Factor Analysis Optimization: Applied on Natural Language Knowledge Discovery

Robert J. Watts[1], Alan L. Porter, Ph.D.[2], Donghua Zhu, Ph.D.[3]

[1] U.S. Army Tank-automotive and Armaments Command, Advanced Vehicle Technologies, Warren, Michigan

[2] Search Technology, Inc.,Technology Policy & Assessment Center, Georgia Tech. Georgia

[3]School of Management & Economics, Beijing Institute of Technology, Beijing, China

Abstract. The Technology Opportunities Analysis of Scientific Information System (*Tech OASIS*) automates the identification and visualization of relationships inherent in sets (i.e., hundreds or thousands) of literature abstracts. An automated *Tech OASIS* algorithm applies principal components analysis (PCA), multi-dimensional scaling (MDS) and a path-erasing algorithm to elicit and display clusters of related concepts. However, cluster groupings and visual representations are not singular for the same set of literature abstracts (i.e., user selection of the items to be clustered and the number of factors to be considered will generate alternative cluster solutions and relationships displays). Our current research, herein documented, seeks to identify and automate selection of a "best" PCA factor analysis solution for a set of literature abstracts. How then can a "best" solution be identified? Research on quality measures of factor/cluster groups indicates that terms/factors selections based on entropy, F-measure and cohesiveness appear promising. Our developed approach applies a composite metric, which strives to minimize the factor grouping entropy and F-measure and maximize each group's cohesiveness, while also considering set coverage. We apply the detailed approach to automatically map conceptual (term) relationships for 1202 abstracts concerning "natural language knowledge discovery."

1. Introduction

Tech OASIS applies "cluster analysis" through the use of principal components analysis (PCA). PCA derives the relatedness of terms and/or phrases in a document set under study to create cluster groups. When creating a *Tech OASIS* map of factor groups, the analyst must select the terms and/or phrases to be clustered and the number of factor groups to be extracted. The guidance provided to the analyst is to "select a reasonable number of the high frequency terms, about 200 or so, and use the default number of factors, which equals the square root of the number of terms selected for the cluster analysis." The analyst's selection of the items to be clustered and the number of factors to be considered will generate alternative cluster solutions. This research strives to aid those analyzing abstract sets by developing an automated algorithm to generate a "best" (i.e. standard) set of term clusters.

Cluster analysis strives to create "highly internally homogenous groups, the members of which are similar to one another, and highly externally heterogeneous groups, members of which are dissimilar to those of other groups" [1].[1] Steinbach, et al., discuss and apply measures of cluster quality, both internal and external measures of "goodness" [2]. Internal measures assess sets of clusters without knowledge of external cluster relationships. External quality measures compare cluster groups to known classes, which we extend to mean other cluster groups. Whether classifying records or portions of text within a free-text document, the clustering goal has remained creating homogenous groups that are heterogeneous in respect to other classified groups. Salton et al (1982) discuss "a clustering process that is designed to identify groups of text excerpts that are closely related to each other, but also relatively disconnected from the rest of the text" [3].

[1] During PCA, there is no "analytical intent" to create dissimilar factor groups (i.e., derived factor groups can be entire or partial sub-sets of other derived factor groups). Only internal group relatedness is analytically derived by PCA. By creating highly internally related factor groups, external group relatedness of the derived PCA groups will tend be minimized

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
29 SEP 2002	**N/A**	

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Factor Analysis Optimization: Applied on Natural Language Knowledge Discovery	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
; ; Watts /Robert,JPorter /Alan,LZhu /Donghua	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
US Army RDECOM-TARDEC 6501 E 11 Mile Rd Warren, MI 48397-5000	**17076**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
TACOM TARDEC	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release, distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

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**	**SAR**	**13**	

This quotation represents a composite assessment derived from several other literature sources [4]. Clearly, the clustering research goal has been to define internally homogenous and externally heterogeneous groups of information.

We will discuss internal and external cluster group quality measures, then present an approach on how to use them to select a "best" PCA factor group. First, however, we introduce the concepts of technology opportunities analysis (TOA) and describe the software tool, *Tech OASIS*, used in this research. We then provide a more detailed explanation of the factor map analytical process that we strive to optimize. The cluster quality measures are developed, then applied to a sample analysis on "natural language knowledge discovery."

2. *Tech OASIS* and Technology Opportunities Analysis

The *Tech OASIS* is software that enables text mining of fixed field literature abstract files.[2] *Tech OASIS*, named "*VantagePoint*" for the commercial market, has been developed as an *MS Windows*-based software suite of tools that combines bibliometrics with content analysis [8]. *Tech OASIS* development reflects a collaboration between Search Technology, Inc., as the prime contractor, and sub-contractors, Georgia Tech Technology Policy and Assessment Center (TPAC) and Intelligent Information Services Corporation (IISC).

Tech OASIS supports the performance of technology assessments by automating the profiling of open-source R&D. It has been used to facilitate the process of "innovation forecasting," which applies bibliometric analyses to enhance traditional technology forecasting techniques [5]. The technology opportunities analysis (TOA) concept originated at Georgia Tech's TPAC. TPAC studies technological innovation processes [6] [7]. The TOA process entails these main steps:

1) *Search and retrieve* text information, typically from large abstract databases on a particular subject. In this paper, we analyze abstracts of research related to "natural language knowledge discovery."
2) *Profile* the resulting search set. *Tech OASIS* applies a combination of machine learning, statistical analyses enhanced by computational linguistics, fuzzy analysis, and principal components analysis, among others, to analyze text. Analyses can be described as one-dimensional, or list-based (e.g., identifying which institutions or authors are most active in a field), and two-dimensional, or matrix-based (e.g., seeking relationships based on co-occurrences of terms, affiliations, or whatever) [9]. Profiling may focus on documents (e.g., "bucketing" documents into related, manageable groups [8, 10]). Or, it may focus on concepts (e.g., "principal components analysis" to group related terms as conceptual clusters [11, 12]). A third choice is a combination – seeking to link documents to concepts (e.g., relevance scoring [13]). Conceptual distinctions and methods are discussed elsewhere [14].
3) *Extract latent relationships*. *Tech OASIS* applies iterative principal components analyses to uncover links among terms and underlying concepts (c.f., examples on the website: http://tpac.gatech.edu [15, 8, 10-12]).
4) *Represent relationships graphically*. The following sections elaborate generation of "maps" [16].
5) *Interpret* the prospects for successful technological development. This typically entails integrating the bibliographic search set analyses with expert domain knowledge (interviews) [11].

