants were on this list (e.g., singles and plurals), and their factor loadings were reasonably close

(Kostoff, 2003d), they were conflated (e.g., 'agent' and 'agents' were conflated into 'agents', and their frequencies were added). All words/ phrases below the ~250 term limit allowed by Excel were eliminated.

RESULTS

Factor Matrix Analysis

For the fourteen factor matrix, the high factor loading words in the dominant tail of each factor are shown in parentheses after the factor number, followed by a brief narrative of the factor theme.

Factor 1 (nuclear, antibodies, extractable, speckled, connective, immuno-fluorescence, antinuclear, tissue, anti-RNP, MCTD, mixed, ribonucleoprotein, swollen, RNP, antibody, antigen, titer, SLE, lupus, erythematosus) focuses on different types of autoantibodies, especially anti-nuclear and extractable nuclear, and their relation to auto-immune diseases.

Factor 2 (double-blind, placebo, mg, daily, weeks, times, agent, nifedipine, trial) focuses on double-blind trials for vasodilators.

Factor 3 (vibration, tools, workers, vibrating, exposure, chain, prevalence, time, exposed, sensory, white, circulatory, complaints) focuses on the impact of vibratory tools on circulation.

Factor 4 (coronary, ventricular, heart, angina, hypertension, myocardial, cardiac, failure, pulmonary) focuses on coronary circulation and blood pressure problems.

Factor 5 (prostaglandin, platelet, E1, prostacyclin, aggregation, infusion, hours, healing, ischaemic, thromboxane, administered, vasodilator, intravenous) focuses on the administration of vasodilators to improve circulation.

Factor 6 (calcinosis, sclerodactyly, esophageal, dysmotility, telangiectasia, anticentromere, variant, diffuse, scleroderma) focuses on scleroderma-spectrum types of autoimmune diseases.

Factor 7 (extremity, sympathectomy, artery, surgery, arteries, upper, occlusions, arterial, brachial, thoracic, operation, surgical, angiography,

occlusive) focuses on surgical solutions to remove constrictions on circulation.

Factor 8 (C, degrees, systolic, pressure, cooling, blood, finger, measured, flow) focuses on blood flow, and associated finger blood pressure and temperature measurements.

Factor 9 (capillaries, capillary, nail-fold, microscopy, capillaroscopy) focuses on the diagnostic use of nail-fold capillary microscopy.

Factor 10 (training, biofeedback, relaxation, stress, outcome, measures, headaches, temperature, conducted, thermal, physiological, responses) focuses on the use of biofeedback training to reduce stress headaches, and raise temperatures through improved circulation.

Factor 11 (vasodilation, peripheral, immersion, calcium, water) focuses on vasodilation of the peripheral circulatory system after immersion, and the role of calcium in this process.

Factor 12 (complexes, immune, circulating, complement, IgG, serum, levels, IgM) focuses on serum levels of circulating immune complexes and immunoglobulins, especially IgG and IgM.

Factor 13 (eosinophilia, fasciitis, fascia, eosinophilic, visceral, hypergammaglobulinemia, absent, scleroderma-like, corticosteroids) focuses on inflammation, especially of the fascia.

Factor 14 (systemic, lupus, RA, erythematosus, PSS, sclerosis, rheumatoid, arthritis, SLE) focuses on autoimmune diseases associated with Raynaud's Phenomenon.

The fourteen factor matrix themes can be divided into the two main thrusts of circulation and autoimmunity, where circulation covers factors 2, 3, 4, 5, 7, 8, 9, 10, and 11, and autoimmunity covers factors 1, 6, 12, 13, 14.

An examination of the words eliminated and those retained showed that most of those retained appeared to have high technical content, and would have been selected by previous manual filtering processes for input to the clustering algorithms. Some of the words in isolation appeared not to have the highest technical content, also as shown above, but it was concluded that

they were important because of their contribution to theme determination in the present clustering application. Similarly, some of the words eliminated by the factor matrix filter appeared to be high technical content, and in previous manual filtering processes might have been selected for the clustering algorithm input (e.g., acrocyanosis, vasomotor, cerebral, gastrointestinal). The conclusion for these words was not that they were unimportant per se. Rather, they did not have sufficient influence in determining the factor themes, and would not make an important contribution to the cluster structure determination. Thus, the context dependency (their influence on factor theme determination) of the words was the deciding factor in their selection or elimination, not only the judgement of their technical value in isolation (independent of factor theme determination), as was done in previous manual filtering approaches.

Word Clustering

The 252 filtered and conflated words were input to the WINSTAT clustering algorithm, and the Average Link option was selected for clustering. Figure 2 is the dendrogram of the 252 words. This is a tree-like structure that shows how the individual words cluster into groups in a hierarchical structure. One axis is the words, and the other axis ('distance') reflects their similarity. The lower the value of 'distance' at which words, or word groups, are linked together, the closer their relation. As an extreme case of illustration, words that tend to appear as members of multi-word phrases, such as 'lupus erythematosus', 'connective tissue', or 'double blind' appear adjacent on the dendrogram with very low values of 'distance' at their juncture.

INSERT FIGURE 2

Now the major structures, or clusters, will be described, following the hierarchical structure of the dendrogram. The capitalized words listed in parentheses after each cluster number are the boundaries of that cluster from the dendrogram. Only the top three hierarchical levels will be described.

The top hierarchical level can be divided into two major clusters. Cluster 1 (PATIENTS-OLD) focuses on autoimmunity, and Cluster 2 (TREATMENT-ACID) focuses on circulation. The second hierarchical level can be divided into four clusters, where Cluster 1 is divided into Clusters 1a and 1b, and Cluster 2 is divided into Clusters 2a and 2b.

Cluster 1a (PATIENTS-NEUROPATHY) focuses on autoimmune diseases and antibodies, while Cluster 1b (LESIONS-OLD) focuses on inflammation, especially fascial inflammation. Cluster 2a (TREATMENT-CONSECUTIVE) focuses on peripheral vascular circulation, while Cluster 2b (PULMONARY-ACID) focuses on coronary vascular circulation. These four high level categories are not computer artifacts, but correspond extremely well to how medical problems with a Raynaud's Phenomenon component are diagnosed and treated in medical practice.

Most of the clusters in the second hierarchical level can be rationally divided into two sub-clusters, to produce the third hierarchical level clusters. Cluster 1a1 (PATIENTS-MARKER) has multiple themes: different types of antibodies, especially anti-nuclear and extractable nuclear, and their relation to autoimmune diseases; sclerotic types of autoimmune diseases; and autoimmune diseases associated with Raynaud's Phenomenon. It incorporates the themes of Factors 1, 6, and 14. Cluster 1a2 (SERUM-NEUROPATHY) focuses on circulating immune complexes, and parallels the theme of Factor 12. Cluster 1b (LESIONS-OLD) is too small to subdivide further, and stops at the second hierarchical level. It parallels the theme of Factor 13.

Cluster 2a1 (TREATMENT-RESERPINE) has multiple themes: double-blind clinical trials for vasodilators; administration of vasodilators to reduce platelet aggregation and improve circulation; blood flow, and associated finger blood pressure and temperature measurements; and occupational exposures, mainly vibrating tools and vinyl chloride, that impact the peripheral and central nervous systems and impact circulation. It incorporates the themes of Factors 2, 3, 5, 7, 8. Cluster 2a2 (CAPILLARY-CONSECUTIVE) focuses on nailfold capillary microscopy as a diagnostic for for microcirculation, and parallels the theme of Factor 9. Cluster 2b1 (PULMONARY-LUNG) focuses on cardiovascular system problems, and parallels the theme of Factor 4. Cluster 2b2 (BIOFEEDBACK-ACID) focuses on biofeedback training to reduce stress and headaches, and increase relaxation, and parallels the theme of Factor 10.

Thus, use of the factor matrix for context-dependent trivial word elimination has produced a taxonomy that is technically defensible. What is the evidence that this taxonomy is improved compared to a taxonomy resulting from non-use of factor matrix filtering? There are two tandem approaches for comparing the quality of taxonomies, quantitative and qualitative. The

quantitative approach identifies metrics for gauging the cohesiveness and uniqueness of clusters, and evaluates the performance of each taxonomy against the metrics. The qualitative approach examines the overall taxonomy structure as well as the individual clusters for reasonableness and technical defensibility.

