9

A MATHEMATICAL APPROACH TO CATEGORIZATION AND LABELING OF QUALITATIVE DATA: THE LATENT CATEGORIZATION METHOD

*Kai R. Larsen**
*David E. Monarchi**

As text databases increasingly become available to researchers, the limits to human cognition are rapidly reached. Focusing on examining objective realities, this paper introduces the latent categorization method, a novel new research method for analysis of large and midsize data sets. This method clusters text artifacts and extracts the words that were most important in creating the clusters. Further, it demonstrates a set of techniques for extracting knowledge from a representative data set involving 6135 abstracts from a variety of business-related journals.

1. INTRODUCTION

Numerous approaches to qualitative research exist. The categorization suggested by Miles and Huberman (1994) splits research approaches into *interpretivism* and *social anthropology*. Interpretive research,

The authors wish to thank the anonymous reviewers of *Sociological Methodology*. Address correspondence to kai.larsen@colorado.edu.

*University of Colorado, Boulder

including phenomenological, semiotic, and hermeneutic approaches, holds strongly that human action and discourse cannot be analyzed by using methods from the natural and physical sciences. In contrast, social anthropological research, including ethnographic methods, grounded theory, and case study approaches, tends to look for regularities in language and artifacts, and tends to believe in an objective reality. In these social anthropological approaches, methods derived from the natural and physical sciences are considered appropriate.

With respect to social anthropology, one of the most desirable approaches to analyzing qualitative data is pattern matching (Yin 2003). Human beings are believed to be especially good at pattern matching and categorization—indeed that categorization is basic to all human intellectual activities (Estes 1994; Lakoff 1990; Rosch 1978); however, the reasoning behind human categorizations is often unclear. In addition, the human ability to categorize and recognize patterns breaks down as data sets grow in size. Lately, the use and importance of computer-aided methods for the management, coding, and retrieving of qualitative data have increased (Mackensen and Wille 1999). However, computer-aided analysis and aid in the categorization and pattern-matching process have been limited.

For academic categorization of quantitative data, several approaches are available, including factor analysis (e.g., Nunnally and Bernstein 1994) and cluster analysis (e.g., Aldenderfer and Blashfield 1984). Even within cluster analysis, however, it is not always clear why one cluster analytic solution should be selected over another (Anderberg 1973). When categorizing qualitative data, the lack of reasoning behind such categorizations is several orders of magnitude more problematic due to inherent human cognition factors.

Because individual human perspectives can vary considerably, approaches to ensuring interrater reliability (e.g., Miles and Huberman 1994) purport to identify categorization quality, and Cohen's alpha standardizes such approaches across data sets (Cohen 1960; Cohen 1968). However, no approach explicates *what* leads to a specific categorization. Furthermore, no reliable and generalizable approach to describing the content of a category with linguistic content has been developed. This situation is cause for concern, given Gordon's (1999) statement that

> there were claims that the main criteria for assessing a [categorization] were its interpretability and usefulness. There are clearly dangers in such an approach: the human brain is quite capable of providing *post hoc* justifications of results of dubious validity.

Related to the interpretation challenge, the labeling of categories has long represented a problem for categorization scholars. Even when other items in a category have labels, explaining the human naming of an additional item is nontrivial. Although exemplar theorists (e.g., Kruschhke 1992) and prototype theorists (e.g., Hampton 1995) expect human labeling of an item to be derived from its similarity to existing labeled items, experiments have shown that similarity between an additional item and already labeled items does not always explain human labeling (Sloman, Malt, and Fridman 2001). The researchers point to linguistic convention—the history and norms that have affected the specifics of a word or label's use—and other factors not yet fully understood as the culprits. Difficulties aside, Richards and Richards (1995: 87–89) provide three principles for hierarchical categorization:

1. The children of a category should be cases in the same sense of the parent.
2. The description of a given category should apply to all the categories in the subtree below it. The subcategories in a tree should not switch partway down; that is, they must remain generic with respect to the higher categories.
3. One topic or idea should occur in only one place in the index system.

Although somewhat vague, principle 1 suggests that in a well-designed structure, links take on a general-to-specific form, where a label higher in the hierarchy is a general case of the labels below it (or at the minimum, identical to them). Principle 2 advocates vertical consistency in labeling, and principle 3 advocates horizontal purity.

The problem of assessing the content of qualitative categories is of extraordinary importance. In an attempt to contribute to research methods in the social anthropological category, this work began with a desire to develop a reproducible representation of artifacts (documents,

interview data, survey data, etc.) and an approach to labeling that representation in a way that would (1) reduce (or delay as much as possible in the analysis process) problems of human interpretation of the data; and (2) allow the application of quantitative techniques based on cardinal, rather than ordinal or nominal, data. The proposed approach is termed the latent categorization method (LCM). Such an advance would offer an alternative as well as a complement to some existing methods for categorizing and labeling qualitative data. LCM promises to greatly reduce the time spent by researchers on data analysis and interpretation, as well as to provide another tool for qualitative researchers to better interpret an objective reality.

2. FOUNDATIONS

2.1. *Theoretical Underpinnings*

The approach is based on latent semantic indexing (LSI), which has its roots in information retrieval research. Experiments have shown LSI to improve the quality of information retrieval over other methods (Bartell, Cottrell, and Belew 1995; Berry, Dumais, and O'Brien 1995). The output of LSI is a numerical representation of the artifacts in the context of the terms they use. More accurately, this approach constructs a representation of the artifacts in the context of the terms that were used or could have been used to write them. In this sense, it latently captures the semantics of the artifacts numerically, thus permitting them to be indexed and retrieved. For an introduction to LSI, see Deerwester et al. (1990).

2.2. *Data Selection*

The work is started by defining a corpus as a collection of artifacts. For simplicity, it is assumed that the artifacts are all written in the same language and that they share a common domain of discourse. In principle, the artifacts themselves could be recursively composed of artifacts. For example, a corpus could comprise documents composed of paragraphs composed of sentences. Hereafter, interviews,

abstracts, documents, or other textual sources of interest to the researcher will be referred to simply as *artifacts*.[1]

2.3. *Data Preparation*

In general, the terms employed in writing the paragraphs can be incorporated into an analysis system in ways ranging from a simple term-index approach to sophisticated techniques that require determining the part of speech (POS) of each word in an attempt to infer the essence of the text. (For a discussion of various approaches used in search engines, see Belew 2000; Berry and Browne 1999.) LSI does not make use of POS information; nor, at the other extreme, does it directly index the terms. LSI begins by treating each paragraph as a bag of words without structure. That is, LSI discards all POS information. A small number of words, however, referred to here as "stop words," occur in a disproportionate amount in English text (Luhn 1957; Zipf 1949); these stop words are discarded. Having little or no meaning when taken out of context, these words serve to provide structure to the sentence, and thus meaning indirectly, but are meaningless, or nearly so, by themselves. These words thus have little value as indexes into the artifacts because they are very poor discriminators. Usually this set of words contains articles, prepositions, pronouns, conjunctions, and common adjectives and adverbs.

A second set of words that are considered irrelevant for the purposes of the analysis at hand is also identified. These words are not intrinsic to the domain of discourse. For example, in interviews, people in a company may frequently refer to other people in the company. Under some circumstances, references to individuals and locations may be ignored, depending upon the goals of the research. This irrelevant set of words is also discarded, just as the stop words were; for convenience, both sets of words are merged and removed from the text.

The next step in data preparation is to "stem" the remaining words to avoid having multiple forms of a word (potentially) represented in the term-artifact table. Stemming converts a word to a

[1]Note that much of the research uses the word *document*, reflecting the original document-retrieval motivation of the development of LSI at Bell Labs in the 1990s.

related form; that is, it "conflates" the word. As examples, removing an "s" or "es" will convert some plurals to singulars (e.g., "books" to "book"), or stemming the words "walk," "walking," and "walked" reduces all three to "walk." Stemming increases recall while reducing precision. It also reduces the size of indexing files. In LSI, it reduces the size of the input matrix, which may have a significant impact on the computational cost of decomposing the matrix. Finally, stemming also allows a system to be more "user friendly" because it is not necessary to know the precise form of the word the author used (the trade-off between recall and precision).

The various stemming algorithms can be divided into four types: (1) affix removal, (2) successor, (3) table lookup, and (4) N-gram. The first of these stemming algorithms was developed by Lovins (1968), but Porter's (1980) stemming algorithm is easily the most popular. It approaches the task by suffix removal.

Along with the advantages to stemming are its disadvantages, primarily due to over- or under-stemming. Stemming of words may lose the morphological information that can hide the differences in the meaning of two similar words. For example, "gravity" and "gravitation" will both stem to the same root by using Porter's algorithm, yet "gravitation" almost certainly deals with "gravitational force," whereas the meaning of "gravity" depends more on context. Hull and Grefenstette (1966) and Hull (1996) analyze several English stemming algorithms and suggest ways to improve them. Krovetz (1993) also notes that stemming is more of a problem for short text passages than for long ones. To deal with these potential problems, the output of the stemming algorithm was examined to ensure a high level of quality.

2.4. *Weighting of Artifacts*

At this point, a set of stemmed words obtained from the paragraphs and the set of paragraphs themselves are obtained; these sets serve as the starting point for representing and manipulating the text. A corpus of d artifacts (e.g., documents or paragraphs) containing t stemmed words (terms) is represented as a $t \mathbf{X} d$ term-frequency matrix **A**. This data structure is a vector space model (e.g., see Salton, Wong, and Yang 1975). Each term of **A**, a_{ij}, is initially the count (term frequency, *tf*) of stemmed word/term i in artifact/paragraph j. However, not all

artifacts and not all terms are "equal." The artifacts may be of unequal length. To the extent that the difference is due to a more or less verbose use of stop words, this inequality will not affect the process. However, a varying number of themes/ideas in the artifacts will adversely impact the subsequent analysis. One problem that can occur involves the implicit heavier weighting of longer artifacts, such as a 100-page document versus a 10-page document. The problem is that a longer artifact can be expected to have a greater number of words as well as more occurrences of the same words. This may be due to the presence of multiple ideas being expressed, or it could simply be a consequence of a more verbose writing style. As Robertson and Walker (1994: 235) state in their scope versus verbosity hypotheses:

> Some documents may simply cover more material than others...(the "Scope hypothesis"). An opposite view would have long documents like short documents but longer; in other words, a long document covers a similar scope to a short document, but simply uses more words (the "Verbosity hypothesis").

Note also that part of the weighting scheme, inverse-document-frequency (IDF), is sensitive to the specification of artifact boundaries. Using paragraphs as the level of analysis addresses that concern. As Belew (2000: 90) states: "the advantage of using the paragraph as the canonical [artifact]... and/or relying on all [artifacts] in the corpus to be of nearly uniform size... is apparent." (See also the OKPAI system, Robertson 1997.) Therefore, in the interest of reducing the effects of artifact length variation and to better perform the types of analyses anticipated, the artifact under consideration will be a paragraph. Naturally, there may be marked variance among the number of paragraphs in any artifact, but if each paragraph expresses a single concept/thought, then it will serve well as the level of granularity.[2] Therefore, paragraphs should be a natural level of analysis, given the need to categorize single thoughts in a larger context.[3]

[2]Paragraphs can also vary in length, but that variance is typically much smaller than at the document level.