The TOA process strives to create knowledge from a "body" of literature beyond that obtainable by digesting individual pieces. We *treat retrieved text as data* [17] to parse text into informative units, count those units, and uncover patterns that can speak to information analysts' interests and management needs. Work on text mining is extremely active. For our purposes, this

[2] The *Tech OASIS* has been developed under the joint sponsorship of the Defense Advanced Research Projects Agency (DARPA) and the U.S. Army Tank-automotive and Armaments Command (TACOM). Sponsorship of such collaborations aspires to address common "functional capability" needs, promote efficient R&D resource utilization, and, thereby, strengthen U.S. military material capabilities.

draws on efforts under several labels, including "KDD" (Knowledge Discovery in Databases --c.f., www.cs.cmu.edu/~dunja/WshKDD2000.html; www.cs.biu.ac.il/~feldman/ijcai-workshop%20cfp.html), and bibliometrics (counting of bibliographic activity -- c.f., sistm.web.unsw.edu.au/conference/issi2001).

This paper focuses on extracting latent relationships through PCA factor analysis and representing the derived relationships graphically. The following section focuses on one type of TOA-based knowledge representation -- technology maps. We exemplify the clustering development using a simple search in the *INSPEC* database. *INSPEC* is a widely available R&D publication database abstracting some 300,000 journal articles and conference papers, annually, from select technical domains. IEE produces *INSPEC* and licenses it to be available through various sources (e.g., through "Dialog" or by subscription). The literature search string -- "KNOWLEDGE (adjacent to) (DISCOVERY OR ACQUISITION OR REPRESENTATION) AND NATURAL (adjacent to) LANGUAGE?)" -- captured 1,202 literature abstracts published between 1990 and 2001in the *INSPEC* database. This "natural language knowledge discovery" search set was retrieved on November 15, 2001. A thorough analysis of this topic would certainly warrant more extensive review of the documented R&D, as well as incorporation of subject matter expert perspectives.

3. Factor Maps[3]

As noted, we seek to identify and represent relationships inherent in sets of abstracts resulting from a database search. This inductive approach does not impose groupings, but instead elicits them from the data. We have developed an automated process to do so based on "co-occurrence" information. Co-occurrence is based on the pattern of terms occurring together in the records. If two terms occur together in the records more frequently than expected, there is a presumption of relationship between them. Terms can include authorship (also organizational affiliation, nationality) or "keywords" (subject index terms), or noun phrases generated from titles or abstracts using our natural language processing (NLP) routine [18, 19].

Principal components analysis (PCA) is a useful technique for extracting the main relationships implicit in a data set. A PCA-based approach called Latent Semantic Indexing (LSI [14, 20]) generates conceptual indices instead of individual words to improve information retrieval. LSI is based on "co-occurrence" information from large text sources, such as entire collections of abstract records. Interesting issues concern the use of grouping techniques such as PCA and LSI, cluster group quality being central to this research, as well as how to represent relationships quickly and effectively [15, 8, 9, 10, 12, 19].

Effective visualization of the basic co-occurrence and correlation matrix information entails a sequence of analyses[18, 16,24]:

- extract the principal components, or factors
- locate the factors as nodes in a 2-dimensional graph (whose axes have no absolute interpretation) using a proprietary two-step multi-dimensional scaling (MDS) algorithm
- connect the nodes using an improved path-erasing algorithm that reflects association better than MDS
- use a routine to determine and display node size (relative frequency of occurrence)
- include a routine to consolidate near-duplicate principal components (in the mapping process)
- apply an algorithm to automatically name principal components
- apply another algorithm to cut off principal components to include only high-loading terms
- automate the mapping process using macro's (scripts) to create maps in *Tech OASIS*, *VantagePoint, Microsoft Word* or *MS PowerPoint*.

The *Tech OASIS*/*VantagePoint* routine generates various maps, such as:

[3] Formally, we are dealing with principal components, but it is somewhat more convenient to call them "factors." Again, formally, these can be distinguished from various "clustering" approaches. We are using PCA to group related terms.

1) terms map [represents the relationships among frequently occurring subject index terms, title phrases, or whatever terms are chosen]
2) affiliations map [represents the relationships of affiliations' research topics, based on terms they use in their documents]
3) authors map [analogous to affiliations map, but for individual researchers]
4) countries map [analogous to affiliations map]
5) sources (e.g., journals) map [analogous to affiliations map]
6) principal components map [represents the relationships among conceptual clusters -- see Figures 1 and 2].

Figures 1 and 2 show factor maps for the “natural language knowledge discovery” topic. Displayed are the most dominant factors abstracted in *INSPEC* for the 1990-2001 time period. The term shown is the highest loading keyword for that cluster of keywords. The size of a node reflects the number of publications sharing the high-loading descriptor terms that constitute the factor. Positioning is determined using the MDS algorithm. MDS is the generally favored approach to reduce n-dimensions to 2-D or 3-D. In MDS, an important parameter called stress is used to control its procedures. The process of generating an MDS map seeks the optimum location for each element in the map by minimizing the stress. Traditionally, the "steepest descent" algorithm is employed in most MDS applications (e.g., SPSS uses this). We have found that the "steepest descent" algorithm is not very effective in a number of text mapping cases, especially for 3-D solutions. The algorithm can often be trapped in a local minimum of "stress space," never reaching the global minimum.

We have devised a "step-by-step" search algorithm. This algorithm is effective at finding the global stress minimum, although it usually consumes more CPU time than the "steepest descent" algorithm. In our mapping algorithm, we bind "step-by-step" and "steepest descent" algorithms in the MDS iteration process. Comparison of a number of cases shows our MDS solution to yield substantially better visual representations than other MDS solutions.

As noted, MDS tries to represent high dimensional spatial relations by displaying the elements in 2-D or 3-D spaces. The resulting distortions tend to become problematic when many elements (hence many dimensions) are involved. Therefore we have added an additional representational element, connecting links, based on a “path-erasing” algorithm. This is built on a proximity matrix among the elements (in Figure 1 and 2, the cluster groups

Our experience is that the path-erasing-based links are easily perceived as dominant proximity representations, with the MDS-based location taken as secondary. Path-erasing removes the weaker links to whatever level is desired, to give an informative map. The MDS axes are essentially arbitrary, so our routine provides four alternative axial perspectives for 3-D views. [We have done 3-D representations, but do not include these in the current software, instead enabling zoom-in viewing options.] We find the resulting representations superior to others in capturing the conceptual entities and visualizing them [21].

