ABSTRACT

Title of Document:	STUDYING THE RELATIONSHIPS OF INFORMATION TECHNOLOGY CONCEPTS
	Chia-jung Tsui, Doctor of Philosophy, 2011
Directed By:	Assistant Professor Ping Wang College of Information Studies

Different information technology concepts are related in complex ways. How can the relationships among multiple IT concepts be described and analyzed in a scalable way? It is a challenging research question, not only because of the complex relationships among IT concepts, but also due to lack of reliable methods. Seeking to meet the challenge, this dissertation offers a computational approach for analyzing, visualizing, and understanding the relationships among IT concepts.

The dissertation contains five empirical studies. The first study employs Kullback-Leibler (KL) divergence to compare the semantic similarity of forty-seven IT concepts discussed in a trade magazine over a ten-year period. Results show that the similarity of IT concepts can be mapped in a hierarchy and similar technologies demonstrated similar discourses. The second study employs co-occurrence analysis to explore the relationships among fifty IT concepts in six magazines over ten years. Results show general patterns similar to those found in the first study, but with interesting nuances. Together, findings from the first two studies imply reasonable validity of this

computational approach. The third study validates and evaluates the approach, making use of an existing thesaurus as ground truth. Results show that the co-occurrence-based IT classification outperforms the KL divergence-based IT classification in agreeing with the ground truth. The fourth study is a survey of information professionals who help evaluate this computational approach. Results are generally consistent with the findings in the previous study. The fifth study explores the co-occurrence analysis further and has generated IT classifications very much similar to the ground truth.

The computational approach developed in this dissertation is expected to help IT practitioners and researchers make sense of the numerous concepts in the IT field. Overall, the dissertation establishes a good foundation for studying the relationships of IT concepts in a representative, dynamic, and scalable way.

STUDYING THE RELATIONSHIPS OF INFORMATION TECHNOLOGY CONCEPTS

By

Chia-jung Tsui

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Advisory Committee:
Assistant Professor Ping Wang, Chair
Associate Professor Kenneth R. Fleischmann
Assistant Professor Yan Qu
Assistant Professor Yue Maggie Zhou
Professor Amy Weinberg, Dean's Representative

© Copyright by
Chia-jung Tsui
2011

Acknowledgements

I am indebted to my advisor and chair of my dissertation committee, Dr. Ping Wang, who advised, supported, and encouraged me throughout this work. I also wish to thank members of the dissertation committee, Dr. Kenneth R. Fleischmann, Dr. Yan Qu, Dr. Yue Maggie Zhou, and Dr. Amy Weinberg for their helpful suggestions and feedback. In particular, I thank Dr. Amy Weinberg for her extensive comments on the draft of the dissertation.

This work would not have been possible without the opportunity given by Dr. Ping Wang to work on his research project: the *Scalable Computational Analysis of the Diffusion of Information Technology Concepts* (PopIT) project funded by NSF grant IIS-0729459. The project helped significantly in shaping the initial ideas of the work. I thank Dr. Ping Wang, Dr. Kenneth R. Fleischmann, Dr. Doug Oard, and Dr. Amy Weinberg for their guidance and ideas in the project. My thanks also go to Asad Sayeed, Lidan Wang, and Hieu Nguyen for their KL divergence scripts, which were applied throughout this work as well as to Tiffany Chao for her great help in collecting, cleaning, and tagging the data on which this work is based.

I am thankful to my fellow doctoral students for the pilot study of the survey. Their suggestions helped improve the final version of the survey. I also thank the survey participants to share their insights and thoughts with me.

Over my years at the University of Maryland, there are many others who have helped me advance intellectually. In particular, I am grateful to Dr. Dagobert Soergel, Dr. Delia Neuman, Dr. Allison Druin, and Dr. Gregory R. Hancock for their instruction and support.

I thank the Graduate School for the prestigious Ann G. Wylie Dissertation Fellowship, which has funded me through the last and the most critical stage of the work.

My greatest appreciation goes to my parents in Taiwan. They have been very supportive for my life and graduate study in the United States. Without their spiritual support, my graduate study in this faraway country would not have been possible.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction and Literature Review

Because of their complication and varieties, different IT concepts are related in complex ways (Wang, 2009). How can the relationships among multiple IT concepts be described and analyzed? It is a difficult question, not only because of the complex relationships among IT concepts, but also due to the lack of reliable methods to describe and analyze the relationships in a scalable way. In the current status of the IT concept literature, most studies employ single-concept research designs, leaving the relationships among IT concepts underexplored (Fichman, 2004). On the other hand, the few multi-concept studies have had to rely on domain experts to evaluate IT concept relationships (Ein-Dor & Segev, 1993; Wang, 2009). Such expert evaluations are difficult to replicate, to generalize to other IT concepts, or to scale up to examine the relationships among a large number of IT concepts. In this dissertation, a computational approach is offered to examine the relationships among multiple IT concepts.