[3]In the future, a hierarchical approach will be developed to capture the semantics of a document as a whole in addition to its paragraphs individually.

The stemmed words occur with varying frequencies. Two extreme examples are (1) a term that appears in only one paragraph and (2) a term that appears the same number of times in each paragraph. Zipf's work (1949) demonstrated that the rank-frequency distribution of the terms in a corpus can be closely approximated by the equation $F(t_r) = C/r^{\alpha}$ for term t where r is frequency rank of term t, $C \approx 0.1$, and $\alpha \approx 1$. (Mandelbrot [1983] generalized this type of phenomenon as fractals.) Not only is this approximation true for the entire set of words in the corpus but also for such subsets as stemmed words and nouns (Belew 2000). Due to this variability, we weighted the a_{ij}.

A variety of term-weighting schemes are in use, dating back to Luhn's work in 1957, and reviewed by Salton and Buckley (1998). With the advent of the World Wide Web, many researchers are employing and exploring weighting schemes to facilitate retrieval. In general, they all consider global effects (g; representing the importance of the term across all artifacts), local effects (l, representing the importance of the term relative to other terms in the artifact), and normalization (n; forcing the length of each column to be 1). Thus $a_{ij} = l_{ij} g_i n$ for the ith term in the jth artifact. The most commonly used weighting scheme is term-frequency inverse-document-frequency (TFIDF) (Berry and Browne 1999). In TFIDF the local weight, l_{ij}, is the term frequency tf_{ij} (i.e., the number of times a stemmed word appears in an artifact). The global weight, g_i, is the inverse artifact frequency ($idf_i = \log(nDocs/nDocs_i)$, where $nDocs$ is the number of artifacts in the corpus, and $nDocs_i$ is the number of artifacts in which term i appears). The lengths here are normalized to 1. (Husbands, Simon, and Ding [2000] question the use of TFIDF when the corpus is extremely large, above 500,000 documents.) So the elements of **A** are computed by

$$a_{ij} = tf_{ij} * \log(nArtifacts/nArtifacts_i). \quad (1)$$

Another commonly used weighting scheme is log-entropy, where the local weight is the logarithm of the term frequency plus one ($\log(tf_{ij} + 1)$), and the global weight is the entropy of the term across all artifacts, $1 - (\sum(p_{ij} \log(p_{ij}))/\log nDocs)$, where $p_{ij} = tf_{ij}/gf_i$ and gf_i is the number of times term i appears in the corpus (e.g., see, Berry and Browne 1999). In this work, TFIDF is used.

2.5. *Numeric Transformation*

A is a sparse matrix; typically much less than 1 percent of the elements are nonzero. In general, t may be greater than or less than d, and the rank of **A**, $r \leq \min(t, d)$. For simplicity, it is assumed that $t > d$, and that the rank of **A** is d. Singular value decomposition (SVD) can be used to separate **A** into three matrices, **U**, **S**, and **V**, such that $\mathbf{A}_{tXd} = \mathbf{U}_{tXr}\mathbf{S}_{rXr}\mathbf{V}^T_{rXd}$, or more simply

$$\mathbf{A} = \mathbf{USV}^T. \tag{2}$$

U and **V** are orthogonal matrices, and **S** is a diagonal matrix containing the singular values of **A** in decreasing size. **U** contains the left-singular vectors of **A**, and **V** contains the right-singular vectors. The columns of **U** form a basis for the row space of **A**, and the columns of **V** form a basis for the column space of **A**.

In practice, matrix **A** is approximated with $\mathbf{A}_k$ by choosing the first k singular values and the corresponding vectors from **U** and **V**. That is,

$$\mathbf{A}_k = \mathbf{U}_k\mathbf{S}_k\mathbf{V}^T_k, \tag{3}$$

where **U** is $t \times k$, **V** is $d \times k$, and **S** is $k \times k$. This approximation, in principle, discards the noise in the matrix (and it is the closest rank-k approximation to **A** according to a theorem by Eckart and Young [1936]). The question of what value to select for k is still a subject of research in the information retrieval community. Most researchers report using a value between 100 and 300, although it is widely acknowledged that this is merely a heuristic with more anecdotal than analytical support.

Thus, the result of using LSI is a representation of the initial text data, the corpus, in a k-dimensional space such that the essence of the original paragraphs has been extracted "latently." The content of the corpus "is modeled by the geometric relationships between the artifact vectors (columns of $\mathbf{A}_k$), not by the individual components of those vectors" (Berry, Drmac, and Jessup 1999: 350). The paragraphs and the terms are both represented in the same k-dimensional space. With the transformation from raw text to a numeric representation complete, the appropriate analysis may be applied to the transformed data.

2.6. *Statistical Processing*

The next step in the analysis involves performing a cluster analysis on the right-singular vectors, $\mathbf{V}_k$, weighted by the singular values, $\mathbf{S}_k$. First, a proximity matrix of the data is computed using squared Euclidian distance as the dissimilarity measure. For two artifacts i and j in $\mathbf{V}_k$, the weighted distance between them can be determined by the law of cosines as

$$\Delta_{ij}^2 = \sum_{l=1}^{k} (s_{ll} v_{il})^2 + \sum_{l=1}^{k} (s_{ll} v_{jl})^2 - 2 \sum_{l=1}^{k} (s_{ll} v_{il})(s_{ll} v_{jl}) \cos\theta_{ij}, \quad (4)$$

where Δ_{ij}^2 is the squared Euclidean distance, s_{ll} is the lth diagonal element from $\mathbf{S}_k$, v_i and v_j are the two k-dimensional rows from $\mathbf{V}_k$ corresponding to i and j, and θ_{ij} is the angle between them. Recognizing that Δ_{ij}^2 can also be written as $\Delta_{ij}^2 = \sum_{l=1}^{k} s_{ll}^2 (v_{il} - v_{jl})^2$ and rewriting (0.4) yields

$$\cos\theta_{ij} = \frac{\sum_{l=1}^{k} s_{ll}^2 v_{il}^2 + \sum_{l=1}^{k} s_{ll}^2 v_{jl}^2 - \sum_{l=1}^{k} s_{ll}^2 (v_{il} - v_{jl})^2}{2\sqrt{\sum_{l=1}^{k} s_{ll}^2 v_{il}^2}\sqrt{\sum_{l=1}^{k} s_{ll}^2 v_{jl}^2}}. \quad (5)$$

Or, in matrix notation,

$$\cos\theta_{ij} = \frac{\mathbf{v}_i^T \mathbf{S}_k \mathbf{S}_k^T \mathbf{v}_j}{\left\|\mathbf{v}_i^T \mathbf{S}_k \mathbf{S}_k^T \mathbf{v}_i\right\|_2 \left\|\mathbf{v}_j^T \mathbf{S}_k \mathbf{S}_k^T \mathbf{v}_j\right\|_2}, \quad (6)$$

where $\mathbf{v}_i$ and $\mathbf{v}_j$ are rows corresponding to artifacts i and j in $\mathbf{V}_k$.

More generally, the similarity matrix for the artifacts is given by

$$\begin{aligned} \mathbf{A}_k^T \mathbf{A}_k &= (\mathbf{U}_k \mathbf{S}_k \mathbf{V}_k^T)^T (\mathbf{U}_k \mathbf{S}_k \mathbf{V}_k^T) \\ &= \mathbf{V}_k \mathbf{S}_k^T \mathbf{U}_k^T \mathbf{U}_k \mathbf{S}_k \mathbf{V}_k^T \\ &= \mathbf{V}_k \mathbf{S}_k^2 \mathbf{V}_k^T, \end{aligned} \quad (7)$$

which becomes a matrix of cosines when divided by the vector norms as in equation (6). Likewise the similarity matrix for the terms is given by

$$\mathbf{A}_k\mathbf{A}_k^T = \mathbf{U}_k\mathbf{S}_k^2\mathbf{U}_k^T. \quad (8)$$

Clustering is performed on the proximity matrix by using the increase in squared Euclidean distance as the criterion. At each step, items are added to clusters in such a way as to minimize the increase in the distance. For a set of n cases, this results in $n - 1$ fusions as the cases are successively grouped into larger clusters. There are n clusters at time zero, and one when the process terminates. The resulting clustering sequence (tree diagram or dendrogram) is used for the next steps in the process. (See Figure 1 for an example dendrogram.)

As clusters are formed, the centroid of the new cluster is computed in k-space, weighting the contributions of each component to the cluster on the basis of the number of items in the components. The leaves of the tree, the individual artifacts, each have weight 1. As the tree is climbed, the weights increase until the last cluster has weight n. Note that because the partitioning is disjoint, the sum of the weights across a set of clusters is always n.

In addition to recomputing the centroid, the union of all the terms in the components of the cluster is formed, and the number of times the terms occur across all the components in the cluster is summed. For example, if two leaves form a cluster, the new cluster would have a weight of 2. If the first leaf had three stemmed terms in it and the second had five, four of which were different from the first, then the cluster would have seven stemmed terms in it. If at a later point two clusters combine to form a third, the weight of the third would be the sum of the weights of the two, and the terms in the third would be determined as above.

3. WORD IMPACT ANALYSIS

Up to this point, most of the process has been easily derived from previous research. This section describes a new development that allows in-depth analysis of the results from the cluster analysis. At any step in the process, it is important to extract the marginal impact on the cluster of a specific term. That is, when two artifacts or clusters joined together, the importance of the presence of this term in the artifact or cluster should be determined as well as the size of the impact of the term on the value of the similarity measure that caused the two to join.

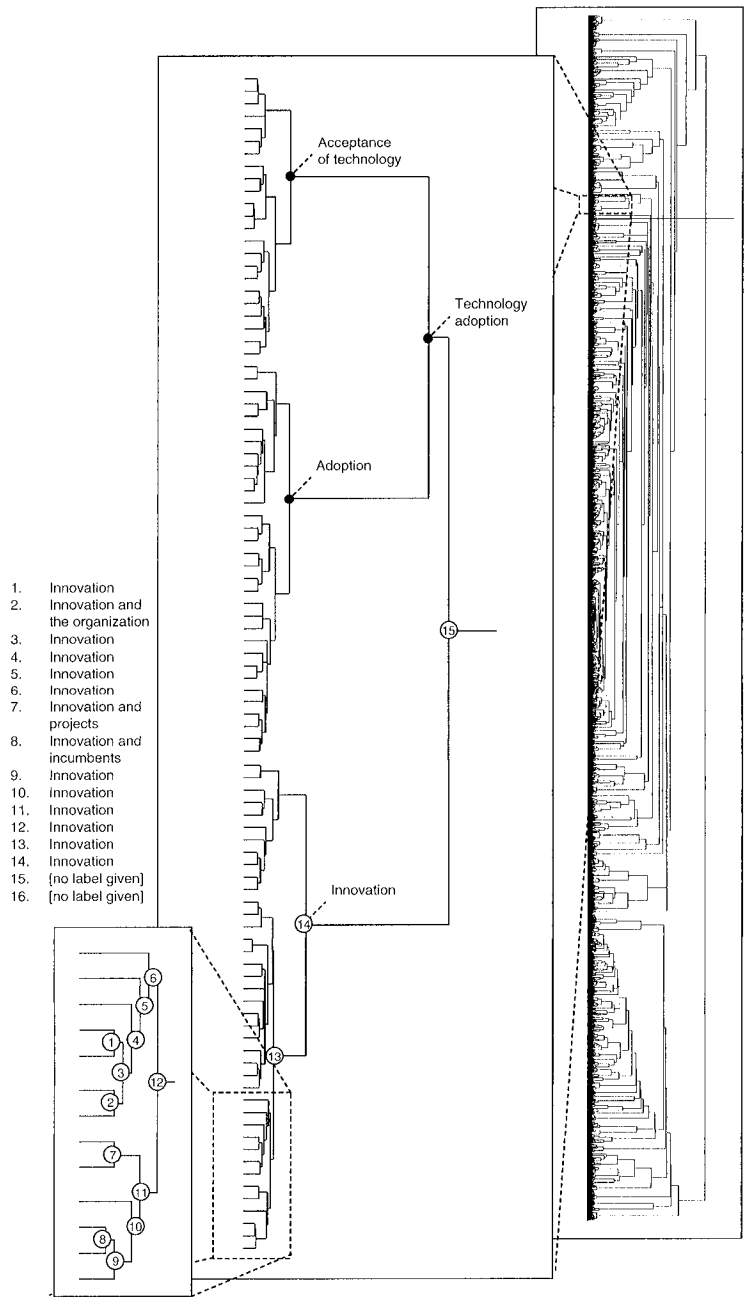

FIGURE 1. Dendrogram of sample similarities.