As noted, the representations are not singular – different numbers of entities mapped and different views give quite different results. Note the differences between Figures 1 and 2, visualizations of the same information. In these two figures, we cluster the same number of descriptors, terms contained in each abstract that define the context and content of the documented research. Only the number of factors (i.e., principle components) considered in the analyses vary for Figures 1 and 2. Figure 1 shows a 14-factor map resulting from a 16-factor request; Figure 2 shows 11 factors, resulting from requesting 12 – the PCA algorithm adjusts to fit.

The high-loading descriptors for each factor define the abstracts contained in each cluster group. For example, in Figure 1, the large “robots” cluster includes 11 descriptors: in addition to “robots,” "learning (artificial intelligence)," "neural nets," "artificial intelligence," "data mining," "software agents," "computer vision," "genetic algorithms," "cooperative systems," "information resources" and "internet." The four descriptors, "knowledge engineering," "robots," computer vision" and engineering computing," define the abstracts contained in the second, smaller, "robots" cluster group. These two "robots" cluster groups have two common "group-defining-terms," "robots" and "computer vision." The Cluster group abstracts are delineated by "or" logic using the "group-defining-terms."

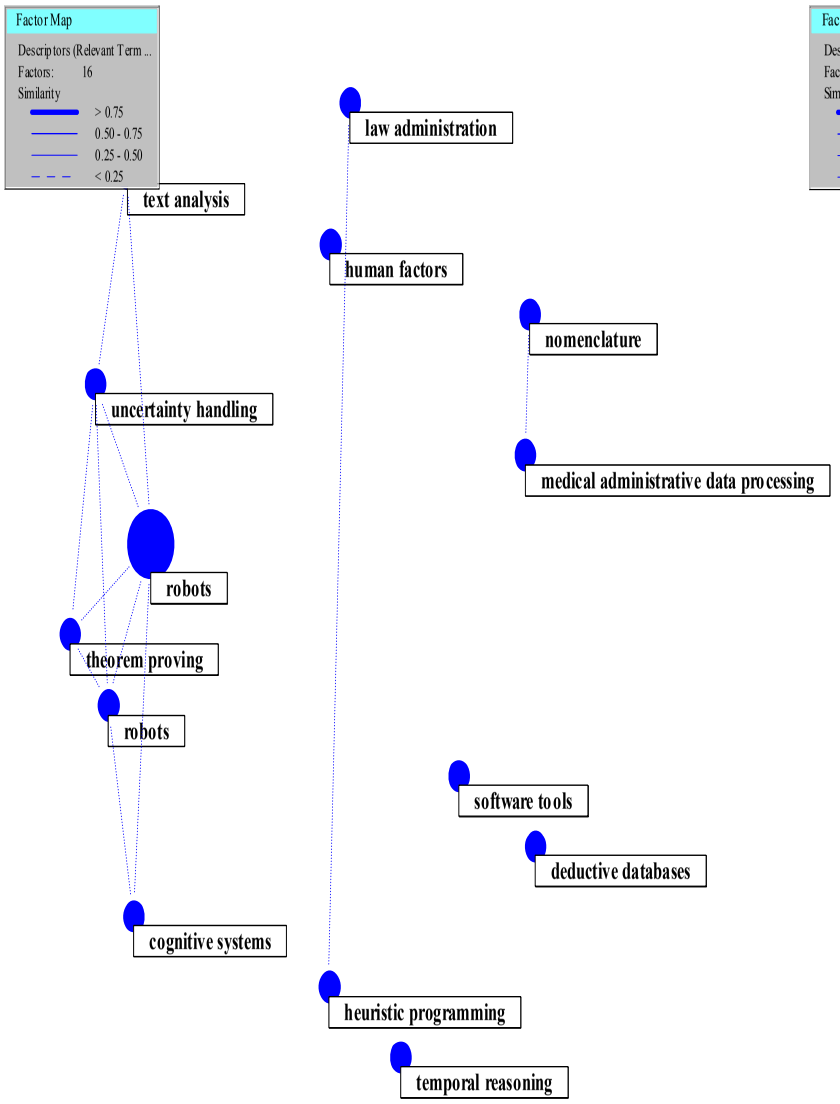


Figure 1 - Natural Language Knowledge Discovery (INSPEC) 16 Factors' Map

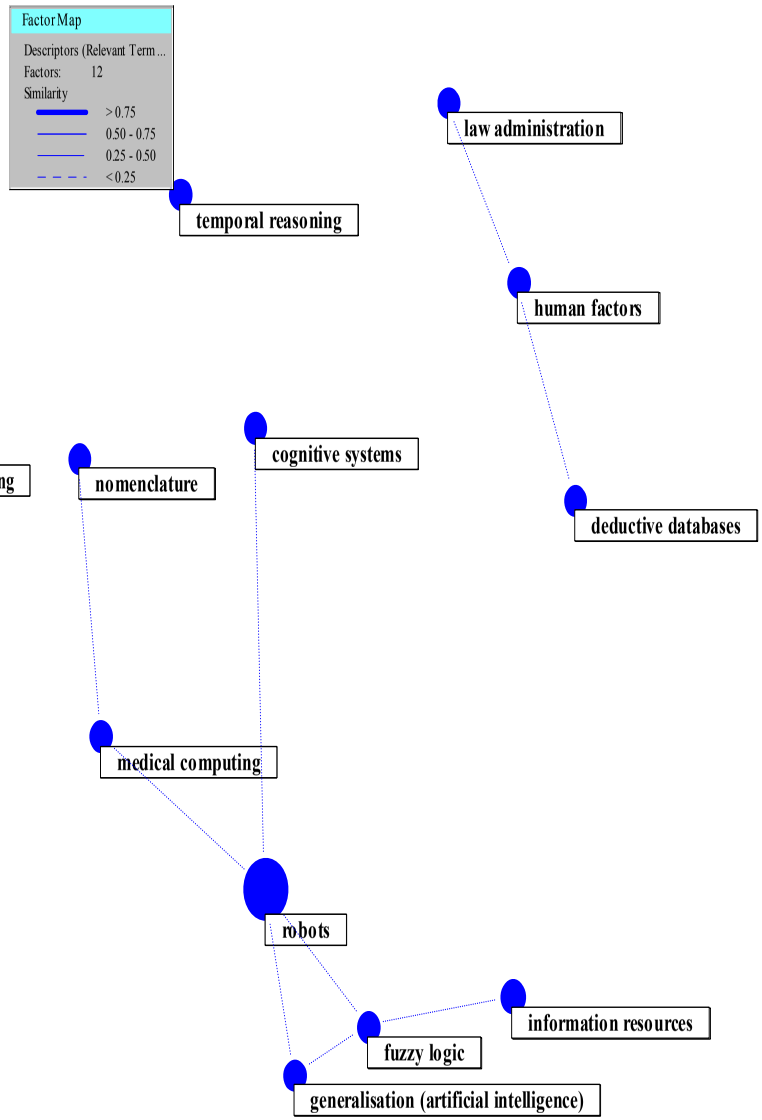


Figure 2 - Natural Language Knowledge Discovery (INSPEC) 12 Factors' Map

4. Cluster Group Quality Measures

The internal quality measure that we apply toward developing a "best" factor analysis approach is cluster cohesion. Cohesiveness emanates from the vector space model of document information cluster analysis. In the vector space model, a term frequency vector represents each document. The terms we choose to represent the documents from the "natural language knowledge discovery" abstract set are those from the "Descriptors" field. Each descriptor occurs only once in each document. Each document, generally, contains five to eight descriptors, which attempt to depict the content of the abstracted paper. Table 1 lists high frequency descriptors (i.e., those occurring in the highest number of abstracts) from the 1202 "natural language knowledge discovery" documents. Terms and phrases that occurred in 8 or fewer abstracts are not shown in Table 1, since, as will be discussed later, they were not used to create the factor groupings. The two highest frequency phrases, "natural languages" and "knowledge representation," have also been excluded from the PCA factor analysis, since, as will be explained later, they offer little discriminating value.

All 1202 document vectors consist of a sequence of 1's and 0's, as representation of inclusion or exclusion of each of the 99 descriptors in Table 1. Each document vector is then normalized to be of unit length. The average pair-wise similarity between cluster group documents depicts the *cohesion* measure for each cluster group generated. The pair-wise similarity is computed by the vector cosine measure, which for unity vectors equals the vectors' dot product. Our "best" factor analysis process shall strive to maximize the cluster groups' cohesion.

Entropy provides an external measure of cluster quality for non-nested clusters or clusters at one level of a hierarchical grouping. For each cluster grouping, we first compute the probabilities, P_{ij} , each representing the probability that a member of cluster j belongs to group i, which we define as the non-common derived cluster groups. These probabilities can be obtained by analyzing the co-occurrence matrix, which has the derived cluster groupings as both the rows and column entries. Table 2 presents such a co-occurrence matrix for the 12-factor PCA. Our algorithm uses a descriptor loading-factor threshold to define the existence (i.e., relevance) of a derived cluster group. The descriptor loading-factors for one factor in the 12-factor analysis did not exceed this threshold, therefore, only 11 cluster groups were generated for the 1202 "natural language knowledge discovery" abstracts. Looking at Table 2, the probability that a record from the "robots" cluster (i.e., j = 1) could belong to the "information resources" cluster (i.e., i = 2) would be 24 / 129 or 0.186. The entropy of each cluster j can then be calculated using the formula,