1.1 Limitations in Current IT Concept Studies

Methodologically, most IT concept studies were designed to examine only one or a few concepts, owing to the difficulty in analyzing large-scale data on multiple concepts (Strang & Soule, 1998). On the other hand, in a few multi-concept study, Ein-Dor and Segev (1993) surveyed 17 IT concepts in the Information Systems literature. They identified the concepts' definition according to 31 attributes and 27 functions, and then described the concepts by two bit-vectors: a vector of attributes and a vector of functions. Further they performed multidimensional scaling (MDS) to visualize the relationships

among the concepts in terms of their relative similarity/dissimilarity. Based on the MDS plot, the 17 concepts were classified and their relationships explored. However, their study faces two challenges. First, the choice of attributes or functions is usually a common problem. According to Mayr (1942), there are three potential problems in the choice of attributes: (a) including irrelevant attributes, (b) omitting important attributes, and (c) redundancy or collinearity of chosen attributes. Besides, when the number of IT concepts increases or when more diversified concepts are included so that more attributes or functions are required to differentiate concepts, the effort drawn from human experts increases substantially.

In view of the classification task of Ein-Dor and Segev's study, I first review the definition of a proximity matrix, an input for visualization such as MDS and hierarchical clustering. After that, a computational approach using either KL-divergence or co-occurrence analysis as a proximity measure is proposed to study multiple IT concepts.

1.2 Proximity Matrix as an Input to Visualization

A proximity matrix is a matrix of similarity or dissimilarity measures between a pair of entities. Generally, a proximity matrix can be generated in two ways (Carroll, Arabie, Chaturvedi, & Hubert, 2004). One is by derived measures which are from raw data coded in a matrix of entities by attributes (variables), then converted into a proximity matrix of entities by entities. Ein-Dor and Segev's (1993) study falls into this category and its limitations are already mentioned above. The other is by "direct measures" of similarity or dissimilarity which can be judgments of pairwise similarities for pairs of entities. For example, Sireci and Geisinger (1992) had three content domain experts rate the similarity of 30 items. The task for each expert was to judge how similar the 30 items

are to each other in a 5-point Likert-type scale ranging from 1, “not at all similar”, to 5, “extremely similar.” The experts were not given any criteria on which to rate the similarity of the items. They rated the similarity of every pair and entered their ratings into a matrix. Since reciprocal comparisons were not necessary, each expert provided a 30 x 30 lower-triangular matrix. Similarly, this way to construct a proximity matrix suffers from the scalability problem regarding the effort drawn from experts. A 30-item already requires 465 (C_2^{30}) pairwise comparisons from each expert. The number of item-item comparisons required will increase exponentially as the number of items increases. Besides, the quality of experts employed is important. For a proximity matrix to be representative, it is crucial that experts have knowledge of a certain content domain.

Contrary to the above limitations in constructing a regular proximity matrix, KL divergence and co-occurrence analysis can be calculated computationally for each pair of entities so that a proximity matrix of “direct measures” of similarity or dissimilarity can be formed. As described in more detail in the following chapters, KL divergence is a semantic similarity measure. Like the name “divergence” suggests, it can be a *dissimilarity* measure for a pair of entities. I did not find any literature using KL divergence to construct a proximity matrix for visualization. However, mutual information, a derivative of KL divergence, was used as a similarity measure for clustering (Kojadinovic, 2004; Kraskov, Stogbauer, Andrzejak, & Grassberger, 2005).

On the other hand, a co-occurrence matrix can be considered as a proximity matrix of *similarity* measures (Burgess, Livesay, & Lund, 1998; Burgess & Lund, 2000; K. Lund, Burgess, & Atchley, 1995). It contains categorical information in that not only concrete nominal concepts but also grammatical concepts and abstract concepts can be

categorized (Burgess, et al., 1998; Kevin Lund & Burgess, 1996). Co-word analysis is of the similar concept used in social science. For example, Ding and her colleagues (2001) visualized the intellectual structure of the field of Information Retrieval during the period of 1987–1997. Co-word analysis was employed in their study to reveal patterns and trends in the IR field. Interestingly, they did not use a co-occurrence matrix directly as a proximity matrix. Instead, a proximity matrix was created based on that two words are rated similar if they have similar co-occurrence profile with all the other words within the co-occurrence matrix.