For simplicity, the process when two artifacts (leaves) are clustered is described, and then the process is extended when two clusters are joined.

First, a pseudodocument (Deerwester et al. 1990), or, in this case, a pseudoartifact, is constructed. The pseudoartifact will have the same term frequencies as one of the existing artifacts that is being clustered, and the same TFIDF weighting transformation as is used on original artifacts is applied. Let $\mathbf{q}_j$ be the pseudoartifact (query) corresponding to artifact j and constructed as above. The jth column of $\mathbf{A}$ as approximated by $\mathbf{A}_k$ in the k-dimensional space may be written as

$$\mathbf{a}_j = \mathbf{U}_k \mathbf{S}_k \mathbf{V}_k^T \mathbf{e}_j, \tag{9}$$

where $\mathbf{e}_j$ is the jth column of a $d \times d$ identity matrix. That is, $\mathbf{V}_k^T \mathbf{e}_j$ is the jth column of the transposed matrix, $\mathbf{v}_{kj}^T$. Rearranging equation (9) yields

$$\mathbf{v}_{kj} = \mathbf{a}_j^T \mathbf{U}_k \mathbf{S}_k^{-1}. \tag{10}$$

So the representation of the pseudoartifact in the k-dimensional space is given by

$$\mathbf{q}_{kj} = \mathbf{q}_j^T \mathbf{U}_k \mathbf{S}_k^{-1} \tag{11}$$

(see Berry, Dumais, and O'Brien 1995: 4).

Then rewriting equation (7) for artifact i and the approximation to artifact j using the pseudoartifact from (0.11), gives

$$\begin{aligned} \mathbf{v}_{ki}^T \mathbf{q}_{kj} &= (\mathbf{e}_i^T \mathbf{V}_k) \mathbf{S}_k^2 \mathbf{q}_{kj}^T \\ &= \mathbf{v}_{ki}^T \mathbf{S}_k^2 (\mathbf{q}_{kj}^T \mathbf{U}_k \mathbf{S}_k^{-1})^T \\ &= (\mathbf{v}_{ki}^T \mathbf{S}_k) \mathbf{S}_k (\mathbf{U}_k \mathbf{S}_k^{-1})^T \mathbf{q}_{kj} \\ &= (\mathbf{v}_{ki}^T \mathbf{S}_k) \mathbf{S}_k \mathbf{S}_k^{-1} \mathbf{U}_k^T \mathbf{q}_{kj} \\ &= (\mathbf{v}_{ki}^T \mathbf{S}_k)(\mathbf{U}_k^T \mathbf{q}_{kj}). \end{aligned} \tag{12}$$

And the cosine between the two is

$$\cos \theta_{ij} = \frac{(\mathbf{v}_{ki}^T \mathbf{S}_k)(\mathbf{U}_k^T \mathbf{q}_{kj})}{\|\mathbf{v}_{ki}^T \mathbf{S}_k\|_2 \|\mathbf{U}_k^T \mathbf{q}_{kj}\|_2}. \tag{13}$$

Before continuing, equation (13) is rewritten for clarity in the subsequent equations, with the understanding that all of this is taking place in the k-dimensional reduced space:

$$\cos\theta_{ij} = \frac{(\mathbf{v}_i^T\mathbf{S})(\mathbf{U}^T\mathbf{q}_j)}{\left\|\mathbf{v}_i^T\mathbf{S}\right\|_2\left\|\mathbf{U}^T\mathbf{q}_j\right\|_2}. \tag{14}$$

To determine the effect of the presence or absence of a term in the pseudoartifact, the numerator of equation (14) is rewritten algebraically:

$$\begin{aligned}(\mathbf{v}_i^T\mathbf{S})(\mathbf{U}^T\mathbf{q}_j) &= \begin{pmatrix} v_{i1} & \cdots & v_{ik}\end{pmatrix}\begin{pmatrix} s_{11} & & 0 \\ & \ddots & \\ 0 & & s_{kk}\end{pmatrix}\begin{pmatrix} u_{11} & \cdots & u_{1t} \\ \vdots & \ddots & \vdots \\ u_{k1} & \cdots & u_{kt}\end{pmatrix}\begin{pmatrix} q_{j1} \\ \cdots \\ q_{jt}\end{pmatrix} \\ &= \begin{pmatrix} s_{11}v_{i1} & \cdots & s_{kk}v_{ik}\end{pmatrix}\begin{pmatrix} \sum_{l=1}^{t} u_{1l}q_{jl} \\ \cdots \\ \sum_{l=1}^{t} u_{kl}q_{jl}\end{pmatrix} \\ &- s_{11}v_{i1}\sum_{l=1}^{t} u_{1l}q_{jl} + \cdots + s_{kk}v_{ik}\sum_{l=1}^{t} u_{kl}q_{jl} \\ &= \sum_{l=1}^{t} q_{jl}\sum_{h=1}^{k} s_{hh}v_{ih}u_{hl}.\end{aligned} \tag{15}$$

The denominator of equation (14) is

$$\sqrt{\sum_{h=1}^{k}(v_{ih}s_{hh})^2}\sqrt{\sum_{h=1}^{k}\left(\sum_{l=1}^{t} u_{hl}q_{jl}\right)^2}. \tag{16}$$

Since $\mathbf{U}$, $\mathbf{S}$, and $\mathbf{V}$ are all known, equation (14) may be rewritten as

$$\cos\theta_{ij} = c_0\frac{\sum_{l=1}^{t} q_{jl}c_i}{\sqrt{\sum_{h=1}^{k}\left(\sum_{l=1}^{t} u_{hl}q_{jl}\right)^2}}, \tag{17}$$

where

$$c_0 = \sqrt{\sum_{h=1}^{k} (v_{ih} s_{hh})^2} = \left\| e_i^T \mathbf{V}_k \mathbf{S}_k \right\|_2$$
$$c_i = \sum_{h=1}^{k} s_{hh} v_{ih} u_{hl} = e_i^T \mathbf{V}_k \mathbf{S}_k \mathbf{U}_k \tag{18}$$

and e_i^T is defined as before.

From equation (17), it may be seen that the impact of the presence or absence of a specific term in the query, q_{jl}, on $\cos\theta_{ij}$ is additive in the numerator but not in the denominator. Since the primary interest here is in the difference rather than the rate of change, partial derivatives will not solve the problem. Consequently, the strategy adopted is to compute $\cos\theta_{ij}$ as well as $\cos\theta_{ij_l}$, the cosine without the lth term in the query. Obviously the smaller the difference, the less the term contributes to the magnitude of the cosine. This difference may be seen as a proportion of $\cos\theta_{ij}$, with the caveat that the proportions will not add to 1 due to the fact that q_{jl} appears in both the numerator and denominator of equation (17). The result is labeled p_{ij_l}

$$p_{ij_l} = (\cos\theta_{ij} - \cos\theta_{ij_l}) / \cos\theta_{ij} \tag{19}$$

For a given artifact i and query (pseudoartifact) j, the p_{ij_l} can be presented in a table of decreasing magnitude. In practice, the first ten terms seem to be more than sufficient to identify the major contributors to the cosine. One final point should be noted. Although the artifact-artifact matrix in equation (7) is symmetric, the matrix of these cosines is not. That is $\cos\theta_{ij} \neq \cos\theta_{ji}$ due to the interaction effects of the term-term matrix. (For a discussion of the transitive effects of terms in LSI, see Kontostathis and Pottenger [2002].) With the major contributors to cosine identified, the researcher may examine the results at each fusion point of the dendrogram and from this build a knowledge structure.

4. APPLICATION OF THE LATENT CATEGORIZATION METHOD

4.1. *Data Collection*

To test the LCM, a set of abstracts was collected. A total of 6135 abstracts distributed according to Table 1 was used as a proxy for a normal qualitative dataset.[4] Abstracts were collected from seven

TABLE 1
Sample Information

Journal	Area	Years	Abstracts
Academy of Management Journal	Management	1996–2001	353
Academy of Management Review	Management	1996–2001	206
Accounting Review	Accounting	1996–2001	148
Administrative Science Quarterly	Management	1996–2001	149
Decision Sciences	Operations research	1994–1999	213
Decision Support Systems	Information systems	1992–2003	242
European Journal of IS	Information systems	1995–2002	117
IEEE Transactions on S.E.	Information systems	1992–2001	29
Information and Management	Information systems	1992–2002	474
Information Systems Research	Information systems	1996–2001	136
Journal of Accounting and Economics	Accounting	1996–2001	156
Journal of Accounting Research	Accounting	2000–2001	41
Journal of Consumer Research	Marketing	1996–2001	184
Journal of Finance	Finance	1996–2001	468
Journal of Financial Economics	Finance	1996–2001	324
Journal of Information Science	Information science	1990–2002	448
Journal of MIS	Information systems	1999–2002	114
Journal of Marketing	Marketing	1996–2001	177
Journal of Marketing Research	Marketing	1996–2001	224
Management Science	Operations research	1998–2001	405
Marketing Science	Marketing	1996–2001	142
MIS Quarterly	Information systems	1996–2001	122
Operations Research	Operations research	1996–2001	264
Organization Science	Management	1992–2002	398
Review of Financial Studies	Finance	1996–2001	213
Strategic Management Journal	Management	1996–2001	388
Total			6135

[4]Business-related journal abstracts rather than sociology-related abstracts were used because of the authors' ability to confirm that results have meaning.

academic areas: (1) accounting, (2) finance, (3) information science, (4) information systems, (5) management, (6) marketing, and (7) operations research.[5] The research method should work equally well on transcripts from interviews and other textual data,[6] but abstracts have the useful property of known content available to most readers.