$$\mathbf{Entropy_j = - Sum,\ i = 1\ to\ m,\ of\ [\ P_{ij}\ log\ (P_{ij})]}$$

where the sum is taken for all groups, excluding each group where i = j. The sum of the weighted entropies for each cluster grouping equals the total entropy:

$$\mathbf{total\ entropy = Sum,\ j=1\ to\ m,\ of\ [\ (n_j * Entropy_j)/n]}$$

where n_j equals the number of abstracts in cluster j, m is the number of clusters and n equals the total number of abstracts in the file (i.e., 1202). The exclusion of the matrix diagonal entries from the analysis attempts to minimize the comparative entropy penalty that a larger number of factor clusters would have versus a smaller number of factor clusters. Our analysis approach attempts to minimize the total factor grouping entropy. However, we do not want to unduly penalize factor groupings that generate a large number of clusters. A larger number of small clusters may have a higher total cohesion than a smaller number of larger clusters.

The *F measure* represents the second external cluster quality measure that we integrate into the "best" factor grouping process. The total F measure for a factor cluster grouping is defined as

$$\mathbf{F = Sum,\ j = 1\ to\ m,\ of\ [\ (n_i / n)\ max\{F(i,j)\}]}$$

Where

Table 1 - Natural Language Knowledge Discovery, INSPEC (1990-2001), Descriptors

# Records	Descriptors (Cleaned)	Relevant Terms
862	natural languages	0
699	knowledge representation	0
402	knowledge acquisition	1
257	knowledge based systems	1
210	computational linguistics	1
207	inference mechanisms	1
187	natural language interfaces	1
124	grammars	1
105	linguistics	1
101	learning (artificial intelligence)	1
83	neural nets	1
77	expert systems	1
72	artificial intelligence	1
67	information retrieval	1
65	formal logic	1
62	language translation	1
58	semantic networks	1
56	word processing	1
53	user interfaces	1
49	deductive databases	1
35	logic programming	1
35	planning (artificial intelligence)	1
32	medical expert systems	1
31	knowledge engineering	1
30	graph theory	1
30	robots	1
29	cognitive systems	1
29	formal languages	1
28	fuzzy logic	1
28	problem solving	1
28	text analysis	1
27	explanation	1
27	intelligent tutoring systems	1
27	interactive systems	1
27	learning systems	1
27	temporal reasoning	1
27	user modelling	1
25	case-based reasoning	1
25	formal specification	1
25	query processing	1
24	computer aided instruction	1
24	database management systems	1
24	fuzzy set theory	1
24	object-oriented programming	1
23	data mining	1
23	software agents	1
21	glossaries	1
21	speech recognition	1
21	uncertainty handling	1
20	computer vision	1
20	document handling	1
20	hypermedia	1
20	software tools	1
18	decision support systems	1
18	nomenclature	1
18	software engineering	1
17	information retrieval systems	1
17	learning by example	1
17	spatial reasoning	1
17	temporal logic	1
16	information analysis	1
16	law administration	1
16	medical administrative data processing	1
15	classification	1
15	data structures	1
15	genetic algorithms	1
14	constraint handling	1
14	cooperative systems	1
14	graphical user interfaces	1
14	pattern recognition	1
13	indexing	1
13	relational databases	1
13	statistical analysis	1
12	common-sense reasoning	1
12	diagnostic expert systems	1
12	engineering computing	1
12	medical information systems	1
12	symbol manipulation	1
12	vocabulary	1
11	information resources	1
11	multimedia computing	1
11	search problems	1
11	thesauri	1
10	biology computing	1
10	computational complexity	1
10	digital simulation	1
10	frame based representation	1
10	Internet	1
10	object-oriented databases	1
10	theorem proving	1
9	computer graphics	1
9	generalisation (artificial intelligence)	1
9	groupware	1
9	heuristic programming	1
9	human factors	1
9	medical computing	1
9	probability	1
9	psychology	1
9	scheduling	1
9	systems analysis	1
9	very large databases	1