Appendix 2 describes the marginal impact on taxonomy quality from the substitution of trivial words for technical content words. Appendix 2 shows that, for the word clustering approach of this paper, the substitution of trivial words for higher technical content words has both quantitative and qualitative consequences. First, the 'distance' of the overall dendrogram (the 'distance' of the final aggregation-highest point on the dendrogram) increases when the trivial words are inserted. The effect of adding trivial words is to reduce the sharpness and uniqueness of individual clusters, and enhance linkages among disparate clusters through the trivial words only. Since 'distances' are low when words in a cluster are very closely related, 'distances' increase as the relations become more diffuse.

Second, the trivial words can act as a magnet, and change the balance of words in a cluster. In Appendix 2, the two trivial words used for replacement (the highest frequency words 'of', 'the') appeared on the dendrogram in the same first level cluster. They, in turn, attracted words formerly in the other first level cluster, and in some cases, these words were shifted to the less defensible technical category.

Appendix 2 shows further that use of the factor matrix filtering for selecting the words input to the cluster above (compared to use of the highest frequency words with no filtering) had two consequences. The overall taxonomy 'distance' decreased with the use of factor matrix filtering, and the word assignment to clusters improved from a technical perspective. Thus, factor matrix filtering improved the quality of the taxonomy and its constituent clusters. The only negative feature of factor matrix filtering is the modest time required for incorporation of this additional analytical step.

DISCUSSION AND CONCLUSIONS

Factor matrix filtering is an effective method for identifying the major themes in a text database, identifying the critical words that define the theme, selecting these critical words in context for clustering, and

identifying which variants of these words can be conflated within the context of the specific database examined.

REFERENCES

Bookstein, A., Klein, S.T., Raita, T. Detecting content-bearing words by serial clustering. (1995). *Proc. 18-th ACM-SIGIR Conf.*, Seattle, WA. 319-327.

Bookstein, A., Klein, S.T., Raita, T. (1998). Clumping properties of content-bearing words. Journal of the American Society for Information Science. 49 (2). 102-114. Feb.

Bookstein, A., Kulyukin, V., Raita, T., Nicholson, J. (2003). Adapting measures of clumping strength to assess term-term similarity. Journal of the American Society for Information Science and Technology. 54 (7). 611-620. May.

Brill, E. (1993). Automatic grammar induction and parsing free text: A transformation-based approach. Proceedings of the 31st meeting of the Association of Computational Linguistics. Columbus, OH.

Brill, E. (1994). Some advances in transformation-based part of speech tagging. Proceedings, Twelfth National Conference on Artificial Intelligence. Seattle, WA.

Browne, M.W. (2001). An overview of analytic rotation in exploratory factor analysis. Multivariate Behavioral Research. 36 (1). 111-150.

Casillas, A, de Lena, M.T.G., and Martinez, R. (2003). Document clustering into an unknown number of clusters using a genetic algorithm. Text, Speech and Dialogue, Proceedings, 2807. Lecture Notes in Artificial Intelligence. 43-49.

Cattell, R.B. (1966). The Scree Test for the number of factors. Multivariate Behavioral Research. 1. 245 –276.

Cooper, J.C.B. (1983). Factor-Analysis - An overview. American Statistician, 37 (2). 141-147.

Coovert, M.D., and McNelis, K. (1988). Determining the number of common factors in factor-analysis - A Review and Program. Educational and Psychological Measurement, 48 (3). 687-692. Fall.

Cutting, D. R., Karger, D. R, Pedersen, J. O. and Tukey, J. W. (1992a). Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR'92). 318-329.

Cutting, D., Kupiec, J., Pederson, J., and Sibun, P. (1992b). A practical part of speech tagger. Presented at the Third International ACL Conference on Applied Natural Language Processing. Trento, Italy.
Davidse, R.J., Van Raan, A.F.J. (1997). Out of particles: impact of CERN, DESY, and SLAC research to fields other than physics. Scientometrics 40:2 . 171-193.
Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O., Rajman, M., Schler, Y., and Zamir, O. (1998). Text mining at the term level. Principles of Data Mining and Knowledge Discovery. Lecture Notes in Artificial Intelligence. 1510. 65-73.
Feldman, R. (1999). Text mining via information extraction. Principles of Data Mining and Knowledge Discovery. 1704. 165-173.
Garfield, E. (1985). History of citation indexes for chemistry - a brief review. JCICS. 25(3): 170-174.
Goldman, J.A., Chu, W.W., Parker, D.S., Goldman, R.M. (1999). Term domain distribution analysis: a data mining tool for text databases. Methods of Information in Medicine. 38. 96-101.
Gordon, M.D., Dumais, S. (1998). Using latent semantic indexing for literature based discovery. Journal of the American Society for Information Science. 49 (8): 674-685.
Greengrass, E. (1997). Information retrieval: An overview. National Security Agency. TR-R52-02-96.
Guerrero-Bote, V.P., Lopez-Pujalte, C., de Moya-Anegon, F., Herrero-Solana, V. (2003). Comparison of neural models for document clustering. International Journal of Approximate Reasoning, 34 (2-3). 287-305. Nov.
Guha, S., Rastogi, R. and Shim, K. (1998). CURE: An efficient clustering algorithm for large databases. In *Proceedings of the ACM-SIGMOD 1998 International Conference on Management of Data* (SIGMOD'98). 73-84.
Hearst, M. A. (1998). The use of categories and clusters in information access interfaces. In T. Strzalkowski (ed.), Natural Language Information Retrieval. Kluwer Academic Publishers.
Hearst, M. A. (1999). Untangling text data mining. Proceedings of ACL 99, the 37th Annual Meeting of the Association for Computational Linguistics. University of Maryland. June 20-26.
Hodge, V.J., and Austin, J. (2002). Hierarchical word clustering - automatic thesaurus generation. Neurocomputing, 48. 819-846. Oct.
Jackson, J. E. (1991). A users guide to principal components. Wiley, New York, NY.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. Educational and psychological measurement. 20. 141-151.
Kankar, P., Adak, S., Sarkar, A., Murali, K., and Sharma, G. (2002). MedMesh summarizer: Text mining for gene clusters. Proceedings of the Second SIAM International Conference on Data Mining. Robert Grossman, Jiawei Han, Vipin Kumar, Heikki Mannila, and Rajeev Motwani, Editors. April. Arlington, VA.
Karypis, G., Han, E.H., and Kumar, V. (1999). Chameleon: A hierarchical clustering algorithm using dynamic modeling. IEEE Computer: Special Issue on Data Analysis and Mining 32(8). 68-75.
Karypis, G. (2002). CLUTO—A clustering toolkit. http://www.cs.umn.edu/~cluto.
Kendall, M.G., and Lawley, D.N. (1956). The principles of factor-analysis. Journal of the Royal Statistical Society Series A-General, 119 (1). 83-84.
Khare, A. (2003). Connecting word clusters to represent concepts with application to web searching. Knowledge-Based Intelligent Information and Engineering Systems, Pt 1, Proceedings, 2773: 816-823. Lecture Notes in Artificial Intelligence.
Kim, W, and Wilbur, W.J. (2001). Corpus-based statistical screening for content-bearing terms. Journal of the American Society for Information Science and Technology. 52 (3). 247-259. Feb 1.
Ko, Y., Kim, K., and Seo, J. (2003). Topic keyword identification for text summarization using lexical clustering. IEICE Transactions on Information and Systems, E86D (9): 1695-1701 Sep.
Kogan, J., Nicholas, C., and Volkovich, V. (2003). Text mining with information - theoretic clustering. Computing in Science & Engineering, 5 (6). 52-59. Nov-Dec.
Kostoff, R.N., Eberhart, H.J., and Toothman, D.R. (1997). Database Tomography for information retrieval. Journal of Information Science. 23:4. 301-311.
Kostoff, R.N., Green, K.A., Toothman, D.R., and Humenik, J.A. (2000a). Database Tomography applied to an aircraft science and technology investment strategy. Journal of Aircraft. 37:4. 727-730.
Kostoff, R.N., Braun, T., Schubert, A., Toothman, D.R., and Humenik, J.A. (2000b). Fullerene roadmaps using bibliometrics and Database Tomography. Journal of Chemical Information and Computer Science. 40(1): 19-39.
Kostoff, R.N., and DeMarco, R.A. (2001a). Science and technology text mining. Analytical Chemistry. 73:13. 370-378A. 1 July.