4.2. *Preprocessing and Initial Statistical Analysis*

There were 906,056 words in the 6135 abstracts examined. As the first step in the analysis, all stop words (e.g., *a, an, the, only, but, and, or*) were removed. In general the 800+ stop words are articles, pronouns, prepositions, conjunctions, and common adjectives and adverbs. The second step consisted of removing the acronyms from the text. Typically these were journal abbreviations such as MISQ, JMR, and so forth. By using Porter's stemming algorithm, 6326 distinct stems were identified in the remaining text. The three most frequently occurring words were *model* (5006 times), *information* (4690 times), and *firm* (4551 times). The most broadly used words were *result* (in 2362 abstracts), *study* (in 2321 abstracts), and *paper* (in 2065 abstracts). Depending on the purpose of a given study, such frequent words as *model, result, study*, and *paper* may have been removed. Due to the exploratory nature of this examination, however, they were retained.

Using the approach outlined in Section 2, the data were processed and cluster analyzed using a hierarchical cluster-analytic process with increase in sum of squares as the criterion. (Specifically, the Centroid method in the software package CLUSTAN was used.) Figure 1 displays the dendrogram of similarities among the 6135 abstracts (right part of figure). The middle cutout displays a small subset of 95 abstracts and their relationships, which after analysis are shown to cluster into two major categories, *technology adoption* and *innovation*. A further cutout (left part of figure) shows the relationships among 13 abstracts in the *innovation* category.

[5]While this sample was collected to examine the method, researchers attempting to understand the structure of a specific area or answer a specific research question should use standard sampling rules found in Babbie (1998).

[6]This will depend on the characteristics of the textual data in question, and further research is required to find for which types of textual data the approach works best.

4.3. *Text Content Analysis*

To follow the analysis fully, the reader is referred to Appendix A, where details about numbers used in the decisions leading to naming of clusters for fusion points 1–16 are available.

Table 2, copied from Appendix A, shows the relationship between two abstracts from the journal *Organization Science* (Cheng and VandeVen 1996; Repenning 2002) that clustered together (as fusion point 1). Important features include a left- and a right-side analysis because the numbers vary slightly (as explained in Section 3). Each side displays the cosine from that side to the other side, which is the main number determining similarity between abstracts. The number of words displayed reflect unique words only, and the number will therefore not grow fast when abstracts are combined. Further, for space reasons, only the top five words are displayed. However, five words were literally always found to be adequate in terms of explaining the similarity between the abstracts. Focusing on the left-side analysis, note that had the word *innovation* been removed from the abstract, the cosine would drop from .5778 to .2367, showing that fully 59.03 percent of the cosine is due to this word. In contrast, the second word, *theory*, explained only 1.89 percent of the cosine, and it is reasonable to argue that this is not enough to include this word in the title of the fusion point. This decision is further supported by the finding that another word occupied the second word location of the right-side analysis.

TABLE 2
Example Result from Appendix A: Fusion Point 1

Left Analysis				**Right Analysis**	
Cosine: .5778–26 different stems			Cosine: .5781–78 different stems		
Stemmed term	**Cosine Without Term**	**% Cosine Explained**	**% Cosine Explained**	**Cosine Without Term**	**Stemmed Term**
Innovation	.2367	59.03	56.51	.2514	Innovation
Theory	.5669	01.89	08.48	.0490	Process
Organization	.5670	01.87	02.31	.0134	Organization
Empirical	.5731	01.82	02.26	.0130	Theory
Provide	.5736	00.73	00.77	.0045	Character

In fusion point 2 (see Appendix A), it may be seen that two words, *innovation* and *organization*, are both major contributors to the cosine, and they were therefore both used in the title of that fusion point. Fusion points 4 and 5 exhibit cases in which the decision to include the second word is a judgment call. However, following Richards and Richards' (1995) principle 1, "the children of a category should be cases in the same sense of the parent," these words were not included in the title. The next fusion point of interest is 15. Note that the cosine for fusion point 15 is down significantly, but not more so than fusion point 10. However, the same word does not appear at the top of both analyses for point 15. Furthermore, knowledge of the field indicates that treating these two fusion points as separate clusters may make sense for the remaining analysis. Fusion point 16 is not displayed in Figure 1, but it is the point at which the whole midsection (fusion point 15) is compared with the closest other cluster (in this case, a cluster named CASE).[7]

4.4. *Postanalysis*

Because conducting the above analysis for every single fusion point in the dendrogram can be time-intensive, especially for large data sets, a system was devised to start the analysis at a higher point in the dendrogram, specifically at the point where each cluster had an average of five leaf-nodes. At this point, there were 1227 clusters that were treated as the starting point for analysis. This increased the speed of analysis fivefold without sacrificing a significant degree of quality, as confirmed by a later examination of the clusters. Appendix B contains one result of this analysis, a 201-cluster solution showing only those clusters containing five or more leaf-nodes. As may be seen, the solution adheres to Richards and Richards' (1995) principle 3, suggesting that a topic should exist in only one place in the indexing system. To simplify the explanation, a smaller solution was extracted, with only those clusters containing $\geq$35 leaf-nodes. The similarity matrix for this

[7]Research on an information system specific research area, computer-aided software Engineering tools.

solution may be found in Appendix C. For expositional ease, multidimensional scaling (MDS) (Kruskal 1964; McLaughlin, Carnevale, and Lim 1991) was performed on the similarity matrix with a highly acceptable stress score of .19 and $R^2 = .90$, leading to a two-dimensional solution.[8] Figure 2 displays that solution and shows which clusters are similar.

Figure 2 shows many finance, accounting, marketing, and operation research topics spread across the bottom, with a slightly more technical topic, decision support systems (DSS), being quite similar.[9] In the middle, a smattering of information systems topics are joined by

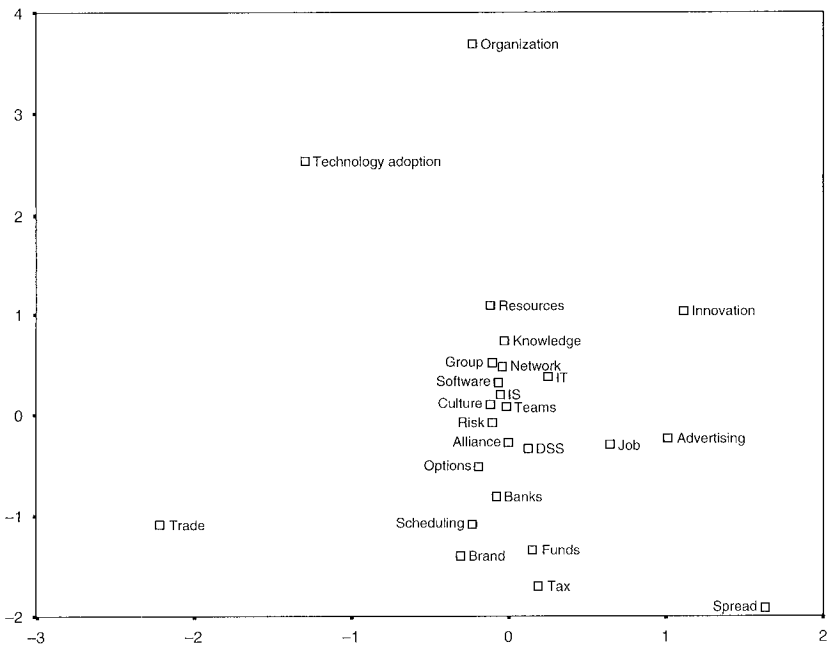

FIGURE 2. Two-dimensional display of 25 clusters.

[8]Note that although the three-dimensional solution displayed slightly better statistics, with a stress score of .16 and $R^2 = .92$, the difference was judged small enough that the simpler two-dimensional solution would be acceptable for this example.

[9]Until the dimensions of the map can be automatically extracted, no dimensional interpretation is attempted.

management topics. Two puzzling findings were the outliers *organization* and *technology adoption*. Further examination would be required to determine why *organization* was an outlier, *technology adoption*'s outlier status (and distance from such clusters as *IT* and *IS*), however, was determined to be an artifact of how *technology adoption* is used in management and operations research papers to specify any kind of technology, whereas the information systems field uses *IT* and *IS* as fairly specific terms mostly referring to computer-related hardware and software.

Table 3 shows how external variables may be used to further examine the findings resulting from the analysis. The *sum* shows how many articles clustered into the 25 categories compared with the total number of articles examined from those areas—in this case, the journals in which articles in the 25 categories were published. Table 1 contains information on area association for each journal. To make definitive statements about the areas and their research interests, the data would need to be more carefully collected. However, in spite of relatively few articles being selected from the marketing area, this area shows a considerable spread in research interests, whereas accounting and finance seem very focused in their research. Similarly, management and IS show a wide spread in topic selection. An artificial sign of wide topic selection may be found in the OR research. Most of this spread was found to due to the interdisciplinary nature of management science and decision sciences. When examining the journal *Operations Research*, only one category contained many papers: *scheduling*.

4.5. *Examination of Word Relationships*

Focus so far has been on developing an overview of the whole data set and its structure. For the method to become workable for qualitative researchers, it must allow the detailed examination of each category. Even though Appendix A contains evidence allowing the examination of some aspects of individual categories, such evidence is limited because it compares subcategories with each other, rather than examining the category as a whole within a larger context. To further examine the individual category, a software tool was developed to allow the examination of word relationships between the category-defining word *innovation* and other words of importance in the data set (Figure 3).

TABLE 3
Relationships Between Categories and Academic Areas

Category	AREA						
	Management	Information System	Operation Research	Accounting	Finance	Marketing	Information Science
1. Advertising	1	1	6			34	1
2. Brand						42	
3. Scheduling		3	32				
4. Group	17	18	2		1	1	
5. Information Systems		38					3
6. Knowledge	26	19	1			4	2
7. Networks	25	6	3			1	1
8. Technology adoption	9	34	9	3			
9. Innovation	24	4			2	8	1
10. Job	20	8	4			3	
11. Software		30	7			1	1
12. Culture	21	3	1			11	3
13. Resources	32	2		1		2	1
14. Organization	41	4	2				1
15. Decision Support Systems		36	2				
16. Teams	32	8	2	1		2	2
17. Alliances	31		3		1	3	
18. IT	3	68	4				3
19. Funds	2				40		1
20. Tax				24	11		
21. Banks	2	5	2	2	36		
22. Trade	1	6			34	2	
23. Spread				1	39		
24. Options	3	4	6	3	24	1	
25. Risk	13	8	6	2	8	6	
Sum	**303/1494**	**305/1234**	**92/882**	**37/345**	**196/1005**	**121/727**	**20/448**

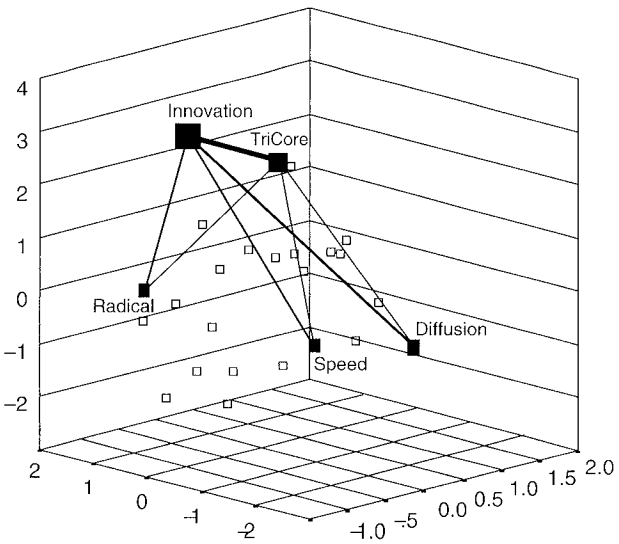

FIGURE 3. Word relationship diagram.