Table 2 – Natural Language Knowledge Discovery 12 Factor Co-Occurrence Matrix

	Descriptors (Relevant Terms) (12 FACTORS)	Descriptors (Relevant Terms) (12 FACTORS)											
	# Records	234	129	123	68	57	54	51	50	48	48	40
# Records		robots	information resources	medical computing	nomenclature	fuzzy logic	deductive databases	cognitive systems	human factors	generalisation (artificial intelligence)	law administration	temporal reasoning
234	robots	234	24	26	10	21	15	23	3	22	4	5
129	information resources	24	129	11	13	10	7		2	7	8	1
123	medical computing	26	11	123	14	8	9	7	7	7	5	3
68	nomenclature	10	13	14	68	3	6	1	2		1	3
57	fuzzy logic	21	10	8	3	57	6	5	3	6		1
54	deductive databases	15	7	9	6	6	54	1	3	5	2	2
51	cognitive systems	23		7	1	5	1	51	1	2	1	2
50	human factors	3	2	7	2	3	3	1	50		2	3
48	generalisation (artificial intelligence)	22	7	7		6	5	2		48	2	
48	law administration	4	8	5	1		2	1	2	2	48	
40	temporal reasoning	5	1	3	3	1	2	2	3			40

$$F(i,j) = (2 * Recall(i,j) * Precision (i,j)) / (Precision(i,j) + Recall(i,j))$$

And

$$Recall(i,j) = n_{ij} / n_i$$

$$Precision(i,j) = n_{ij} / n_j$$

n_{ij} equals the number of members of group i in cluster j, n_j is the number of members of cluster j, n_i equals the number of members of group i, and n is the number of documents. Referring to Table 2, we again designate the columns as the i groups and the rows as j clusters. F(2,1)=(2*(24/129)*(24/234))/((24/234)+(24/129)) = 0.1322311. As with the entropy calculations, the diagonal values are also excluded from the process analysis. The "best" factor analysis process attempts to minimize the total F measure and the total entropy, while maximizing the total cohesion of the derived factor cluster groupings.

5. Process Automation and Evaluation

To create a factor map, the analyst must first select the terms to be analyzed. Calculations involving matrix manipulations mandate moderation on term inclusion. Prior guidance suggested selecting 200 to 300 items for inclusion in the PCA. However, this guidance served only as a rule-of-thumb, which attempted to prevent inexperienced analysts from attempting to perform a factor analysis on a complete list of several thousands of items. Such an analysis would be time-intensive and would create a large number of hard-to-comprehend cluster groups. To assist inexperienced analysts in selecting the group of items to be analyzed, a simple linear regression algorithm was developed to analyze the Zipf distribution of the occurrence frequencies of the candidate list of items. Figure 3 depicts the Zipf distribution for the "occurrence frequencies" of the complete descriptors list for the 1202 "natural language knowledge discovery" abstracts. The subset, "relevant data" in Figure 3, designates those term frequencies nominated for factor analysis inclusion. In Table 1, terms deemed relevant are so indicated. Note that in a more substantively considered analysis, and certainly for less on-target term sets, we would likely assess terms individually for relevance.

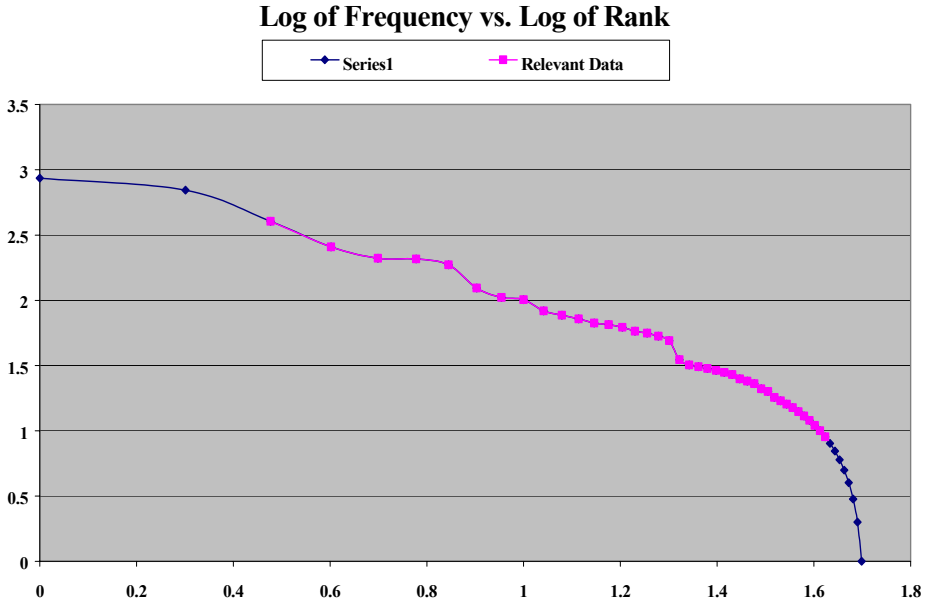
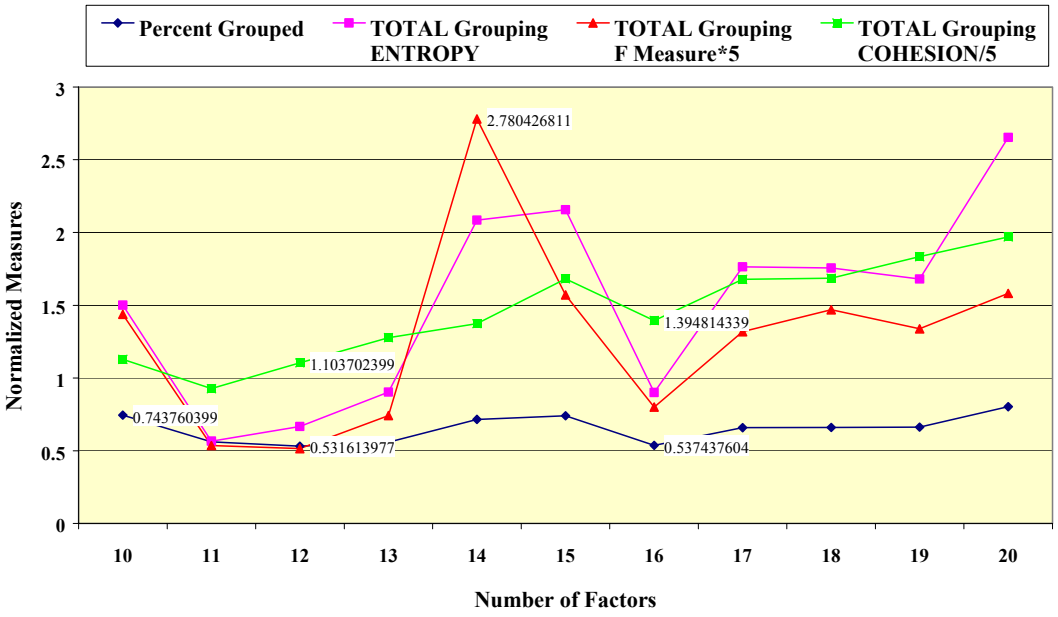


Figure 3 – Natural Language Knowledge Discovery, INSPEC (1990-2001), Descriptors' Log Frequencies

Once the items to be included in the PCA have been selected, the analyst must then specify the number of principal components to be formed. *Tech OASIS* defaults to the square root of the number of items to be analyzed (note, this default can be over-ridden by the analyst). The current research indicates that this default may be too low, since our derived "best" factor analysis, most often, uses a number of factors greater than 1.5 times the square root of the number of items to be analyzed.