Kostoff, R.N., Del Rio, J.A., García, E.O., Ramírez, A.M., Humenik, J.A. (2001b). Citation mining: integrating text mining and bibliometrics for research user profiling. Journal of the American Society for Information Science and Technology. 52:13. 1148-1156.
Kostoff, R.N., Toothman, D.R., Eberhart, H.J., and Humenik, J.A. (2001c). Text mining using database tomography and bibliometrics: A review. Technology Forecasting and Social Change. 68:3. November.
Kostoff, R.N., Tshiteya, R., Pfeil, K.M., and Humenik, J.A. (2002). Electrochemical power source roadmaps using bibliometrics and database tomography. Journal of Power Sources. 110:1. 163-176.
Kostoff, R.N. (2003a). Text mining for global technology watch. In Encyclopedia of Library and Information Science, Second Edition. Drake, M., Ed. Marcel Dekker, Inc. New York, NY. Vol. 4. 2789-2799.
Kostoff, R.N. (2003b). Stimulating innovation. International Handbook of Innovation. Larisa V. Shavinina (ed.). Elsevier Social and Behavioral Sciences, Oxford, UK.
Kostoff, R.N. (2003c). Bilateral asymmetry prediction. Medical Hypotheses. 61:2. 265-266.
Kostoff, R. N. (2003d). The practice and malpractice of stemming. JASIST. 54:10. 984-985. August.
Kostoff, R.N., Shlesinger, M.F., Malpohl, G. (2004a). Fractals roadmaps using bibliometrics and database tomography. Fractals. 12:1. 1-16.
Kostoff, R.N., Shlesinger, M., and Tshiteya, R. (2004b). Nonlinear dynamics roadmaps using bibliometrics and Database Tomography. International Journal of Bifurcation and Chaos. 14:1. 61-92.
Kostoff, R.N., Bedford, C.W., Del Rio, J. A., Cortes, H., and Karypis, G. (2004c). Macromolecule mass spectrometry: Citation mining of user documents. Journal of the American Society for Mass Spectrometry. 15:3. 281-287. March.
Kostoff, R.N., and Shlesinger, M.F. (2004d). CAB-Citation-assisted background. Scientometrics. In Press.
Kuhnhold, M. (2000). The concept of "text mining". Wirtschaftsinformatik, 42 (2). 175-179. Apr.
Lagus, K., Honkela, T., Kaski, S., and Kohonen, T. (1999). Websom for textual data mining. Artificial Intelligence Review. 13(5-6). 345-364. December.
Losiewicz, P., Oard, D., and Kostoff, R.N. (2000).Textual Data Mining to Support Science and Technology Management. Journal of Intelligent Information Systems. 15.

McArdle, J.J. Principles versus Principals of Structural Factor-Analyses. Multivariate Behavioral Research, 25 (1). 81-87. Jan 1990.
Narin, F. (1976). Evaluative bibliometrics: the use of publication and citation analysis in the evaluation of scientific activity (monograph). NSF C-637. National Science Foundation. Contract NSF C-627. NTIS Accession No. PB252339/AS.
Narin, F., Olivastro, D., Stevens, K.A. (1994). Bibliometrics theory, practice and problems. Evaluation Review. 18(1). 65-76.
Perrin, P. and Petry, F.E. (2003). Extraction and representation of contextual information for knowledge discovery in texts. Information Sciences, 151. 125-152. May.
Porter, M.F. (1980). An algorithm for suffix stripping. Program, 14(3). 130-137.
Rasmussen, E. (1992). *Clustering Algorithms*. In W. B. Frakes and R. Baeza-Yates (eds.). Information Retrieval Data Structures and Algorithms, Prentice Hall, N. J.
Salton, G. and Buckley, C. (1998). Term-weighting approaches in automatic text retrieval. Information Processing and Management. 24:5. 513-523.
Schenker, A., Last, M., Bunke, H., and Kandel, A. (2003). Graph representations for web document clustering. Pattern Recognition and Image Analysis, Proceedings, 2652. 935-942.
Schroeder, M. (1991). Fractals, chaos, power laws: Minutes from an infinite paradise. (W.H. Freeman, New York, NY).
Schubert, A., Glanzel, W., Braun, T. (1987). Subject field characteristic citation scores and scales for assessing research performance. Scientometrics. 12 (5-6): 267-291.
Stegmann, J., and Grohmann, G. (2003). Hypothesis generation guided by co-word clustering. Scientometrics, 56 (1). 111-135.
Steinbach, M., Karypis, G., and Kumar, V. A. (2000). Comparison of document clustering techniques. Technical Report #00--034. Department of Computer Science and Engineering. University of Minnesota.
Stensmo, M. (2002). A scalable and efficient probabilistic information retrieval and text mining system. Artificial Neural Networks. ICANN 2002. 2415. 643-648.
Swanson D.R. (1986). Fish oil, Raynaud's Syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine. 30 (1). 7-18. Fall.

Swanson, D.R., Smalheiser, N.R. (1997). An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence. 91 (2). 183-203.
TREC (2003). (Text Retrieval Conference), Home Page, http://trec.nist.gov/.
Trybula, W.J. (1999). Text mining. Annual Review of Information Science and Technology. 34. 385-419.
Viator, J.A., Pestorius, F.M . (2001). Investigating trends in acoustics research from 1970-1999. Journal of the Acoustical Society of America. 109 (5): 1779-1783 Part 1.
Visa, A. (2001). Technology of text mining. Machine Learning and Data Mining in Pattern Recognition. 2123. 1-11.
Wang, B.B., McKay, R.I., Abbass, H.A., and Barlow, M. (2003). A comparative study for domain ontology guided feature extraction. Twenty-Fifth Australian Computer Science Conference [ACSC2003], Adelaide, Australia, Vol. 16, Michael Oudshoorn, ed.
Weeber M., Klein H., Aronson A.R,, Mork J.G, de Jong-van den Berg, L.T.W, and Vos R. (2000). Text-based discovery in biomedicine: The architecture of the DAD-system. Journal of the American Medical Informatics Association. 903-907 Suppl. S.
Weiss, S.M., Apte, C., Damerau, F.J., Johnson, D.E., Oles, F.J., Goetz, T., and Hampp, T. (1999). Maximizing text-mining performance. IEEE Intelligent Systems & Their Applications. 14:4. 63-69. July-August.
Wilbur W. J. and Sirotkin K. (1992). The automatic identification of stop words. Journal of Information Science. 18 (1). 45-55.
Wilbur W.J. and Yang Y.M. (1996). 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). 209-222. May.
Willet, P. (1988). Recent trends in hierarchical document clustering: A critical review. *Information Processing and Management.* 24:577-597.
Yalcin, I. and Amemiya, Y. (2001). Nonlinear factor analysis as a statistical method. Statistical Science, 16 (3). 275-294. Aug.
Zamir, O. and Etzioni, O. (1998). Web document clustering: A feasibility demonstration. In: *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR'98). 46-54.
Zhu, D.H. and Porter, A.L. (2002). Automated extraction and visualization of information for technological intelligence and forecasting. Technological Forecasting and Social Change. 69:5. 495-506. June.

(*The views in this paper are solely those of the authors and do not represent the views of the Department of the Navy or its components, or the views of Rush Medical College. We acknowledge the valuable contributions of Dr. Wendy Martinez (ONR) to the statistical portions of the paper*).

APPENDIX 1 – SCREE PLOT INTERPRETATION

A parametric analysis of factor number determination as a function of Scree Plot scale was performed, to test whether factor number determination was scale dependent. The results follow.