To derive the most important words, *innovation* was used as a starting point, and a list of the 24 other words in the term-term matrix with the highest cosines (degree of similarity) to *innovation* was extracted. With these 25 words (a user may select a higher or lower number), a subset of the term-term matrix was extracted with all relationships between the words. Multidimensional scaling was then used to create the three-dimensional map of the distances between terms in Figure 3 (stress = .12, $R^2 = .89$). Each square represents a term, and the distance map is a result of the whole data set, and as such not just representative of the category in question.

The *innovation* category contained 40 abstracts. A simple count of how many of those 40 abstracts in which a specific term would appear was also extracted. For the activated terms, this is represented by the sizes of the squares in Figure 3. The *innovation* term existed in all 40 abstracts, the *TriCore* term appeared the second most frequently, and *diffusion* was the third most frequent term. These terms were clearly important, but no action was taken based on this count.

For each of the 25 terms selected for the MDS map, the abstracts were examined to find how many times sets of 2 of the 25 terms (300 total sets) occurred together in one of the abstracts. Figure 3 shows the relationships between those terms that occurred together in more than 5

percent of the abstracts; these terms are referred to as *activated* within the context of this category.[10] The activated terms were *innovation, TriCore, radical, diffusion,* and, *speed.* The selection of these terms has a high level of face validity. All are terms related to *innovation.*[11] The strength of relationship is represented by the thickness of lines in Figure 3, with the relationship between *innovation* and *diffusion* and between *innovation* and *TriCore* being the strongest. In the network, *innovation* and *TriCore*, the two innovation-related terms, are the only terms to connect directly to all other activated terms. This makes intuitive sense, given that the Tri-Core model is a theory of innovation, whereas *radical* and *speed* are attributes of innovations. *Diffusion* is another major area of study, and as such relates to the two innovation area terms. The specifics of this network diagram (Figure 3) provide a powerful and visual overview of the important contents of a category in the context of the whole data set.

At the end of the analysis, additional useful information related to the categorization and pattern-matching process is available to the researcher: (1) information on the structure of the data set, in this case displayed through a set of clusters and their distance relationships (see Appendixes A and C); (2) detailed information on each cluster in terms of how the subdendrogram behaved and which words were important in those relationships; and (3) the membership information on each of the data points subjected to the analysis. By examining the available data and findings further, the researcher may select whether a general or detailed level of analysis is desired. This paper scratches only the surface of the rich findings available when using this new method; the reader should have enough information to extrapolate from this example to an in-depth analysis as well as to other data sets.

5. CONCLUSION

The importance of creating and assessing the content of qualitative categories is beyond question. By proposing a new process for the

[10]The 5 percent cut-off point, while somewhat arbitrary, was arrived upon after a careful examination of the characteristics of this data set. A different cut-off point may be needed for other data sets.

[11]The word *TriCore* may be the least known term for those unfamiliar with the innovation literature. The term derives from Tri-Core model of organizational innovations (Swanson 1994).

automatic clustering of text units and for determining what terms were important in that clustering process, this paper may be the first to offer a comprehensive and reproducible approach to not only the clustering of qualitative data but also, and more important, the labeling of clusters through the identification of the relative importance of terms.

The next steps for this research include developing techniques for examining the relationships among categories as well as the content of a category through network diagrams of the most important words in a category. Once the method has been fully developed for inductive approaches, research should be conducted to expand the method such that confirmatory research may be possible. Later work may tie this method into data visualization techniques such as displaying findings and relationships in a virtual reality environment. Further, the method may be tested against manual categorization work, such as Subramani and Walden's (2001) attempt at predicting the market value of firms based on their announcements.

Through techniques this paper has only started to explore, the move from small to large qualitative data sets is facilitated with logarithmic rather than linear or exponential increases in workload growth. Among the uses for the method is one that may be obvious: a replacement or complementary method to co-citation analysis, thus removing the need for access to bibliography data or buying (limited) data from the Social Science Citation Index. Other uses may include determining the structure of qualitative data sets, including interview, secondary data, and open-ended survey information. By combining the resulting categorical structure with such external variables as time, the method may contribute to explanatory research.

Although it is not appropriate for all types of qualitative research, LCM bears promise for ethnographers and case study researchers, as well as researchers looking for patterns in large data sets.[12] In its current form, LCM may be especially appropriate when grounded researchers are looking for patterns where none existed. These early tests provided promising results, and future use will determine LCM's ultimate usefulness.

[12]Further work is required to extract which characteristics of texts (length, type, etc.) are most appropriate for LCM. We acknowledge the help of an anonymous reviewer on this point.

APPENDIX A: WORD SCORES FOR FIGURE 1

This appendix presents an example of what led to the labeling of one specific cluster. Figure 4 contains a cut-out from Figure 1—namely, information on the first 12 fusion points described in this appendix.

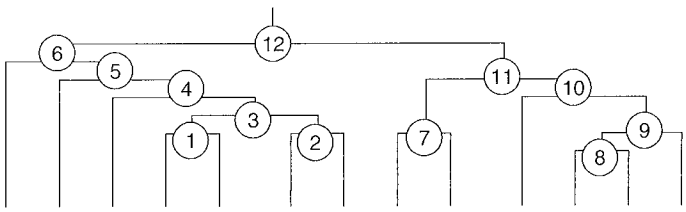

FIGURE 4. Cut-out from Figure 1.

Fusion Point 1: Innovation

Left Analysis			Right Analysis		
Cosine: .5778–26 different stems			Cosine: .5781–78 different stems		
Stemmed Term	**Cosine without Term**	**% Cosine Explained**	**% Cosine Explained**	**Cosine without Term**	**Stemmed Term**
Innovation	.2367	59.03	56.51	.2514	Innovation
Theory	.5669	01.89	08.48	.0490	Process
Organization	.5670	01.87	02.31	.0134	Organization
Empirical	.5731	01.82	02.26	.0130	Theory
Provide	.5736	00.73	00.77	.0045	Character

Fusion Point 2: Innovation and the Organization

Left Analysis			Right Analysis		
Cosine: .5572–55 different stems			Cosine: .5574–68 different stems		
Stemmed Term	**Cosine without Term**	**% Cosine Explained**	**% Cosine Explained**	**Cosine without Term**	**Stemmed Term**
Innovation	.3132	43.80	50.72	.2747	Innovation
Organization	.4732	15.09	15.52	.4709	Organization
Firm	.5353	03.94	03.91	.5356	Firm
High	.5475	01.75	01.02	.5518	Found
Association	.5524	00.87	00.91	.5524	Process

Fusion Point 3: Innovation

Left Analysis			Right Analysis		
Cosine: .6163–112 different stems			Cosine: .6258–83 different stems		
Stemmed Term	**Cosine without Term**	**% Cosine Explained**	**% Cosine Explained**	**Cosine without Term**	**Stemmed Term**
Innovation	.2204	64.24	61.05	.2437	Innovation
Process	.5911	04.10	03.06	.6067	Organization
Theory	.6079	01.36	01.05	.6193	Theory
Learning	.6103	00.98	00.77	.6210	Benefit
Organization	.6105	00.95	00.68	.6216	Paradox

Fusion Point 4: Innovation

Left Analysis			Right Analysis		
Cosine: .6024–43 different stems			Cosine: .6084–177 different stems		
Stemmed Term	**Cosine without Term**	**% Cosine Explained**	**% Cosine Explained**	**Cosine without Term**	**Stemmed Term**
Innovation	.2457	59.21	45.98	.3286	Innovation
Organization	.5500	08.70	13.59	.5257	Organization
Empirical	.5991	00.55	00.85	.6032	Environment
Stream	.5991	00.54	00.80	.6035	Practical
Find	.6	00.35	00.73	.6037	Framework

Fusion Point 5: Innovation

Left Analysis			Right Analysis		
Cosine: .6172–34 different stems			Cosine: .6118–203 different stems		
Stemmed Term	**Cosine without Term**	**% Cosine Explained**	**% Cosine Explained**	**Cosine without Term**	**Stemmed Term**
Innovation	.2247	63.59	52.10	.2931	Innovation
Organization	.5666	08.20	05.38	.5789	Organization
Process	.6090	01.81	05.38	.5862	Implement
Outcome	.6126	01.17	01.17	.6047	Commitment
Climate	.6133	00.60	00.56	.6081	Model

Fusion Point 6: Innovation

Left Analysis			Right Analysis		
Cosine: .5470–222 different stems			Cosine: .5436–54 different stems		
Stemmed Term	**Cosine without Term**	**% Cosine Explained**	**% Cosine Explained**	**Cosine without Term**	**Stemmed Term**
Innovation	.3139	42.62	59.21	.2217	Innovation
Organization	.5323	02.70	06.66	.5074	Organization
Variable	.5334	02.49	01.31	.5365	Theory
Theory	.5335	02.47	00.96	.5384	Firm
Firm	.5362	01.97	00.59	.5405	Model

Fusion Point 7: Innovation and projects

Left Analysis			Right Analysis		
Cosine: .6656–62 different stems			Cosine: .6652–65 different stems		
Stemmed Term	**Cosine without Term**	**% Cosine Explained**	**% Cosine Explained**	**Cosine without Term**	**Stemmed Term**
Innovation	.4388	34.07	35.23	.4309	Innovation
Project	.4846	27.20	19.62	.5347	Project
Management	.6414	03.63	04.21	.6372	Management
Control	.6425	03.46	00.66	.6608	Develop
Size	.6624	00.48	00.52	.6617	Study

Fusion Point 8: Innovation and incumbents

Left Analysis			Right Analysis		
Cosine: .6551–50 different stems			Cosine: .6527–86 different stems		
Stemmed Term	**Cosine without Term**	**% Cosine Explained**	**% Cosine Explained**	**Cosine without Term**	**Stemmed Term**
Innovation	.4715	28.02	25.42	.4868	Incumbent
Incumbent	.5138	21.56	07.09	.6064	Innovation
Firm	.6357	02.96	06.00	.6135	Product
Product	.6376	02.66	02.70	.6351	Market
Author	.6453	01.50	02.18	.6385	Firm

Fusion Point 9: Innovation

Left Analysis			Right Analysis		
Cosine: .6618–124 different stems			Cosine: .6330–40 different stems		
Stemmed Term	**Cosine without Term**	**% Cosine Explained**	**% Cosine Explained**	**Cosine without Term**	**Stemmed Term**
Innovation	.4161	37.12	38.87	.3869	Innovation
Firm	.5978	09.68	07.85	.5833	Product
Product	.6168	06.81	06.06	.5946	Radical
Radical	.6369	03.76	03.58	.6103	Firm
Author	.6467	02.28	01.87	.6211	Author