Figure 4 displays the cluster quality measures -- cohesion, entropy and F measure, for the "natural language knowledge discovery" factor groupings. The number of factors requested was varied between 10 and 20. [The square root of the number of descriptors (99) in Table 1 was 10, which served as the beginning number of factors to be considered in the quest for the "best." The comparative analysis ended at two times the square root number, or 20.] As a general observed trend in Figure 4, increasing the number of factors results in increases to all quality measures. However, variations, realized by single increments to the number of factors requested of the PCA analysis, do not always follow this trend (note in Figure 4 that the F measure is maximum at 14 PCA factors and the other two measures reach maximum at 20 factors). The relative changes in the three quality measures also vary from PCA factor grouping to factor grouping. The cluster groupings for 11 and 12 factors have the lowest entropy and F measures, one of our desired goals. However, in respect to our second cluster quality goal, the highest cohesion values result from

Figure 4 - Cluster Groupings' Quality Measures - Natural Language Knowledge Discovery (INSPEC 1990-2001)



cluster groupings of 19 and 20 factors.

The quality measures have been weighted in figure 4 to allow their display on the same scale. The numbers of factors have been ordered to depict positive slopes of the quality measure curves. The "best" solution must represent a compromise between the minimization and maximization objectives. Also important, but not reflected in the factor rank order of Figure 4, is the percentage of the 1202 abstracts clustered in each factor analysis. The percentage of the abstracts clustered must be known and considered by those viewing a factor map. Figures 1 and 2 display the factor maps for 16 and 12 factors requested, respectively – the two "best" ranked abstract cluster groupings. Both the 16 and 12 factor, PCA analyses, cluster approximately 54% of the 1202 records. When viewing these Figures, the analyst must recognize that about 46% of the records in the "natural language knowledge discovery" file have not been captured in these clusters.

Comparing the two “best” rated factor maps, Figures 1 and 2, one observes apparently large differences in the depiction of the same information. However, the co-occurrence matrix, shown in Table 3, of the two sets of abstract cluster groupings, derived by the PCA analyses requesting 16 and 12 factors, shows significant record commonality between them. Note the diagonal record matches between the two sets of cluster groupings and the relative sparseness off this diagonal. The primary differences in cluster groupings reside within the largest cluster groups of the 12-factor PCA. The 12-factor "robots" cluster group has been divided into two groups in the 16-factor PCA, as mentioned earlier. The "information resources" cluster group has similarly been sub-divided into the "robots" and "text analysis" clusters of the 16-factor PCA. The most significant difference lies in the dispersion of the abstracts in the 12-factor PCA cluster group, "medical computing." The 123 abstracts of "medical computing" have been primarily placed in the cluster categories "medical administrative data processing," "software tools," "robots" and "cognitive systems." The Table 3 cluster groups' record comparisons, then, indicate that the two sets of abstracts' cluster groups are really similar.

Figure 1 represents the "best" "natural language knowledge discovery" cluster map, based on a composite index that strives to maximize internal cluster groups' cohesion and minimize external cluster groups' entropy and F measures. Recall though, that this map, Figure 1, reflects 54% (i.e., 646 abstracts) of the information contained in the 1202 abstracts file. What cluster groupings exist in the remaining 47% of the "natural language knowledge discovery" abstracts' sub-file? Figures 5 and 6 depict the "best" factor maps for the non-clustered abstracts of Figure 1. Of the 556 abstracts not clustered in Figure 1, 283 (i.e. 51% of the non-clustered abstracts and 24% of the original 1202 abstract file) get clustered, as depicted in the Figure 5, following the same quality measure index optimization process as used to derive the groups in Figure 1. Similarly, the abstracts not clustered in Figure 5 have been clustered in Figure 6, again using the quality measure index optimization process. The Figure 6 cluster groups contain only 100 of the 273 abstracts not clustered in Figure 5 (i.e., 37% of Figure 5 non-clustered abstracts or 8% of the original 1202 abstracts.). Figures 1 and 5 then portray 78% of the information contained in the 1202 "natural language knowledge discovery" abstracts' file.

Table 3 – Natural Language Knowledge Discovery 16 and 12 Factor Co-Occurrence Matrix

	Descriptors (Relevant Terms) (16 FACTORS)	Descriptors (Relevant Terms) (12 FACTORS)										
	# Records	234	129	123	68	57	54	51	50	48	48	40
# Records		robots	information resources	medical computing	nomenclature	fuzzy logic	deductive databases	cognitive systems	human factors	generalisation (artificial intelligence)	law administration	temporal reasoning
263	robots	215	68	23	12	27	17	21	8	24	6	5
71	robots (2)	66	5	15	2	6	4	9	2	10	1	1
43	text analysis	14	43	2	5	3	2		1	1	2	
62	medical administrative data processing	11	7	39	36	6	9	1	1	2	1	2
44	nomenclature	6	11	5	44	2	2	1	1		1	1
75	uncertainty handling	26	10	8	3	57	6	5	3	9	2	2
54	deductive databases	15	7	9	6	6	54	1	3	5	2	2
71	cognitive systems	27	2	12	2	5	1	51	2	2	1	2
73	human factors	9	3	9	2	4	4	2	50	2	5	3
48	theorem proving	22	7	7		6	5	2		48	2	
48	law administration	4	8	5	1		2	1	2	2	48	
52	temporal reasoning	5	2	4	3	1	3	3	4			40
56	software tools	10	5	24	1	5	5	2	5	2	2	2
31	heuristic programming	7	6	3	2	2	2	1		1	4	2