1) Demonstration of Fractality

Factor matrices with different numbers of factors specified were computed. Eigenvalues were generated by Principal Components Analysis, and these eigenvalues represented the variance accounted for by each underlying factor. Figure 1 shows the factor eigenvalue-factor number plot for the 659 un-rotated factors on a linear scale. The 'elbow', or break point, of the curve appears to be about fourteen factors. To improve resolution, the curve was stretched in the x direction by halving the number of factors shown on one page. The curve had a similar shape to the 659 factor case, but the factor termination point appeared to decrease. The halving process was repeated until ten factors were plotted on one page, and the resolution effectively increased by an order of magnitude overall.

Figure 3 shows the ten factor plot. The elbow of the curve appears to be about two factors. Thus, the number of factors selected based on significant slope change decreased from fourteen in the 659 factor plot to two in the ten factor plot.

INSERT FIGURE 3

In fractal analysis, a fractal object has a number of characteristics (Schroeder, 1991). Among these are self-similarity (similar to itself at different magnifications), and adherence to a scaling relationship (the

measured value of a property will depend on the resolution used to make the measurement). The Scree Plot appears to have these two fractal properties. As the resolution increases, more structure appears, and the value of the break point changes.

The simplest and most common form of the scaling relationship is that of a power law. When such a power law is plotted on a log-log scale, the scaling relationship appears as a straight line. Figure 4 is a plot of the break point on a linear scale, and Figure 5 is a re-plot of Figure 4 on a log-log scale. The log-log plot is approximately linear, reflects power law scaling, and validates the break point selection as a fractal process. This observation about the fractal-like nature of the Scree Plot analysis process does not appear to have been reported in the literature previously. Further, this sensitivity of factor number to scale confirms the reviewer's concerns about the precision of the Scree Plot approach to determining number of factors to be selected.

INSERT FIGURE 4

INSERT FIGURE 5

The reviewer also raised the concern about the impact of the inconsistency among analysts in interpreting the Scree Plot on the final taxonomy. Essentially, this issue concerns the consequences of selecting a few more or less factors to run.

The considerations involved in selecting the number of factors to run are similar to those involved in selecting the number of clusters for the cluster analysis. Basically, the issue revolves around the level of resolution desired. More clusters, or more factors, provide more resolution, detail, and information. In addition, more factors capture a greater fraction of total variance. The cost of more factors or clusters is more computer running time, and especially more time for interpretation of results. Thus, the number of clusters and/ or factors to be selected depends on the objectives of the study.

As an example of the need for resolution, the first author has performed text mining of both homogeneous databases and heterogeneous databases in the past few years. The homogeneous databases derive from monodiscipline

studies, where the topical material is closely related (e.g., the anthrax database). The heterogeneous databases derive from multi-discipline databases, where the topical material can be very disparate (e.g., China's research output).

The single discipline studies require relatively few clusters to capture the main themes of the database, because the topical variation is modest, and relatively few hierarchical taxonomy levels are required for the same reason. The multi-discipline studies require many clusters to resolve the disparate themes, and many taxonomy levels may be required to portray the structure accurately.

For example, suppose it were desired to examine the research outputs of a major country whose research budget was one billion dollars per year, whose research program consisted of 100 different disciplines, and whose research output database consisted of 10000 records. If two clusters were used for the analysis, only one hierarchical taxonomy level would be possible, and each cluster (on average) would cover 500 million dollars per year worth of research, fifty technical disciplines, and 5000 records. Obviously, the results and corresponding insights would be very generic and aggregated, and probably of very limited utility. If, however, 1000 clusters were used, then ten hierarchical taxonomy levels would be possible, and each elemental cluster (on average) would cover one million dollars worth of research, 1/10 of a technical discipline, and ten records. Much more detailed understanding of the country's research would be obtained, down to approximately the individual program level. Obviously, much more work would be required to generate useful information from the 1000 cluster case than the two cluster case.

In the present study, the issue of factor number variations was examined by the authors in the initial phase of the study by performing a parametric study of number of factors verses taxonomy structure. Factor matrices ranging from two factors in size to fourteen factors were run, and the results analyzed. The main themes did not change, and the context-dependent trivial words identified remained essentially the same. The main changes were the number of hierarchical levels that could be generated, and the resolution afforded by each factor.

For the two factor and fourteen factor taxonomies, the main themes are still autoimmunity and circulation. The fourteen factor taxonomy allows more

structural detail to be shown, as displayed in the Results section. More levels in the taxonomy could have been generated with the fourteen factor taxonomy, if desired, and more sub-themes could have been generated with more detailed specificity. Obviously, if there had been three main themes instead of two for this topical area, then the two factor case would have missed one of the themes. In practice, the analyst should always perform some type of sensitivity study on number of factors, to insure that no main themes are being missed with the final number of factors chosen.

APPENDIX 2 – FACTOR MATRIX FILTERING

The purpose of this Appendix is to show how two word clusters, generated using the WINSTAT multi-link hierarchical aggregation technique, can be compared. The comparison method is a combination of quantitative and qualitative approaches.

This Appendix contains two sections. The first section shows the impact on the 'distance' metric, and on the assignment of words to clusters, of substituting the most trivial words for higher technical content words in the selection of cluster input words. The second section shows the impact on the 'distance' metric, and on the assignment of words to clusters, of substituting non-factor matrix-selected words by factor matrix-selected words in the selection of cluster input words.

Impact of Trivial Words on Cluster Quality

The impact was shown by conducting a simple experiment where trivial words were substituted for higher technical content words, and the change in cluster quality evaluated. The same 930 record database that was used in the main text of the study was used in this Appendix, for consistency. The words contained in the Abstract were extracted using the TechOasis software package, subject to the standard StopWord list filtering. This StopWord list contains words that are trivial in almost every context (e.g., the, if, or, and, etc).

In the first case, the 252 highest frequency words extracted from the Abstract were used as input for the WINSTAT clustering process. In the second case, the two highest frequency trivial words on the StopWord list (of, the) were substituted for the two lowest frequency words of the 252 words used for the first case (abnormalities, C). While these two words are

not of the highest technical content, they are certainly of higher content than ‘of’, ‘the’.

Clusters were run using the 8, 16, 32, 64, 128, and 253 highest frequency words, respectively. For display purposes, the 64 word dendrograms were selected. Figure 6 shows the non-trivial word dendrogram, while Figure 7 shows the dendrogram that includes the two trivial words. The overall ‘distance’ metric value associated with Figure 6 is 132.56, while the overall ‘distance’ metric value associated with Figure 7 is 139.54. For the 252 word cases, the respective overall ‘distance’ metric values are 513.81 and 596.32. As expected, substitution of the trivial words ‘of’, ‘the’ for the higher technical content words ‘abnormalities’, ‘C’ results in an increase in the value of the ‘distance’ metric. As the constituent clusters become more diffuse with the addition of trivial words, the ‘distance’, a measure of cluster cohesiveness, increases.

INSERT FIGURE 6

INSERT FIGURE 7

Of equal importance is the impact on assignment of words to clusters. The cluster thematic differences are most pronounced at the highest taxonomy levels, and these differences become more subtle as the lowest taxonomy levels are accessed. As shown in the text, and known from medical experience, the first taxonomy level of the Raynaud’s Phenomenon literature has the two major categories of auto-immunity and circulation. The literatures, and key phrases, associated with each category tend to be distinct. How did the substitution of trivial words for higher content words impact the assignment of words to the highest level categories in the taxonomy?

On Figure 6, there are two obvious first level clusters. One ranges from the words Raynaud’s to Diagnosis, and the other ranges from the words Blood to Disorders. The Raynaud’s-Diagnosis cluster focuses on Auto-immunity, while the Blood-Diagnosis cluster focuses on Circulation. The words in the clusters range from very specific (e.g., lupus, scleroderma, vascular, arterial) to quite general (e.g., test, years, cases, severe). The presence of the general words is a consequence of not using any filtering (manual, factor matrix, etc) for the demonstration in this section, other than the StopWord list and the frequency cut-off.