1.3 Visualization of a Proximity Matrix

With the help of either KL divergence or co-occurrence analysis, a proximity matrix can be calculated computationally. After that, the matrix can be visualized for classification. Multidimensional scaling and hierarchical clustering are techniques used to visualize proximities in a low dimensional space and a tree representation respectively. As the two techniques will be described in more detail in the following chapters, the benefits of the techniques are described below.

In his introduction to the multidimensional scaling, Shepard (1972) argued that the value of spatial scaling techniques like MDS lies in their capability "(a) to uncover the hidden pattern or structure of perceptions, and (b) to represent this structure in a form that is much more accessible to the human eye - as a picture or map". On the other hand, clustering is considered a way of learning (Manning & Schütze, 1999). In clustering, similar objects are grouped together in a cluster. Therefore, we can generalize from what we know about some members of the cluster to other members we are not sure about.

1.4 Classification and Relationships of IT Concepts

According to Sokal (1974), classification is defined as “the ordering or arrangement of objects into groups or sets on the basis of their relationships” (p. 1116). The relationships can be based on “observable or inferred properties” (Sokal, 1974, p. 1116). In addition, Sokal (1974) stated about the purpose of a classification:

> The paramount purpose of a classification is to describe the structure and relationship of the constituent objects to each other and to similar objects, and to simplify these relationships in such a way that general statements can be made about classes of objects. (p. 1116)

As for classification techniques, they include cluster analysis and ordination (Sokal, 1974, p. 1123). As a result, the visualization techniques are classification techniques as well; hierarchical clustering belongs to cluster analysis and multidimensional scaling is an ordination technique. In the study, the techniques are applied to a proximity matrix constructed by either KL divergence or co-occurrence analysis to study the relationships of IT concepts. As a result, relationships here are not meant to be those between a pair of IT concepts, such as competition, complement, and substitution, etc. Instead, relationships in the study are about general statements made about classes of objects in classification. For example, the relationships in a list of IT concepts may include enterprise IT-related class and Web2.0-related class.

1.5 Summary

According to the literature review about IT concept studies, most studies examined only one or a few concepts due to the difficulty in analyzing large-scale data.

Besides, others studying multiple concepts constantly faced the limitations in scalability due to expert effort. As a result, the dissertation aims to make a contribution in applying KL divergence or co-occurrence analysis in a computational approach to study *multiple* IT concepts in terms of their similarity and relationships.

The following chapters contain five empirical studies, unified under one overall research question: How can the relationships among multiple IT concepts be described and analyzed in a representative, dynamic, and scalable way? The first study employs Kullback-Leibler (KL) divergence to compare the semantic similarity of forty-seven IT concepts discussed in a trade magazine over a ten-year period. Using hierarchical clustering, I have found that the similarity of the concepts can be mapped in a hierarchy and similar technologies demonstrated similar discourses. The second study employs co-occurrence analysis to explore relationships among fifty IT concepts discussed in six magazines over ten years. Results from hierarchical clustering and multidimensional scaling show general patterns similar to those found in the first study, but with interesting nuances. Together, findings from the first two studies imply reasonable validity of this scalable computational approach. The third study makes use of an existing thesaurus as ground truth to rigorously validate and evaluate this approach. Results show that the co-occurrence-based classification outperforms the KL divergence-based classification in agreeing with the ground truth classification. The fourth study conducts a survey to compare the three classifications: two automatic classifications and the ground truth. Results from the survey correspond to those in the third study. The fifth study further explores the co-occurrence analysis and makes an improvement to the co-occurrence-based classification.

Chapter 2: Exploring the Relationships among IT concepts: a Scalable Computational Approach Using KL Divergence and Hierarchical Clustering

This chapter is based on computational analysis of discourse to examine the relationships among IT concepts. Specifically, Kullback-Leibler (KL) divergence is employed to compare the semantic similarity of forty-seven IT concepts in a trade magazine InformationWeek over a decade. Using hierarchical clustering, the similarities of the technologies can be depicted in hierarchies, and that similar technologies can be clustered into meaningful groups. The results establish the validity of the approach and demonstrate its scalability and richness.

2.1 Introduction

Practitioners who consider adopting and using new IT concepts and scholars who study the diffusion of IT concepts face a constant challenge: At any one time, we confront numerous seemingly promising IT concepts. Some of them become widely adopted and used, making significant contributions to economic prosperity and social welfare; whereas others fade away, leaving little trace behind. While it has been argued that various IT concepts are related to varying degrees and so are their diffusion trajectories (Wang, 2009), it is difficult to make sense of the relationships among IT concepts.