6. Conclusions

We introduced the concepts of technology opportunities analysis (TOA) and described the software tool, *Tech OASIS* / *VantagePoint*, which was used to perform this research. We then explained the factor map analytical process that we strive to optimize. We discussed the Steinbach, et al., internal and external cluster group quality measures and presented an approach on how to use them to select the 'best" (i.e. standard) PCA factor cluster groups. The derived cluster quality measures, obtained for a series of PCA's, while varying only the number of factors, were presented and discussed in regards to their variance over the range of selected PCA factors. We finished with the automated process description and our preliminary assessment of the derived "best" factor maps. The three "best" factor maps sequentially clustered approximately 86% of the 1202 "natural language knowledge discovery" abstracts. The context of Figure 1 appeared to center on "robots," "law administration" and "medical administrative data processing." The focus of Figure 6 appeared to be "language translation," "speech analysis and processing" and "information retrieval." The final 8% of the abstracts profiled, Figure 6, centered on "planning (artificial intelligence)," "operating systems (computers)" and "mobile robots." Obviously, the observed or interpreted context of the cluster maps depends on the items selected on which to perform the PCA. Recall that we used the Zipf distribution of the list item occurrence frequencies as a basis for analysis selection.

The use of the terms designated by the Zipf distribution analysis represents a sub-optimal solution to that that might be obtained by an alternative approach, such as a term frequency, inverse document frequency ("tf-idf") type of analysis. Under a tf-idf approach, a descriptor that occurred more frequently in the search set than in the database searched (e.g., INSPEC) would be considered more relevant than one that occurred with the same or lower frequency in the search set than the database searched, irrespective of the term's search set occurrence frequency. Using the Zipf distribution approach, all list items are considered relevant that have occurrence frequencies between the linear regression determined start and end occurrence frequencies. We will be investigating the use of a tf-idf approach during the next two years.

We assess use of a thesaurus for term selection for a "best" factor analysis elsewhere [23]. The current implementation represents a viable trade-off among computation time, information availability, ease of use and effective representation. The database term frequency information has yet to be obtainable from database suppliers. Therefore, the Zipf distribution application provides a standardized approach that could be integrated with the cluster group quality measure analysis.

In sum, we offer a way to obtain an arguably "best" cluster representation of sizable sets of abstract texts. This approach lends itself to scripting. In turn, that means that these text mining routines can be used quickly and cheaply to inform technology management decision processes. The automation is not, and should not be, total. The analyst exerts discretion in searching, term set determination, depiction (e.g., whether to show pull-downs of the individual terms associated with each factor), and involving subject matter experts to aid in interpretation. We believe this approach to make text analyses accessible is critical to their utilization in decision-making.