On Figure 6, the first level Auto-immunity category contains 31 words (48%), and the first level Circulation category contains 33 words (52%). The key auto-immunity single words and obvious word combinations (e.g., systemic sclerosis, lupus erythematosus) are in the Auto-immunity category, and the key circulation single words and obvious word combinations (e.g., blood flow, finger temperature) are in the Circulation category.

On Figure 7, there is a major change in the taxonomy structure. The first level of the taxonomy splits into two categories. One is a very small category, ranging from Of to Clinical (in abbreviated form, Of-Clinical), and the other is a much larger category (Systemic-Disorders). The small category is a very generic overview of the database. The larger category combines the Auto-immunity and Circulation categories.

The larger category splits into two second level sub-categories. One has an Auto-immunity focus (Systemic-3), and the other has a Circulation focus (Blood-Disorders). The Auto-immunity sub-category contains 33 words (59%), while the Circulation sub-category contains 23 words (41%). Thus, for the major technical sub-division between Auto-immunity and Circulation, Auto-immunity has gone from 48% of the words to 59% when the trivial words are substituted, while Circulation has decreased from 52% to 41%, a noticeable change.

What are the words that switched categories? Most were generic, concentrated in a small Circulation cluster near the far end of the dendrogram on Figure 6 (Group-3). The other two were more specific (peripheral, vascular). They switched from the Circulation category on Figure 6 to the Auto-immunity category on Figure 7. While it could be argued that the generic terms that switched categories were only weakly linked thematically to the Circulation category initially, the switch of the more specific terms (peripheral, vascular) is less defendable.

There are 52 records in the 930 Raynaud's record database that contain the words peripheral and vascular. A sampling of the 52 records shows that the combination of peripheral and vascular is always used in the context of circulation. For primary Raynaud's, there is only the association with circulation. For secondary Raynaud's, where circulatory problems may be a symptomatic spin-off of the auto-immune disease, such articles might include the underlying auto-immune disease in conjunction with the

circulatory problem. However, the main association of the peripheral-vascular combination is with circulation problems. Thus, in addition to changing the first level structure of the taxonomy, an additional thematic effect of substituting the two trivial words into the clustering input has been to re-assign the combination peripheral and vascular from the more appropriate technical category to the less appropriate category.

The number of trivial words used in this substitution experiment was small (two). As the number of words used to form the clusters increases, the effect of the two substituted trivial words on the results is expected to decrease. In the 252 word case that includes the substituted trivial words, the combination peripheral and vascular reverts to the Circulation category in the clusters with the substituted trivial words. However, the separate generic first level category containing the substituted trivial words and the more generic words remains.

Additionally, in the 252 word case that includes the substituted trivial words, the combination vascular and peripheral reverts to the appropriate high level Circulation category. However, the individual words vascular and peripheral are in different lower level categories relative to the 252 word case that does not include the substituted trivial words. In the latter case, peripheral and vascular are adjacent. Thus, the simple substitution of two trivial words has changed the taxonomy structure at the highest level, and more pervasively at the lowest levels.

Impact of Factor Matrix-Selected Words on Cluster Quality

The impact was shown by comparing the word dendrogram from the main text (whose input words were obtained with factor matrix filtering) with a word dendrogram whose input words were not obtained with factor matrix filtering. The comparison results will depend strongly on the method used for word selection in the latter case. To minimize human intervention, and the potential for arbitrary bias, the highest frequency words after StopWord list filtering were selected for clustering, including conflation of closely related words.

The 252 word dendrograms are shown on Figures 2 and 8. Figure 2 incorporates factor matrix filtering, while Figure 8 does not. The overall distance metric associated with Figure 2 is 513.06, while the overall distance metric associated with Figure 8 is 527.28. For the 64 word cases, the

respective ‘distance’ metric values are 134.08 and 140.5. As expected, substitution of factor matrix filtering for non-factor matrix filtering results in a decrease in the value of the ‘distance’ metric. As the constituent clusters become sharper with the removal of context-dependent trivial words from factor matrix filtering, the ‘distance’, a measure of cluster cohesiveness, decreases.

INSERT FIGURE 8

As in the previous example, the assignment of words to clusters is of equal importance to the change is particular metrics. How did the addition of factor matrix filtering impact the assignment of words to the highest level categories in the taxonomy?

To re-iterate the structure of Figure 2, there are the two first level categories of Auto-immunity (Patients-Old) and Circulation (Treatment-Acid). Auto-immunity can be subdivided into its second level categories of Auto-immune Diseases/ Antibodies (Patients-Neuropathy) and Inflammation (Lesions-Old), and Circulation can be sub-divided into its second level categories of peripheral vascular circulation (Treatment-Consecutive) and coronary vascular circulation (Pulmonary-Acid). The categories are sharp, correspond to medical diagnosis and treatment, and the contents are appropriately placed.

Figure 8 contains a different higher category level structure. The first level can be divided into two categories, a very generic small category (13 words) centered around Raynaud’s Phenomenon and scleroderma (Raynaud’s-Symptoms), and the other being large (239 words) and including both Auto-immunity and Circulation (Systemic-Abnormality). The small generic first level category can be subdivided into its second level categories centered around Raynaud’s Phenomenon/ Scleroderma (Raynaud’s-Severity) and very generic terms (One-Symptoms). The larger second level category can be sub-divided into its second level categories of Auto-immunity (Systemic-Signs) and Circulation (Treatment-Abnormality).

Not only is the structure of the first level different between the two figures, but, for example, the word ‘scleroderma’ is in different first level categories. In the factor matrix filtered case (Figure 2), scleroderma is in the Auto-immunity category, closely linked to both relatively generic terms (patients, diseases) and a very specific term (progressive systemic sclerosis, PSS).

This is due to 1) PSS being a pseudonym for scleroderma and 2) scleroderma being a sign of a group of diseases that involve the abnormal growth of connective tissue, which supports the skin and internal organs, and therefore used as a more generic umbrella term for these disorders. In the non-factor matrix filtered case (Figure 8), scleroderma is in the generic first level category, coupled to the generic terms 'patients, diseases', as in Figure 2, but de-coupled from its pseudonym PSS.

A third level sub-division is required for Figure 8 for comparison with the second level sub-division of Figure 2. The second level Auto-immunity category of Figure 8 (Systemic-Signs) sub-divides into third level categories of Auto-immune Diseases/ Antibodies (Systemic-Comparison) and Inflammation (Lesions-Signs). The Inflammation category contains 13 words, of which perhaps 3 relate specifically to inflammation (lesions, vasculitis, inflammatory). Contrast this with the second level Inflammation category from Figure 2. This Inflammation category contains 16 words (similar in magnitude), of which 8 relate specifically to inflammation (lesions, corticosteroids, eosinophilia, fasciitis, hypergammaglobulinemia, scleroderma-like, inflammatory, polyarthritis). The difference in level of detail is striking.

The second level Circulation category of Figure 8 (Treatment-Abnormality) sub-divided into third level categories of Circulation (Treatment-Data) and a very generic unfocused category (Criteria-Abnormality). In the third level Circulation category, peripheral vascular circulation is intermingled with coronary artery circulation, and lower taxonomy levels need to be accessed before these circulation sub-categories can be differentiated.

On Figure 8, the circulation sub-category of Coronary Artery Circulation (Combined-Five) contains 12 terms, of which 3 are relatively specific (hypertensive, cardiac, heart). On Figure 2, the second level category of Coronary Artery Circulation (Pulmonary-Acid) contains 25 terms, of which about 20 are relatively specific (pulmonary, fibrosis, hypertension, cardiac, cardiovascular, heart, coronary, myocardial, ventricular, angina, necrosis, spasm, chest, lung, biofeedback, training, relaxation, stress, migraine, headaches). The biofeedback thrust was not even mentioned on Figure 8. Additionally, Figure 2 provides much more detail about the physical consequences of insufficient coronary artery circulation, whereas Figure 8 alludes to the general area with no specific detail.

There are many other differences in structure and content between the two dendrograms. In summary, the factor matrix filtering provides a lower value of overall ‘distance’ (translating into a more sharply defined overall taxonomy. The clusters are improved from a medical perspective, and the contents are far more detailed.