For example, here is a partial list of contemporary IT concepts: Service-oriented architecture (SOA), web services, open source software (OSS), web 2.0, YouTube,

iPhone, blogs, and utility computing. How are they related? How are their diffusion trajectories related? These are difficult questions. Indeed, IT concepts are related in complex ways. First, a broader concept may be comprised of narrower, more specific concepts. Second, different concepts may represent the same core idea. Third, concepts may compete with each other as alternative solutions to similar problems or for the attention from the same group of people or organizations. Finally, concepts may complement each other to accomplish common tasks. Over time, these relationships may change, making it even harder to interpret.

Researchers of IT concept are not well equipped to describe and analyze the complex and evolving relationships among IT concepts. On the one hand, most studies employ single-concept research designs, leaving the relationships among IT concepts underexplored (Fichman, 2004). On the other hand, the few multi-concept studies have had to explicitly or implicitly rely on domain experts to evaluate IT concept relationships (Ein-Dor & Segev, 1993; Wang, 2009). Such expert evaluations are difficult to replicate, to generalize to other IT concepts, or to scale up to examine the relationships among a large number of IT concepts. Therefore, considering the current status of the IT concept literature, this research question is raised: *How can the relationships among a large number of IT concepts be described and analyzed in a scalable way?*

2.2 Data Collection

This chapter focuses on a particular IT trade magazine, InformationWeek, as the data source. All of the articles were downloaded during a ten-year period (1998-2007) in InformationWeek using the Lexis/Nexis online database. Meanwhile, a list of 47 IT concepts was compiled (Table 2.1), ranging from enterprise software (e.g., CRM) to

personal gadgets (e.g., iPod), from abstract concepts (e.g., artificial intelligence) to concrete products/services (e.g., YouTube), and from highly popular (e.g., e-business) to less well-known concepts (e.g., digital subscriber line – DSL). Admittedly, this list is ad hoc, but it serves the illustration purpose well because the list covers a broad range of IT concepts in the examination period. Then, all paragraphs that contain any of the IT concepts on the list were extracted from the InformationWeek articles. In doing so, possible labels for each concept, plural forms, and acronyms unique to the concept were considered. In total, 71,113 paragraphs were extracted, with about 1,500 paragraphs on average for each concept.

Table 2.1: List of IT Concepts

AI	Artificial Intelligence	**MP3**	MP3 player
ASP	Application Service Provider	**MySpace**	MySpace
ATM	Automated Teller Machine	**OLAP**	Online Analytical Processing
BI	Business Intelligence	**OSS**	Open Source Software
Blog	Blog	**Outsource**	Outsourcing
Bluetooth	Bluetooth	**PDA**	Personal Digital Assistant
CAD	Computer Aided Design	**RFID**	Radio Frequency Identification
CRM	Customer Relationship Management	**SmartCard**	Smart Card
DigiCam	Digital Camera	**SCM**	Supply Chain Management
DLearn	Distance Learning	**SFA**	Sales Force Automation
DSL	Digital Subscriber Line	**SocNet**	Social Networking
DW	Data Warehouse	**SOA**	Service-Oriented Architecture
eBiz	eBusiness	**Telecommute**	Telecommuting
eCom	eCommerce	**TabletPC**	Tablet PC
EDI	Electronic Data Interchange	**UtiComp**	Utility Computing
ERP	Enterprise Resource Planning	**Virtualization**	Virtualization
GPS	Global Positioning System	**VPN**	Virtual Private Network
Grpware	Groupware	**Web2.0**	Web 2.0
IM	Instant Messaging	**WebServ**	Web Services
iPhone	iPhone	**WiFi**	Wi-Fi
iPod	iPod	**Wiki**	Wiki
KM	Knowledge Management	**Wikipedia**	Wikipedia
Linux	Linux	**YouTube**	YouTube
Multimedia	Multimedia		

2.3 Data Analysis

To make sense of the relationships among the IT concepts, this chapter explores the similarity of the concepts. The approach is to infer concept similarity from the semantic similarity of the discourses about the concepts. Specifically, Kullback-Leibler (KL) divergence, a probabilistic measure for differences in the pattern of word choices, is employed as a proxy for comparison of the semantic similarity of any two collections of paragraphs extracted from InformationWeek. Based on KL divergence results, hierarchical clustering analysis is used to aggregate the concepts in a hierarchical structure.