Fusion Point 10: Innovation

Left Analysis			Right Analysis		
Cosine: .4174–38 different stems			Cosine: .4233–152 different stems		
Stemmed Term	**Cosine without Term**	**% Cosine Explained**	**% Cosine Explained**	**Cosine without Term**	**Stemmed Term**
Innovation	.1084	74.04	70.82	.1235	Innovation
Diffuse	.4122	01.25	01.78	.4157	Industry
Industry	.4124	01.19	01.41	.4173	View
Special	.4130	01.05	01.11	.4186	Article
Strategy	.4133	00.98	00.88	.4196	Complex

Fusion Point 11: Innovation

Left Analysis			Right Analysis		
Cosine: .5322–181 different stems			Cosine: .5401–119 different stems		
Stemmed Term	**Cosine without Term**	**% Cosine Explained**	**% Cosine Explained**	**Cosine without Term**	**Stemmed Term**
Innovation	.1845	65.34	70.68	.1584	Innovation
Radical	.5228	01.76	02.55	.5263	Radical
Industry	.5243	01.48	01.67	.5311	Firm
Size	.5275	00.87	01.34	.5329	Industry
Large	.5301	00.39	00.68	.5364	Literature

Fusion Point 12: Innovation

Left Analysis			Right Analysis		
Cosine: .7077–268 different stems			Cosine: .7289–252 different stems		
Stemmed Term	**Cosine without Term**	**% Cosine Explained**	**% Cosine Explained**	**Cosine without Term**	**Stemmed Term**
Innovation	.2427	65.71	64.44	.2592	Innovation
Organization	.6934	02.03	01.90	.7150	Firm
Firm	.7024	00.76	00.46	.7255	Industry
Radical	.7027	00.70	00.43	.7257	New
Variable	.7047	00.46	00.41	.7261	Variable

Fusion Point 13: Innovation

Left Analysis			Right Analysis		
Cosine: .8812–491 different stems			Cosine: .8878–447 different stems		
Stemmed Term	**Cosine without Term**	**% Cosine Explained**	**% Cosine Explained**	**Cosine without Term**	**Stemmed Term**
Innovation	.3677	58.27	63.57	.3235	Innovation
Firm	.8709	01.17	00.88	.8800	Firm
Organization	.8727	00.96	00.70	.8816	Product
Project	.8789	00.25	00.48	.8835	Radical
Radical	.8801	00.20	00.47	.8837	Technology

Fusion Point 14: Innovation

Left Analysis			Right Analysis		
Cosine: .7050–750 different stems			Cosine: .6846–430 different stems		
Stemmed Term	**Cosine without Term**	**% Cosine Explained**	**% Cosine Explained**	**Cosine without Term**	**Stemmed Term**
Innovation	.3717	47.28	60.20	.2725	Innovation
Adopt	.6855	02.78	01.22	.6763	Technology
Diffuse	.6970	01.14	01.00	.6777	Product
Technology	.6991	00.85	00.87	.6787	Organization
Organization	.6998	00.59	00.62	.6804	Speed

Fusion Point 15: [no label given]

Left Analysis			Right Analysis		
Cosine: .4363–939 different stems			Cosine: .4493–1006 different stems		
Stemmed Term	**Cosine without Term**	**% Cosine Explained**	**% Cosine Explained**	**Cosine without Term**	**Stemmed Term**
Adopt	.3558	18.44	22.15	.3498	Innovation
Technology	.3905	10.51	06.30	.4210	Adopt
Diffuse	.4230	03.04	02.94	.4361	Technology
Firm	.4288	01.71	02.00	.4403	Diffuse
Model	.4310	01.21	01.75	.4414	Firm

Fusion Point 16: [no label given]

Left Analysis			Right Analysis		
Cosine: .2945–1408 different stems			Cosine: .2789–322 different stems		
Stemmed Term	**Cosine without Term**	**% Cosine Explained**	**% Cosine Explained**	**Cosine wwithout Term**	**Stemmed Term**
Adopt	.2351	20.15	21.13	.2199	Adopt
Perceive	.2751	06.58	10.28	.2502	Innovation
CASE	.2826	04.03	06.21	.2615	Technology
Accept	.2858	02.95	03.97	.2678	Perceive
Ease	.2868	02.62	02.05	.2731	Accept

1: Innovation process

(Cheng and VandeVen 1996)

This paper reports the first empirical findings demonstrating the presence of chaotic processes in innovation process. The paper provides an empirical basis for distinguishing between orderly periodic stages in the innovation process, random sequences, and chaotic patterns. The finding that the initial innovation development process exhibits chaotic patterns will have very important implications for theories of organization learning and for the structuring of "exploration" processes in organizations.

and

(Repenning 2002)

The history of management practice is filled with innovations that failed to live up to the promise suggested by their early success. A paradox currently facing organizational theory is that the failure of these innovations often cannot be attributed to an intrinsic lack of efficacy. To resolve this paradox, in this paper I study the process of innovation implementation. Working from existing theoretical frameworks, I synthesize a model that describes the process through which participants in an organization develop commitment to using a newly adopted innovation. I then translate that framework into a formal model and analyze it using computer simulation. The analysis suggests three new constructs—reversion, regeneration, and the motivation threshold—characterizing the dynamics of implementation. Taken together, the constructs provide an internally consistent theory of how seemingly rational decision rules can create the apparent paradox of innovations that generate early results but fail to produce sustained benefit.

2: Innovation and the organization

(Sorensen and Stuart 2000)

This paper investigates the relationship between organizational aging and innovation processes to illuminate the dynamics of high-technology industries, as well as to resolve debates in organizational theory about the effects of aging on organizational functioning. We test hypotheses based on two seemingly contradictory consequences of aging for organizational innovation: that aging is associated with increases in firms' rates of innovation and that the difficulties of keeping pace with incessant external developments causes firms' innovative outputs to become obsolete relative to the most current environmental demands. These seemingly contradictory outcomes are intimately related and reflect inherent tradeoffs in organizational learning and innovation processes. Multiple longitudinal analyses of the relationship between firm age and patenting behavior in the semiconductor and biotechnology industries lend support to these arguments.

and

(Whittington, Pettigrew, Peck, Fenton, and Conyon 1999)

This paper addresses three weaknesses in the literature on new organizational forms: the limited mapping of the extent of contemporary organizational change; confusion about how contemporary changes link together; and the lack of systematic testing of the performance

consequences of this kind of change. Drawing on a large-scale survey of organizational innovation in European firms, the paper finds widespread but not revolutionary change in terms of organization structure, processes, and boundaries. Using the economics notion of complementarities, the paper develops contingency and configurational approaches to suggest that organizational innovations will tend to cluster in particular ways and that the performance benefits of these innovations depend on their clustering. Complementarities in performance are explored from both inductive and deductive perspectives. Consistent with the expectations of complementarity theory, high-performing firms appeared to be innovating more and differently than low-performing firms. Again consistent with complementarities, piecemeal changes—with the exception of IT—were found to deliver little performance benefit, while exploitation of the full set of innovations was associated with high performance. Though few European firms were found to exploit the complementarities of new organizational practices, those that did enjoyed high-performance premia.

3: Innovation
1: Innovation process
and
2: Innovation and the organization

4: Innovation
(Mone, McKinley, and Barker 1998)
An examination of the diverse literature on organizational decline shows that there is disagreement regarding the effects of decline on innovation. Some research streams suggest that organizational decline interferes with an organization's capacity to innovate, whereas other research implies just the opposite: organizational decline stimulates innovation. In this article we integrate the inconsistent perspectives and findings in these research streams by developing a contingency model. The model identifies variables at the environmental, organizational, and individual levels of analysis that determine whether organizational decline inhibits or stimulates innovation. We summarize the moderating effects of these variables with empirically testable propositions and discuss implications of the framework for future research and management practice.

and
3: Innovation

5: Innovation
(Klein and Sorra 1996)
Implementation is the process of gaining targeted organizational members' appropriate and committed use of an innovation. Our model suggests that implementation effectiveness—the consistency and quality of targeted organizational members' use of an innovation—is a function of (a) the strength of an organization's climate for the implementation of that innovation and (b) the fit of that innovation to targeted users' values. The model specifies a range of implementation outcomes (including resistance, avoidance, compliance, and commitment); highlights the equifinality of an organization's climate for implementation; describes within- and between-organizational differences in innovation-values fit; and suggests new topics and strategies for implementation research.
and
4: Innovation

6: Innovation
(Monge, Cozzens, and Contractor 1992)
This paper reports on research designed to test a dynamic model of the causes of organizational innovation. Two communication variables (level of information and group communication) and three motivational variables (perceptions of equity, expectations of benefits, and perceived social pressure) were derived from equity theory, expectancy theory and the theory of reasoned action. These variables were used to predict the number of innovative ideas contributed by members of the organizations. Weekly data were collected for over a year from five firms and were analyzed with multivariate time series techniques. The results indicated that the communication variables were causes of organizational innovation but the motivational variables were not. Across the five firms, the variance explained by the model ranged from a low of 30 percent to a high of 78 percent. In four of the five firms, the forecast accuracy for the amount of individual innovation ranged from a low of 77 percent to a high of 85 percent.

and
5: Innovation

7: Innovation and projects
(Libutti 2000)
The paper reports on the work and the results of the IMTIC (Innovation Management Techniques for Industry Clusters) Project supported by the European Commission's Innovation Programme. The purpose of the project was to make industrial small and medium-sized enterprises aware of the possibilities offered by innovation management techniques (IMTs) in planning and implementing long-term business strategies. IMTs were presented and outlined to a number of industrial clusters in five Italian regions in the areas of: (i) marketing of innovation, (ii) Technology Watch (TW), (iii) Technology Search (TS), (iv) management of intellectual property rights and (v) quality management. The Institute for Studies on Research and Scientific Documentation (ISRDS) of the National Research Council (CNR) of Italy was one of the subcontractors for the project. The main task assigned to ISRDS/CNR was to set up the methodological framework for two innovative techniques: TW and TS. In particular, the methodological process for setting up TW and TS is described. Monitoring by ISRDS of the project phases and the control of the results are also described.
and
(Cardinal 2001)

The literature on the management of R&D professionals strongly advocates managing R&D projects on a project-by-project basis. This literature suggests that projects should be managed differently depending upon project characteristics such as risk, ambiguity, and nonroutineness. While the primary emphasis of the R&D professional literature has been on project teams, the purpose of this study is to examine the impact of organization-wide controls on innovativeness at the firm level. In a sample of 57 pharmaceutical firms, this study investigates the influence of organizational controls on the research and development activities of R&D professionals. This study is one of a handful of studies that simultaneously explores the use of input, behavior, and output controls. Two categories of innovation are considered as dependent variables: incremental innovations in the form of drug enhancements and radical innovations in the form of

new drugs. Contrary to existing theory and hypotheses developed in this study, the results show that input, behavior, and output control enhanced radical innovation, and input and output controls enhanced incremental innovation. These results challenge several important features of existing models of R&D management and diverge from common beliefs about R&D management at the project level. While it is commonly accepted that incremental and radical innovation should be managed differently, the results of this study suggest otherwise. In this instance, the management of R&D activities may be considered more similar than previously thought.