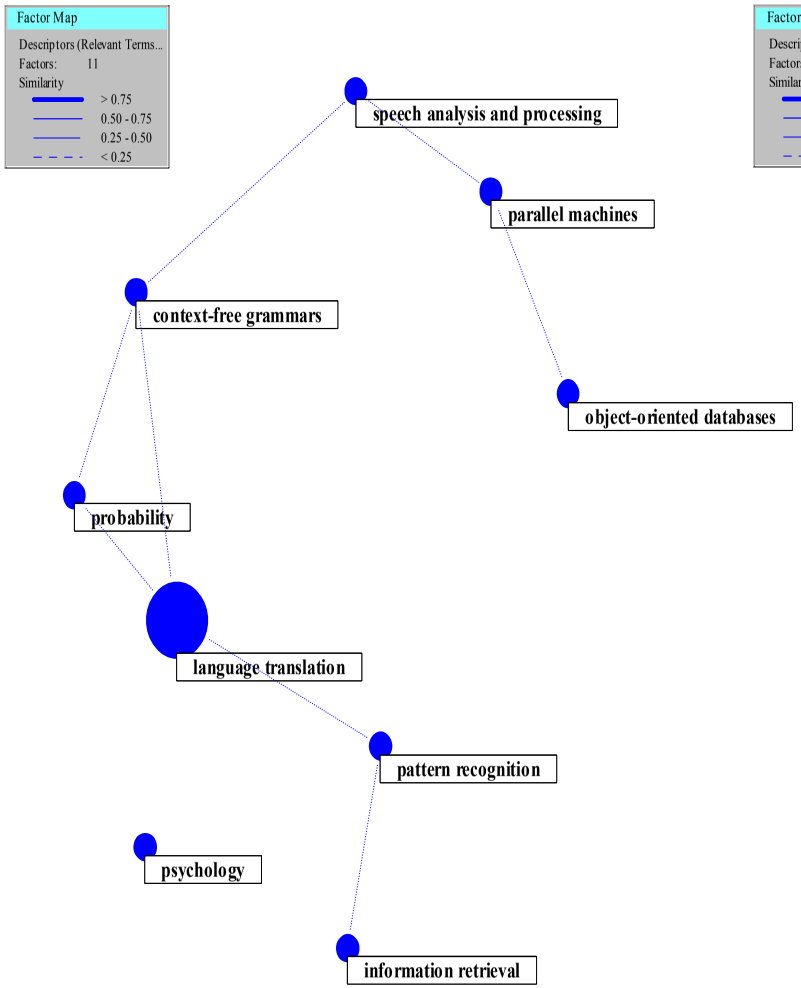


Figure 5 – Natural Language Knowledge Discovery; Other "Best" Factor Map

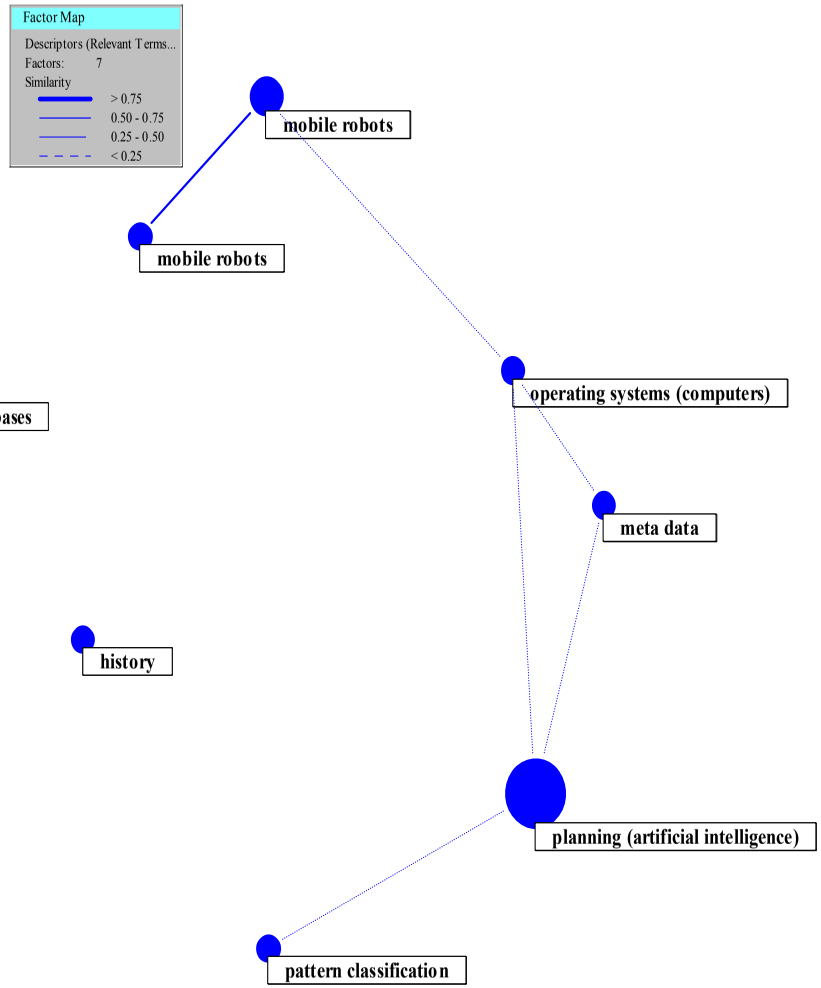


Figure 6 – Natural Language Knowledge Discovery; Other-Other "Best" Factor Map

References

[1] Katy Borner, Chaomei Chen, Kevin W. Boyack, "Visualizing Knowledge Domains," submitted ARIST, Volume 37 09/30/2001

[2] Michael Steinbach, George Karypis, Vipin Kumar, "A Comparison of Document Clustering Techniques" University of Minnesota, Technical Report #00-034 (2000). http://www.cs.umn.edu/tech_reports/

[3] Gerard Salton, James Allan, Chris Buckley, Amit Singhal, "Automatic Analysis, Theme Generation and Summarization of Machine-Readable Texts," Science, Volume 264, 3 June 1994 1421-1426

[4] F. Murtagh, Comput. J 26 354 (1982), W.B. Croft, J. Am. Soc. Int. Sci 28 341 (1977) and G. Salton and A. Wong, ACM Trans Database Syst. 3 321 (1978)

[5] Porter, A.L., Roper, A.T., Mason, T.W., Rossini, F.A., and Banks, J.: *Forecasting and Management of Technology*. Wiley, New York, NY, 1991.

[6] Porter, A.L., Jin, X-Y., Gilmour, J.E., Cunningham, S.W., Xu, H., Stanard, C., and Wang, L., 1994. Technology Opportunities Analysis: Integrating Technology Monitoring, Forecasting & Assessment with Strategic Planning, *SRA Journal (Society of Research Administrators)* 21(2), 21-31.

[7] Porter, A.L., and Detampel, M.J.: Technology Opportunities Analysis, *Technological Forecasting and Social Change*49 (2), 237-255 (1995).

[8] Watts, R.J., Porter, A.L., Cunningham, S.W., and Zhu, D.: *VantagePoint* Intelligence Mining: Analysis of Natural Language Processing and Computational Linguistics, in *Principles of Data Mining and Knowledge Discovery (First European Symposium, PKDD'97, Trondheim, Norway)*, J. Komorowski and J Zytkow, eds., p. 323-335: Springer, 1997.

[9] van Raan, A.F.J. (1992). "Advanced Bibliometric Methods to Assess Research Performance and Scientific Development: Basic Principles and Recent Practical Applications," Research Evaluation, 3(3): 151-166 See also website: http://sahara.fsw.leidenuniv.nl/cwts/nnmap0.html)

[10] Watts, R.J., Porter, A.L., Courseault, C.: Functional Analysis: Deriving Systems Knowledge from Bibliographic Information Resources, *Information, Knowledge, Systems Management* 1(1), 45-61 (1999).

[11] Watts, R.J., Porter, A.L., and Newman, N.C.: Innovation Forecasting Using Bibliometrics, *Competitive Intelligence Review* 9(4), 1-9 (1998).

[12] Watts, R.J., and Porter, A.L.: Innovation Forecasting, *Technological Forecasting and Social Change* 56, 25-47 (1997).

[13] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, D.: Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science* 41, 391-407 (1990).

[14] Carlisle, J.P., Cunningham, S.W., Nayak, A., and Porter, A.L., Related Problems of Knowledge Discovery, *Hawaii International Conference on System Sciences [HICSS] Proceedings on CD – Modeling Technologies and Intelligent Systems Track; Data Mining and Knowledge Discovery Mini-track*, January, 1999.

[15] Zhu, D., Porter, A.L., Cunningham, S., Carlisle, J., Nayak, A.: A Process for Mining Science & Technology Documents Databases, Illustrated for the Case of "Knowledge Discovery and Data mining, *Cienciaa da Informacao* 28(1), 1-8. (1999).

[16] Zhu, D., Porter, A.L., Cunningham, S.W., Carlisle, J., and Nayak, A.: "A Process for Mining Science & Technology Documents Databases, Illustrated for the Case of 'Knowledge Discovery and Data Mining'," Internal TOA Paper #94 [available on request].

[17] Losiewicz, P., Oard, D.W., and Kostoff, R.N.: Textual data mining to support science and technology management, *Journal of Intelligent Information Systems* 15(2), 99-119 (2000).

[18] Zhu, D. and Porter, A.L.: *TOA: Illustrated for the Case of Knowledge Discovery in Databases and Data Mining* ; http://tpac.gatech.edu.

[19] Porter, A.L.: Text Mining for Technology Foresight, *Futures Research Methodology,* J. Glenn and T. Gordon, eds., AC/UNU, to appear.

[20] Kostoff, R.N.: various reports on bibliometrics, http://www.scicentral.com/G-scipol.html#reports; http://www.dtic.mil/dtic/kostoff/index.html

[21] Berry, M.W., Dumais, S.T., and Letsche, T.A.: Computational Methods for Intelligent Information Access, http://www.cs.utk.edu/~berry/sc95/sc95.html, Feb, 1995.

[22] Schvaneveldt, R.W. (Ed.): *Pathfinder Associative Networks: Studies in Knowledge Organization*, ISBN 0-89391-624-2, 1990.

[23]Watts, Robert and Porter, Alan: Tracking the Evolution of Management of Technology (MOT), *International Association for Management of Technology (IAMOT) 2002 Conference,* (Submitted)

[24] Zhu, D. and Porter, A.L.: Automated Extraction and Visualization of Information for Technological Intelligence and Forecasting, *The 21st International Symposium on Forecasting, Pine Mountain, Georgia, June 17-20, 2001.*