APPENDIX 3 – THE PRACTICE AND MALPRACTICE OF STEMMING

BACKGROUND

Stemming algorithms remove the common morphological and inflexional endings from words, converting word variants to a common form (which may or may not be a real word). The process of stemming is important to the operation of classifiers and index builders/searchers because it makes the operations less dependant on particular forms of words and therefore reduces the potential size of vocabularies which might otherwise have to contain all possible forms.

Stemming reduces the dimensionality of a system, but if the combination process merges words that have very different meanings, errors will result. The severity of the errors depends on the application context. Almost all authors who discuss the consequences of errors provide the examples of the close singular/ plural variant conflation as being essentially error-free, and more diverse variants as being more problemmatical (e.g., relative/ relativistic).

Many stemming algorithms have been generated, probably the most widely used of which is the Porter algorithm (A3-1). Almost every stemming algorithm, including Porter's, consists of conversion rules that are independent of specific corpuses, contexts, and applications. Croft and Xu (A3-2), and Xu and Croft (A3-3) designed a corpus-based filter to select word variants for stemming. Their "basic hypothesis is that the word forms that should be conflated for a given corpus will co-occur in documents from that corpus." (A3-3).

The purpose of this appendix is to show that word stemming can be strongly context and application dependent, and that selection of word variants for stemming should be context/ application dependent. In addition, this appendix will show that the conflation filter rule proposed in (A3-3) does not have a strong rational basis.

ANALYSIS

The first author has used stemming for information retrieval and text mining (A3-4, A3-5), mainly phrase clustering in text mining. A simple experiment was run, as part of a larger text mining study on the Fractals literature, to test the effect of word stemming on cluster theme definition. A Fractals-based query retrieved 4389 Science Citation Index records containing Abstracts, covering the period 2001-October 2002. All the single Abstract words were extracted, and the highest frequency highest technical content words (820) were selected for phrase clustering. A two step clustering process was used, where a factor matrix was generated initially with no word combination required, then a hierarchical clustering was performed using word combinations based on the factor matrix results.

The factor matrix generator in the TechOasis software package used a correlation matrix of the uncombined 820 words as input. The generator produced a 29 factor matrix (820 x 29), where each factor represented a theme of the Fractals database. The value of each matrix element Mij was the factor loading, the contribution of word i to factor j.

For the analysis of each factor, the factor column was sorted in descending numerical order. Each factor had two tails, one with large positive value and one with large negative value. The tails were not of the same absolute value size; one of the tails was always dominant. The theme of each factor was determined by the highest absolute value terms in the dominant tail.

For purposes of this appendix, the interchangeability of the singular and plural variants only will be reported and discussed, although the results of interchangeability of all the word variants in the 820 word list were used to determine the word combinations input to the hierarchical clustering algorithm. All words that had both singular and plural forms represented in the 820 words were examined, especially where at least one of the variants was contained in the dominant tail of a factor and thereby was influential in determining the theme of the factor. Singular and plural forms that could be conflated credibly should be interchangeable. They should be located in close proximity in the dominant tail (similar factor loadings), and should have similar influence in determining the cluster theme. Otherwise, they are being used in different contexts, and their conflation has the effect of artificially merging themes or clusters to produce erroneous groupings.

One benchmark for how well the factor matrix algorithm spots interchangeability is its numerical performance with multi-word phrases. In

the Fractals literature, there are multi-word phrases that appear frequently, where each word in the multi-word phrase is either exclusive to the phrase, or used frequently in the phrase. Examples are: Atomic Force Microscopy and its acronym AFM, Scanning Electron Microscopy and its acronym SEM, Thin Film, Fractional Brownian Motion and its acronym FBM, and Monte Carlo. The component words of these strong multi-word phrases should appear close to each other in the dominant tail, if the clustering is viewing them as a unit. The dominant factor tails that include the multi-word phrases above, and the word factor loadings (in parenthesis) are as follows.

Factor 6: microscopy (-.59), atomic (-.58), AFM (-.58), force (-.52); scanning (-.47), microscope (-.44), electron (-.40), SEM (-.34); film (-.34), thin (-.31)

Factor 8: Brownian (-.68), fractional (-.64), motion (-.62), FBM (-.50)

Factor 3: Monte (-.47), Carlo (-.46)

The threshold absolute value for high factor loading across all factors was about .20, and the highest absolute value for factor loading across all factors was about .70. All the words above were well above the threshold and at or near the end of the dominant tail in their respective factor. All the multi-word phrase components had high factor loadings in close proximity, with words relatively unique to the multi-word phrase being in very close proximity.

Now the performance of singular and plural variants will be examined. There was a continuum of relative values between the singular and plural variants, and only the extremes will be shown to illustrate the main points. Singular/ plural variants had a high absolute value factor loading in one factor only, and that value will be displayed. Low value factor loadings do not determine the factor theme, and will not be shown. However, it was clear that variants closely related in their dominant tail appearance also tended to be closely related in most of their appearances in other factors, and variants not closely related in their dominant tail appearance tended not to be closely related in appearances in other factors.

Sample closely-related singular-plural variants, accompanied by their factor loadings/ factors in parenthesis, are as follows: avalanche (.453/10),

avalanches (.502/10); earthquake (.599/17), earthquakes (.541/17); gel (.539/18), gels (.495/18); island (.42/24), islands (.38/24); network (.49/21), networks (.45/21).

Sample disparately-related singular-plural variants include: angle (.31/23), angles (.08/23); control (-.25/21), controls (-.01/21); electron (-.40/6), electrons (-.02/6), force (-.52/6), forces (.01/6), state (-.26/10), states (-.01/10).

Thus, the closely-related singular-plural variants had similar high factor loadings, and could be conflated with minimal impact on the clustering results, since they are acting interchangeably in the clustering context. The disparately-related singular-plural variants had one high and one low factor loading, and could not be justifiably conflated, since they are operationally different concepts with similar superficial appearance.

It should be strongly emphasized that the metric used for conflation justification was interchangeability, not co-occurrence of the variants in the same document, as proposed by (A3-3). While intra-document co-occurrence may be operable under some scenarios, there is no a priori reason that it should be stated as a condition, metric, or requirement. One could easily envision a corpus where singular-plural variants never co-occur in the same document, yet behave interchangeably (or don't behave interchangeably). For example, a corpus of small documents, such as Titles or Abstracts, might not contain word variants in the same document, but could have word variants behaving interchangeably even though they are in different documents. The condition to require is that the variants should correlate or co-occur similarly with other words in the corpus for the purpose of the application context. Thus, their variant is transparent from the perspective of the other words in the specific context of the application. Reference (A3-3) would have had a much more credible condition had the metric been co-occurrence similarity of each word variant with other (non-variant) words in the text, rather than high co-occurrence with other forms of the variant.

Once the conflation-justified variants were identified by the factor matrix filter, they were then combined to lower the dimensionality of the system, and used to generate a co-occurrence matrix. This 250 word square matrix was imported into an Excel statistical package add-in named WINSTAT

(Excel has an approximate 250 column limitation), and used as the basis for a multi-link clustering algorithm.

In summary, credible conflation is context and application sensitive. The metric for determining conflation credibility should be driven by the context and application. For the clustering application described in this letter, correlation-driven interchangeability is the appropriate metric, rather than the variant co-occurrence-based metric proposed in (A3-3).

REFERENCES

A3-1. Porter,M.F. "An algorithm for suffix stripping". Program. 14(3). Pages 130-137. July 19 1980.

A3-2. Croft, W. B. and Xu, J. *Corpus-specific stemming using word form co-occurence.* In Proceedings for the Fourth Annual Symposium on Document Analysis and Information Retrieval, Pages 147—159. Las Vegas, Nevada. April 1995.

A3-3. Xu, J. & Croft, W.B. Corpus-based stemming using coocurrence of word variants. *ACM Transactions on Information Systems*. 16(1). Pages 61-81.