8: Innovation and incumbents

(Kuester, Homburg, and Robertson 1999)

In this article, the authors focus on the defense strategies that firms pursue when threatened by rival new products in their markets. They investigate retaliation as a multidimensional construct. The integrative framework combines the analysis of the marketing instrument used to react and the speed and breadth of retaliation. Results emphasize the importance of the rival product's innovativeness in generating a reciprocal retaliation (a move in kind), though innovativeness slows the incumbent's reaction time. Market growth encourages rapid retaliation, especially on the product mix, whereas in concentrated markets, firms react less strongly on the product mix and exhibit slower reactions. The study also captures the phenomenon of incumbent inertia: Larger incumbents retaliate less strongly and more slowly.

and

(Chandy and Tellis 2000)

A common perception in the field of innovation is that large, incumbent firms rarely introduce radical product innovations. Such firms tend to solidify their market positions with relatively incremental innovations. They may even turn away entrepreneurs who come up with radical innovations, though they themselves had such entrepreneurial roots. As a result, radical innovations tend to come from small firms, the outsiders. This thesis, which we term the "incumbent's curse," is commonly accepted in academic and popular accounts of radical innovation. This topic is important, because radical product innovation is an engine of economic growth that has created entire industries and brought down giants while catapulting small firms to

market leadership. Yet a review of the literature suggests that the evidence for the incumbent's curse is based on anecdotes and scattered case studies of highly specialized innovations. It is not clear if it applies widely across several product categories. The authors reexamine the incumbent's curse using a historical analysis of a relatively large number of radical innovations in the consumer durables and office products categories. In particular, the authors seek to answer the following questions: (1) How prevalent is this phenomenon? What percentage of radical innovations do incumbents versus nonincumbents introduce? What percentage of radical innovations do small firms versus large firms introduce? (2) Is the phenomenon a curse that invariably afflicts large incumbents in current industries? Is it driven by incumbency or size? and (3) How consistent is the phenomenon? Has the increasing size and complexity of firms over time accentuated it? Does it vary across national boundaries? Results from the study suggest that conventional wisdom about the incumbent's curse may not always be valid.

9: Innovation
8: Innovation and incumbents
and
(Chandy and Tellis 1998)
Why are some firms more successful at introducing radical product innovations than others? Following Schumpeter (1942), many researchers have suggested that firm size is the key organizational predictor of radical product innovation. The authors provide an alternate view and argue that one key variable that differentiates firms with strong radical product innovation records from others is the firms' willingness to cannibalize their own investments. The authors identify three organizational factors that drive a firm's willingness to cannibalize. Results from a survey of three high-tech industries tend to support the alternate view that willingness to cannibalize is a more powerful driver of radical product innovation than firm size is. These results suggest a need to reconsider conventional wisdom on firm size, cannibalization, and organizational synergy.

10: Innovation
(Drazin and Schoonhoven 1996)
Two distinct themes emerge from the Special Research Forum on Innovation and Organizations. One group of articles develops an

expanded view of the influence of context on organizations' ability to innovate. Together, the articles offer a complex multilevel view of context as including elements ranging from the dominant strategy of an organization to the social-psychological antecedents of creativity. A second group of articles provides a community and population perspective on the diffusion of innovations. We suggest the possibility of a union between the context and industry dynamics approaches.
and
9: Innovation

11: Innovation
10: Innovation
and
7: Innovation and projects

12: Innovation
11: Innovation
and
6: Innovation

13: Innovation
12: Innovation
and
:Innovation

14: Innovation
13: Innovation
and
:Innovation adoption

15: [no label given]
14: Innovation
:Technology adoption

16: [no label given]
15: [no label given]
:CASE

APPENDIX B

This appendix lists the names of all clusters that contained five or more abstracts. A total of 201 such clusters were discovered. It should be noted that because of the interdisciplinary nature of the data set, not all cluster names will make sense to all readers. Excluded from the listing is the detail structure of each cluster's children. Note that Figure 1 presents an example detail listing of clusters 49 (technology adoption) and 50 (innovation).

1. Advertising
2. Segments
3. Promotion
4. Preference
5. Attributes and preferences
6. Consumption
7. Consumers
8. Purchasing
9. Goals
10. Emotional
11. Brand
12. Store
13. EDI
14. Supply chains
15. Inventories
16. Retail
17. Demand
18. Discount
19. Stage
20. Capacity
21. Scheduling
22. Algorithms
23. Bound
24. Trust
25. Citations
26. Journal
27. GSS
28. Communication
37. Knowledge
38. Neural networks
39. Networks
40. Users
41. End users
42. Instruments
43. Databases
44. Documents
45. Retrieval
46. Indexing
47. Web
48. Searching
49. Technology adoption
50. Innovation
51. Subsidiaries
52. Customer service
53. Wait time
54. Quality
55. Service
56. Customers
57. Satisfaction
58. Job
59. Software
60. Projects
61. Mergers
62. Acquisition
63. Executives
64. Learning
72. Resources
73. Industry and competition
74. Competition
75. Object oriented
76. Orientation
77. Employees
78. Leadership and leaders
79. Agents
80. Cognition
81. Decisions
82. Ethics
83. Environment
84. Product lines
85. New product success
86. Products
87. Methods
88. Methodologies
89. Components
90. Design
91. Social structure
92. Capital
93. Identity
94. Manufacturing
95. Implementation
96. Technology
97. Change
98. Institutions
99. Internet

29. GDSS, group
30. Consensus
31. Meeting
32. Group
33. Diversity
34. Outsourcing
35. IS
36. Planning
65. Ventures
66. Experts, ES
67. Culture
68. Suppliers
69. Incumbents and entrance
70. Entries
71. Collaboration
100. Attitudes
101. Information
102. Libraries
103. Journal research
104. Profession
105. Franchises
106. Industry
107. Control
108. Rules
109. Order
110. Techniques
111. Justice
112. Participation
113. Incentives
114. Companies
115. Tasks
116. Reviews
117. Paradigm
118. Economies
119. Business strategies
120. Competitors
121. Firms
122. Data
123. Models
124. Power
125. Benefits
126. Organization
127. Process
128. Auction
129. Efficiency
130. Market
131. Function
132. Patent
133. Group members
134. Costs
135. Relationships
136. Responses
137. Corporate
138. Family
139. Conflict
140. Stakeholders
141. BPR
142. Creativity
143. DSS
144. Teams
145. Alliances
146. IT
147. Funds
148. Tax
149. Dividends
150. Banks
151. Disclosure
152. Liquidity
153. Traders
154. Trade
155. Insiders
156. Spread
157. Orders
158. Specialists
159. Audits
160. Auditors
161. Analysts
162. Forecasting
163. Volatility
164. Options
165. IPOs
166. Auctions
167. Cash flow
168. Earnings
169. Diversification
170. Ownership
171. Debt
172. Risk
173. Rates
174. Foreign
175. Investments
176. Investors
177. Portfolios
178. Returns
179. Stock
180. Takeovers
181. Shareholders
182. Announcements
183. Estimation
184. Value
185. Insurance
186. Accounting
187. Country
188. Bundling
189. Contracts
190. Price
191. Transfer
192. Issues
193. Selling
194. Transaction costs
195. Governance
196. Financing
197. Public
198. Sales
199. Assets
200. Boards
201. Compensation

Appendix C:
Similarity Matrix for ≥35 Leaf-Nodes Solution

Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. Advertising	1.00	0.62	0.90	0.87	0.90	0.90	0.91	0.86	0.88	0.89	0.88	0.86	0.87	0.82	0.92	0.92	0.93	0.91	0.93	0.95	0.93	0.88	0.92	0.91	0.89
2. Brand	0.62	1.00	0.96	0.92	0.93	0.92	0.93	0.88	0.91	0.93	0.91	0.87	0.91	0.88	0.94	0.94	0.94	0.93	0.95	0.96	0.95	0.91	0.93	0.92	0.92
3. Scheduling	0.90	0.96	1.00	0.91	0.93	0.91	0.89	0.91	0.93	0.68	0.88	0.94	0.85	0.88	0.88	0.93	0.96	0.94	0.95	0.96	0.96	0.92	0.94	0.91	0.92
4. Group	0.87	0.92	0.91	1.00	0.87	0.81	0.83	0.80	0.84	0.88	0.82	0.87	0.81	0.74	0.86	0.78	0.90	0.88	0.93	0.94	0.92	0.91	0.93	0.92	0.89
5. IS	0.90	0.93	0.93	0.87	1.00	0.86	0.90	0.78	0.82	0.87	0.84	0.87	0.83	0.79	0.89	0.88	0.93	0.83	0.94	0.96	0.94	0.92	0.94	0.92	0.89
6. Knowledge	0.90	0.92	0.91	0.81	0.86	1.00	0.79	0.80	0.81	0.91	0.86	0.84	0.73	0.74	0.86	0.86	0.85	0.84	0.94	0.95	0.94	0.91	0.94	0.92	0.90
7. Networks	0.91	0.93	0.89	0.83	0.90	0.79	1.00	0.81	0.81	0.90	0.88	0.86	0.80	0.74	0.91	0.86	0.81	0.87	0.94	0.94	0.90	0.90	0.93	0.91	0.90
8. Technology adoption	0.86	0.88	0.91	0.80	0.78	0.80	0.81	1.00	0.56	0.82	0.79	0.81	0.77	0.65	0.84	0.87	0.86	0.68	0.92	0.93	0.88	0.85	0.88	0.88	0.83
9. Innovation	0.88	0.91	0.93	0.84	0.82	0.81	0.81	0.56	1.00	0.89	0.84	0.86	0.80	0.69	0.91	0.87	0.86	0.84	0.94	0.94	0.89	0.90	0.93	0.90	0.88
10. Job	0.89	0.93	0.68	0.88	0.87	0.91	0.90	0.82	0.89	1.00	0.89	0.88	0.88	0.81	0.92	0.89	0.94	0.90	0.95	0.96	0.94	0.92	0.94	0.93	0.92
11. Software	0.88	0.91	0.88	0.82	0.84	0.86	0.88	0.79	0.84	0.89	1.00	0.87	0.85	0.78	0.86	0.88	0.91	0.86	0.93	0.94	0.92	0.89	0.92	0.89	0.82
12. Culture	0.86	0.87	0.94	0.87	0.87	0.84	0.86	0.81	0.86	0.88	0.87	1.00	0.84	0.63	0.91	0.82	0.91	0.87	0.95	0.95	0.95	0.91	0.93	0.91	0.87
13. Resources	0.87	0.91	0.85	0.81	0.83	0.73	0.80	0.77	0.80	0.88	0.85	0.84	1.00	0.68	0.88	0.88	0.83	0.80	0.91	0.93	0.90	0.89	0.92	0.88	0.86
14. Organization	0.82	0.88	0.88	0.74	0.79	0.74	0.74	0.65	0.69	0.81	0.78	0.63	0.68	1.00	0.82	0.78	0.82	0.77	0.90	0.91	0.86	0.85	0.88	0.87	0.81
15. DSS	0.92	0.94	0.88	0.86	0.89	0.86	0.91	0.84	0.91	0.92	0.86	0.91	0.88	0.82	1.00	0.92	0.95	0.90	0.95	0.96	0.94	0.93	0.95	0.93	0.91
16. Teams	0.92	0.94	0.93	0.78	0.88	0.86	0.86	0.87	0.87	0.89	0.88	0.82	0.88	0.78	0.92	1.00	0.93	0.90	0.95	0.96	0.95	0.93	0.96	0.94	0.91
17. Alliances	0.93	0.94	0.96	0.90	0.93	0.85	0.81	0.86	0.86	0.94	0.91	0.91	0.83	0.82	0.95	0.93	1.00	0.91	0.95	0.96	0.93	0.93	0.95	0.93	0.91
18. IT	0.91	0.93	0.94	0.88	0.83	0.84	0.87	0.68	0.84	0.90	0.86	0.87	0.80	0.77	0.90	0.90	0.91	1.00	0.93	0.94	0.89	0.88	0.93	0.83	0.89
19. Funds	0.93	0.95	0.95	0.93	0.94	0.94	0.94	0.92	0.94	0.95	0.93	0.95	0.91	0.90	0.95	0.95	0.95	0.93	1.00	0.94	0.90	0.83	0.90	0.91	0.87
20. Tax	0.95	0.96	0.96	0.94	0.96	0.95	0.94	0.93	0.94	0.96	0.94	0.95	0.93	0.91	0.96	0.96	0.96	0.94	0.94	1.00	0.91	0.91	0.93	0.92	0.92
21. Banks	0.93	0.95	0.96	0.92	0.94	0.94	0.90	0.88	0.89	0.94	0.92	0.95	0.90	0.86	0.94	0.95	0.93	0.89	0.90	0.91	1.00	0.87	0.90	0.90	0.85
22. Trade	0.88	0.91	0.92	0.91	0.92	0.91	0.90	0.85	0.90	0.92	0.89	0.91	0.89	0.85	0.93	0.93	0.93	0.88	0.83	0.91	0.87	1.00	0.50	0.81	0.82
23. Spread	0.92	0.93	0.94	0.93	0.94	0.94	0.93	0.88	0.93	0.94	0.92	0.93	0.92	0.88	0.95	0.96	0.95	0.93	0.90	0.93	0.90	0.50	1.00	0.83	0.87
24. Options	0.91	0.92	0.91	0.92	0.92	0.92	0.91	0.88	0.90	0.93	0.89	0.91	0.88	0.87	0.93	0.94	0.93	0.83	0.91	0.92	0.90	0.81	0.83	1.00	0.80
25. Risk	0.89	0.92	0.92	0.89	0.89	0.90	0.90	0.83	0.88	0.92	0.82	0.87	0.86	0.81	0.91	0.91	0.91	0.89	0.87	0.92	0.85	0.82	0.87	0.80	1.00

REFERENCES

Aldenderfer, Mark S., and Roger K. Blashfield. 1984. *Cluster Analysis*, edited by M. S. Lewis-Beck. Newbury Park, CA: Sage Publications.

Anderberg, Michael R. 1973. *Cluster Analysis for Applications*, edited by Z. W. Birnbaum and E. Lukacs. New York: Academic Press.

Babbie, Earl. 1998. *The Practice of Social Research*. Belmont, CA: Wadsworth.

Bartell, B. T., G. W. Cottrell, and R. K. Belew. 1995. "Representing Documents Using an Explicity Model of Their Similarities." *Journal of the American Society for Information Science* 46:251–71.

Belew, Richard K. 2000. *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW*. Cambridge, England: Cambridge University Press.

Berry, Michael W., and Murray Browne. 1999. *Understanding Search Engines: Mathematical Modeling and Text Retrieval*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Berry, M. W., Z. Drmac, and E. R. Jessup. 1999. "Matrices, Vector Spaces, and Information Retrieval." *SIAM Review* 41:335–62.

Berry, M. W., S. T. Dumais, and G. W. O'Brien. 1995. "Using Linear Algebra for Intelligent Information Retrieval." *SIAM Review* 37:573–95.

Cardinal, L. B. 2001. "Technological Innovation in the Pharmaceutical Industry: The Use of Organizational Control in Managing Research and Development." *Organization Science* 12:19–36.

Chandy, R. K., and G. J. Tellis. 1998. "Organizing for Radical Product Innovation: The Overlooked Role of Willingness to Cannibalize." *Journal of Marketing Research* 35:474–87.

———. 2000. "The Incumbent's Curse? Incumbency, Size, and Radical Product Innovation." *Journal of Marketing* 64:1–17.

Cheng, Y. T., and A. H. VandeVen. 1996. "Learning the Innovation Journey: Order Out of Chaos?" *Organization Science* 7:593–614.

Cohen, J. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20:37–46.

———. 1968. "Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit." *Psychological Bulletin* 70:213–20.

Deerwester, S., S. Dumais, G. Furnas, T. Landauer, and R. Harshman. 1990. "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science* 41:391–407.

Drazin, R., and C. B. Schoonhoven. 1996. "Community, Population, and Organization Effects on Innovation: A Multilevel Perspective." *Academy of Management Journal* 39:1065–83.

Eckart, C., and G. Young. 1936. "The Approximation of One Matrix by Another of Lower Rank." *Psychometrika* 1:211–18.

Estes, W. K. 1994. *Classification and Cognition*, Vol. 22. New York: Oxford University Press.

Gordon, A. D. 1999. *Classification*. London, England: Chapman and Hall/CRC.

Hampton, J. A. 1995. "Testing the Prototype Theory of Concepts." *Journal of Memory and Language* 34:686–708.

Hull, D. A. 1996. "Stemming Algorithms — A Case Study for Detailed Evaluation." *Journal of the American Society for Information Science* 47:70–84.

Hull, D. A., and G. Grefenstette. 1966. *A Detailed Analysis of English Stemming Algorithms*. Rank Xerox Research Centre Europe, Meylan, France.

Husbands, Parry, Horst Simon, and Chris Ding. 2000. "On the Use of Singular Value Decomposition for Text Retrieval." In *SIAM Comp. Info. Retrieval Workshop*. Raleigh, NC: Society for Industrial and Applied Mathematics.

Klein, K. J., and J. S. Sorra. 1996. "The Challenge of Innovation Implementation." *Academy of Management Review* 21:1055–80.

Kontostathis, April, and William M. Pottenger. 2002. "Detecting Patterns in the LSI Term-Term Matrix." Presented at the Workshop on Foundations of Data Mining and Discovery at the IEEE International Conference on Data Mining, Maebashi City, Japan.

Krovetz, Robert. 1993. "Viewing Morphology as an Inference Process." Presented at the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, PA.

Kruschhke, J. K. 1992. "ALCOVE: An Exemplar Based Connectionist Model of Category Learning." *Psychological Review* 99:22–44.

Kruskal, J. B. 1964. "Multidimensional Scaling by Optimizing Goodness-of-Fit to a Nonmetric Hypothesis." *Psychometrika* 29:28–42.

Kuester, S., C. Homburg, and T. S. Robertson. 1999. "Retaliatory Behavior to New Product Entry." *Journal of Marketing* 63:90–106.

Lakoff, George. 1990. *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. Chicago, IL: University of Chicago Press.

Libutti, L. 2000. "Building Competitive Skills in Small and Medium-Sized Enterprises Through Innovation Management Techniques: Overview of an Italian Experience." *Journal of Information Science* 26:413–19.

Lovins, J. B. 1968. "Development of a Stemming Algorithm." *Mechanical Translation and Computational Linguistics* 11:22–31.

Luhn, H. P. 1957. "A Statistical Approach to Mechanized Encoding and Searching of Literary Information." *IBM Journal of Research and Development* 1:309–17.

Mackensen, Karsten, and Uta Wille. 1999. "Qualitative Text Analysis Supported By Conceptual Data Systems." *Quality and Quantity* 33:135–56.

Mandelbrot, Benoit. 1983. *The Fractal Geometry of Nature*. New York: Freeman.

McLaughlin, Mary E., Peter Carnevale, and Rodney G. Lim. 1991. "Professional Mediators' Judgments of Mediation Tactics: Multidimensional Scaling and Cluster Analyses." *Journal of Applied Psychology* 76:465–72.

Miles, Matthew B., and A. Michael Huberman. 1994. *Qualitative Data Analysis: An Expanded Sourcebook*. Beverly Hills, CA: Sage Publications.

Mone, M. A., W. McKinley, and V. L. Barker. 1998. "Organizational Decline and Innovation: A Contingency Framework." *Academy of Management Review* 23:115–32.

Monge, P. R., M. D. Cozzens, and N. S. Contractor. 1992. "Communication and Motivational Predictors of the Dynamics of Organizational Innovation." *Organization Science* 3:250–74.

Nunnally, Jum C., and Ira H. Bernstein. 1994. *Psychometric Theory*. New York: McGraw-Hill.

Porter, M. F. 1980. "An Algorithm for Suffix Stripping." *Program* 14:130–37.

Repenning, N. P. 2002. "A Simulation-Based Approach to Understanding the Dynamics of Innovation Implementation." *Organization Science* 13:109–27.

Richards, Tom, and Lyn Richards. 1995. "Using Hierarchical Categories in Qualitative Data Analysis." Pp. 80–95 in *Computer-Aided Qualitative Data Analysis: Theory, Methods and Practice*, edited by U. Kelle. London, England: Sage Publications.

Robertson, S., and S. Walker. 1994. "Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval." Pp. 232–41 in *Proceedings of the 17th Annual International ACM SIGIR Conference on R&D in Information Retrieval*, edited by W. Croft and C. van Rijsbergen: New York: Springer-Verlag.

Robertson, S. E. 1997. "Overview of the OKAPI Projects." *Journal of Documentation* 53:3–7.

Rosch, E. 1978. *Cognition and Categorization*. New York: Lawrence Erlbaum.

Salton, G., A. Wong, and C. Yang. 1975. "A Vector Space Model for Automatic Indexing." *Communications of the ACM* 18:613–20.

Salton, Gerard, and Christopher Buckley. 1998. "Term-Weighting Approaches in Automatic Text Retrieval." *Information Processing and Management* 24:513–23.

Sloman, Steven A., Barbara C. Malt, and Arthur Fridman. 2001. "Categorization Versus Similarity: The Case of Container Names." Pp. 73–86 in *Similarity and Categorization*, edited by U. Hahn and M. Ramscar. Oxford, England: Oxford University Press.

Sorensen, J. B., and T. E. Stuart. 2000. "Aging, Obsolescence, and Organizational Innovation." *Administrative Science Quarterly* 45:81–112.

Subramani, Mani, and Eric A. Walden. 2001. "The Impact of E-Commerce Announcements on the Market Value of Firms." *Information Systems Research* 12:135–54.

Swanson, E. Burton. 1994. "Information Systems Innovation Among Organizations." *Management Science* 40:1069–92.

Whittington, R., A. Pettigrew, S. Peck, E. Fenton, and M. Conyon. 1999. "Change and Complementarities in the New Competitive Landscape: A European Panel Study, 1992–1996." *Organization Science* 10:583–600.

Yin, R. K. 2003. *Case Study Research, Design and Methods*. Beverly Hills, CA: Sage Publications.

Zipf, G.K. 1949. *Human Behavior and the Principle of Least Effort*. Reading, MA: Addison-Wesley.