A3-4. Kostoff, R. N., and DeMarco, R. A. Science and Technology Text Mining. Analytical Chemistry. 73:13. Pages 370-378A. 1 July 2001.

A3-5. Kostoff, R. N., Tshiteya, R., Pfeil, K. M., and Humenik, J. A. Electrochemical Power Source Roadmaps using Bibliometrics and Database Tomography. Journal of Power Sources. 110:1. Pages 163-176. 2002.

Figure 1

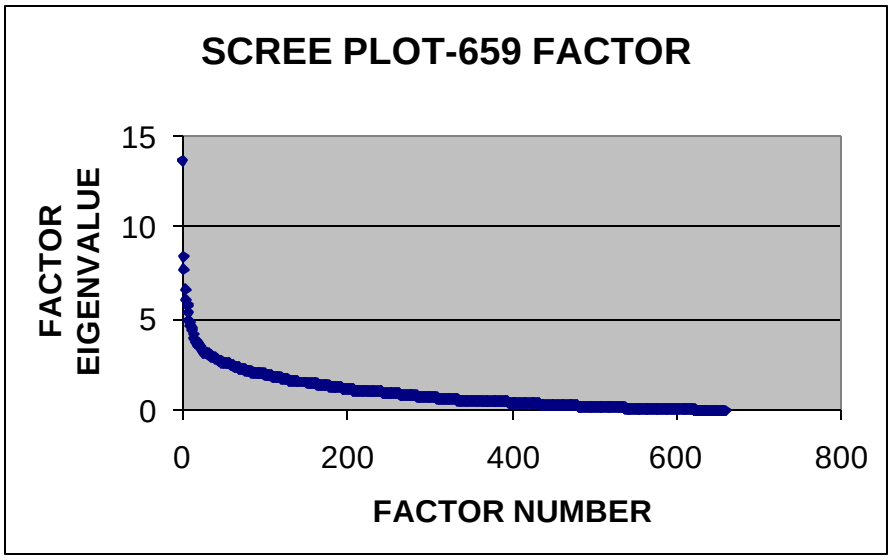
SCREE PLOT-659 FACTOR
FACTOR EIGENVALUE
15
10
5
0
0
200
400
600
800
FACTOR NUMBER

Figure 2

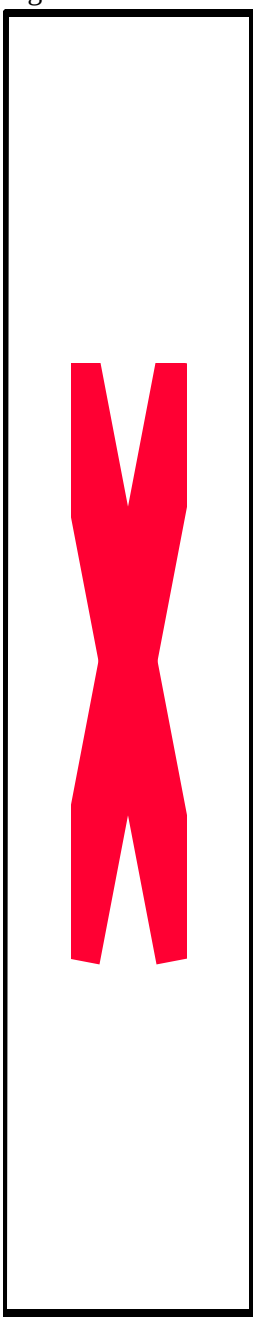

Figure 3

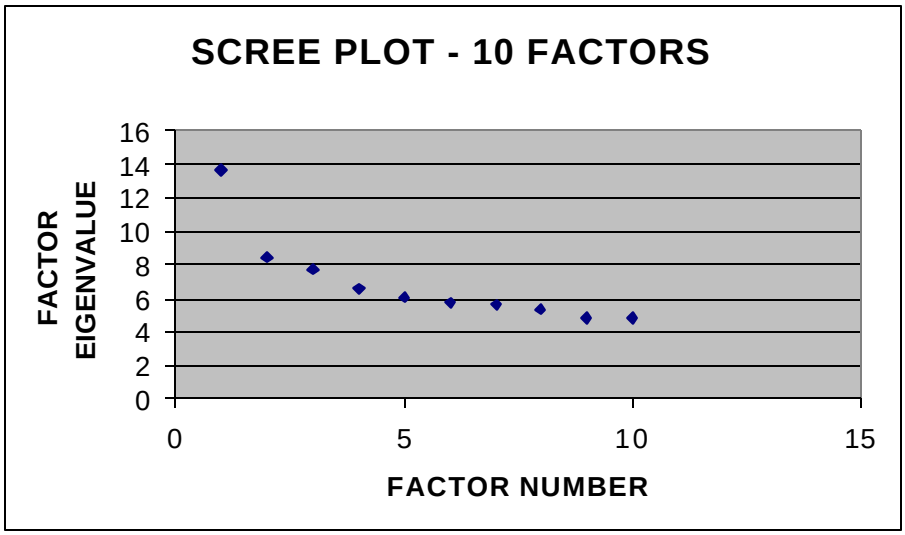
SCREE PLOT - 10 FACTORS
FACTOR EIGENVALUE
16
14
12
10
8
6
4
2
0
0
5
10
15
FACTOR NUMBER

Figure 4

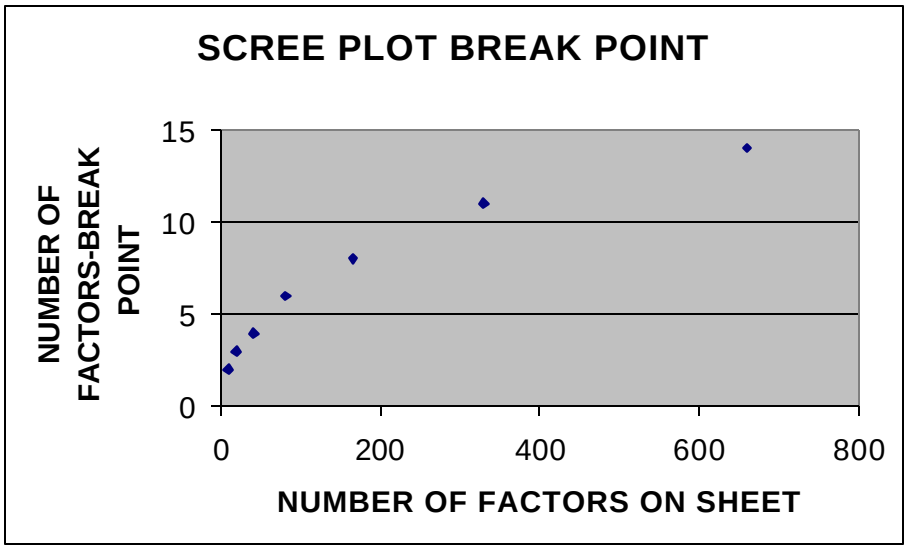
SCREE PLOT BREAK POINT
NUMBER OF FACTORS-BREAK POINT
15
10
5
0
0
200
400
600
800
NUMBER OF FACTORS ON SHEET

Figure 5

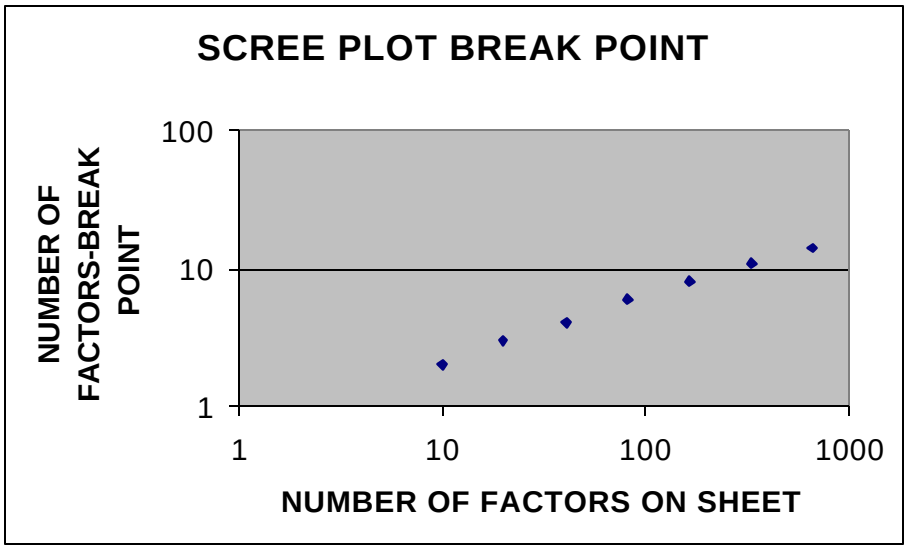
SCREE PLOT BREAK POINT
NUMBER OF FACTORS-BREAK POINT
100
10
1
1
10
100
1000
NUMBER OF FACTORS ON SHEET

Figure 6

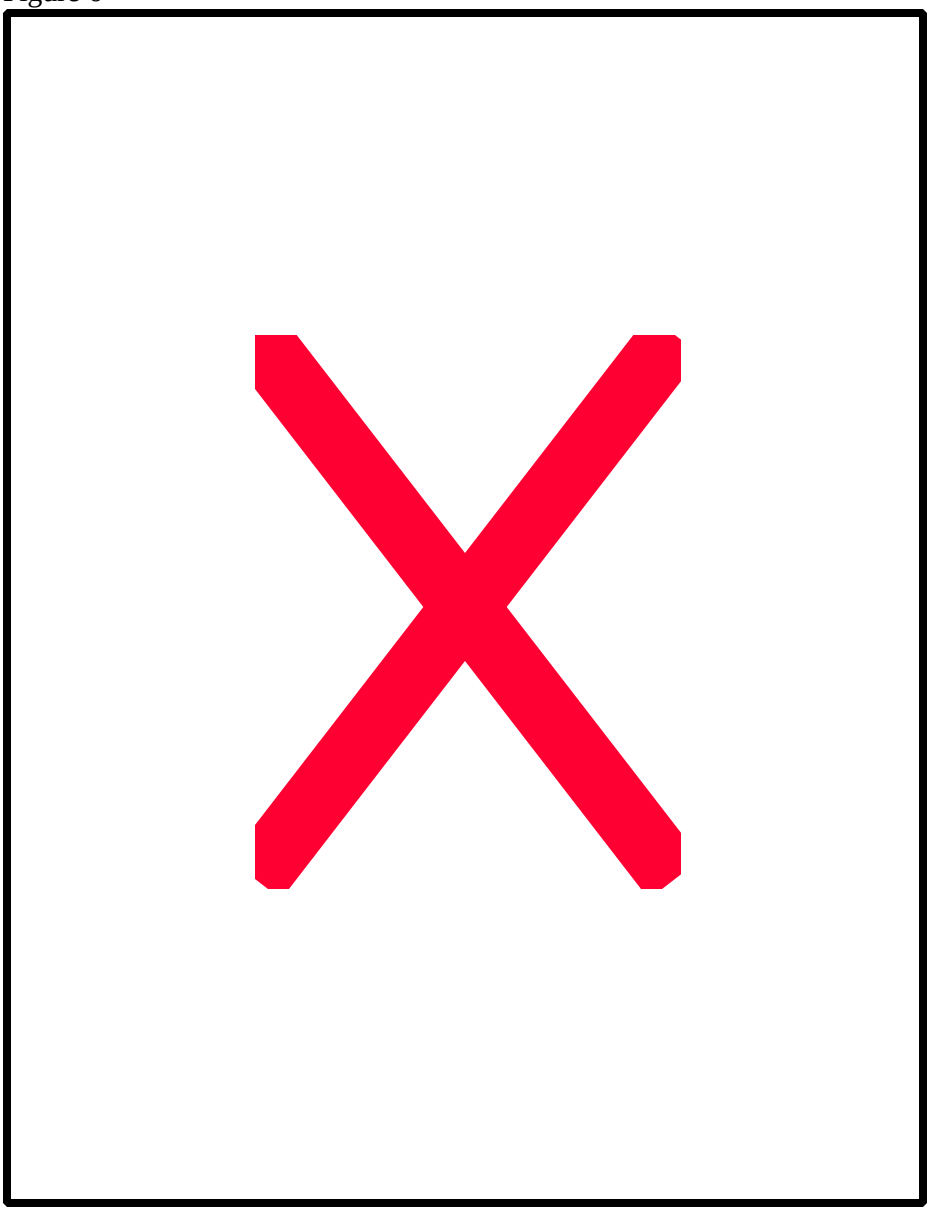

Figure 7

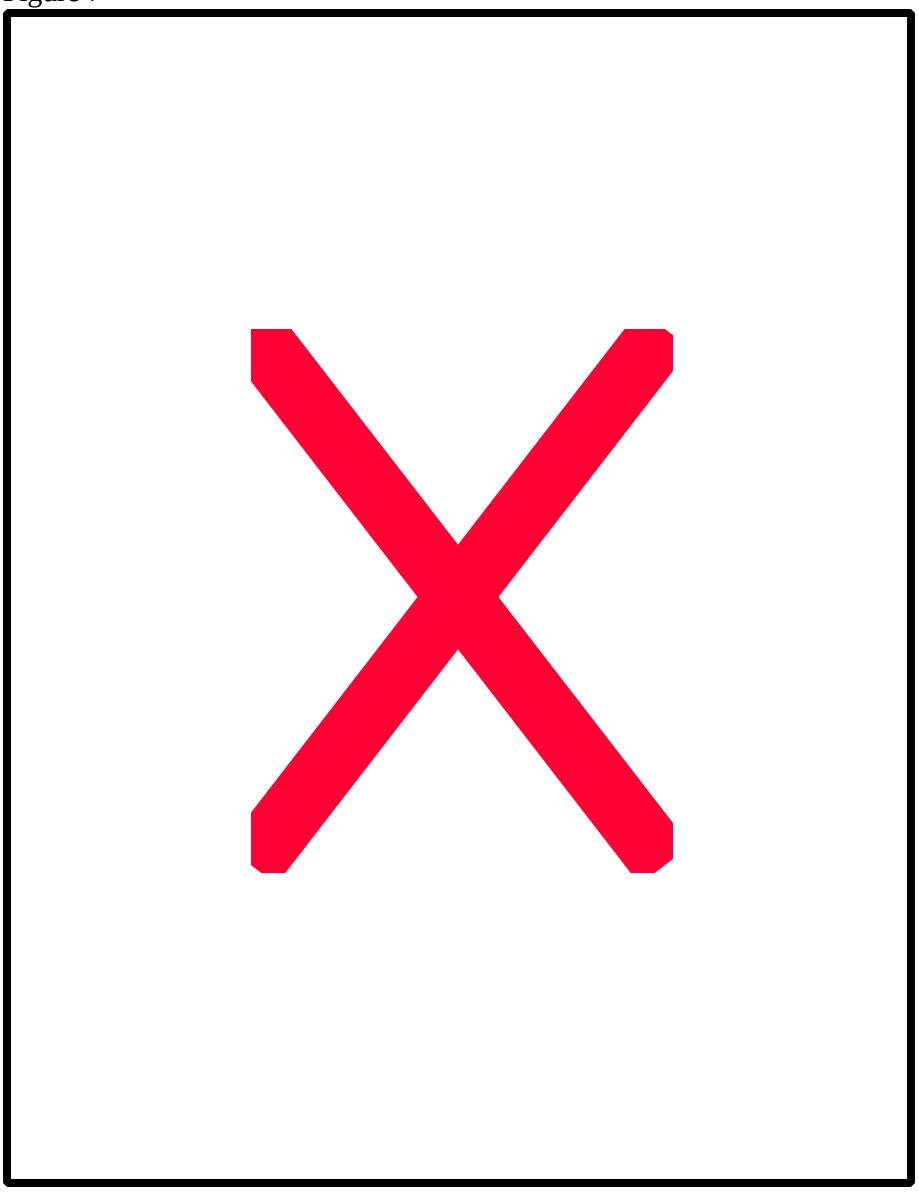

Figure 8