2.3.1 KL Divergence

Originally introduced in 1951 (Kullback & Leibler, 1951) and considered a foundation of information theory (Cover & Thomas, 1991), KL divergence is a statistic that quantifies in bits how close a probability distribution *P* is to another distribution *Q*. For probability distributions of discrete random variables, the KL divergence of *Q* from *P* is defined as: $D_{KL}(P \parallel Q) = \sum_i P(i) \log(P(i) / Q(i))$.

In the dataset, each IT concept is represented by concatenating all of the paragraphs mentioning the concept. The use of language in the paragraphs constitutes a probability distribution of normalized unigram word counts and the KL divergence for each pair of IT concepts was calculated. The calculation generated an asymmetric 47x47 matrix with each column and row representing one of the 47 IT concepts. After symmetrization (by averaging the KL divergence in each direction, i.e., $\frac{D_{KL}(P \parallel Q) + D_{KL}(Q \parallel P)}{2}$), the value in each cell of the matrix defines a distance between a pair of IT concepts. In

order to group the concepts and visualize their relationships, hierarchical clustering analysis was performed on the symmetrized KL divergence matrix.

2.3.2 Hierarchical Clustering

Cluster analysis is the process of grouping objects into unknown clusters such that the within-group variation is minimized and the between-group variation maximized (Everitt, Landau, & Leese, 2001). The agglomerative hierarchical clustering method groups objects on a series of levels, from the finest partition, in which each individual object forms its own cluster, and successively combines smaller clusters into larger ones until all objects are in one cluster. Agglomerative hierarchical clustering employs an aggregation criterion, or "linkage rule," to determine how the distance between two clusters should be calculated based on the distance scores of pairs of objects. The most well known aggregation criteria are single link, complete link, and average link (Hansen & Jaumard, 1997). The distance between two clusters is represented by the minimum, maximum, or average distance between any pair of objects, one object from each cluster. In single link clustering, two clusters with the smallest minimum pairwise distance are merged in each step. In complete link clustering, two clusters with the smallest maximum pairwise distance are merged in each step. And average link clustering is a compromise between the other two methods. The complete link was used in this study because it produces small and tight clusters (Manning & Schütze, 1999; van Rijsbergen, 1979).

2.4 Results

The clustering analysis generated a hierarchy of clusters in a dendrogram (Figure 2.1), where vertical lines show joined clusters and the position of the lines on the

horizontal scale from 1 to 25 indicates the distance at which clusters are merged. By inspecting the dendrogram, five natural clusters are identified, all of which merged between 15 and 20 in the scale. These clusters are indicated by the five intersection points between the dendrogram and the vertical dotted line in Figure 2.1.

Take Cluster 1 in the dendrogram as an example. It includes 26 IT concepts. Most of them are enterprise IT applications. The hierarchical structure of this large cluster is shown in the dendrogram. For example, at the next granular level (around 15 in the horizontal scale), there exist two sub-clusters: one consisting of service-oriented IT concepts such as OSS and web services and the other representing more traditional IT concepts, which may be further differentiated at lower levels. Within the latter sub-cluster, for example, the discourse on e-business is very similar to that on e-commerce. Similar relationships seem to exist in concept pairs such as CRM and ERP, and knowledge management (KM) and groupware.

Then, the number of paragraphs mentioning each IT concept was counted every year. The number of paragraphs about a concept indicates the prevalence or popularity of the concept in the discourse. For example, Figure 2.2 shows that the popularity curves of the pair of very similar concepts (e-business and e-commerce) followed very similar patterns: both concepts enjoyed peak popularity around 2000 and then have lost much momentum since the dot-com crash. In contrast, the popularity curves of other very similar concepts followed very different patterns. Consider Figure 2.3, which shows the evolutionary trajectories of web services and SOA. The negatively correlated curves in the figure seem to suggest that the newer SOA replaced the older web services.

Cluster 2 includes five IT concepts: DSL and virtual private network (VPN) are telecommunication technologies which can be applied to the other three concepts in the cluster. As Figure 2.4 shows, DSL and VPN had very similar popularity trajectories.

Only three concepts form Cluster 3 and four concepts form Cluster 5. These two clusters correspond to the so-called web 2.0 technologies that have become highly popular in recent years. Concepts in Cluster 3 share social network as a common feature. Cluster 5 represents text-based web 2.0 applications with user generated contents. The popularity curves in Figures 2.5 and 2.6 show that the concepts in these two clusters have generally experienced dramatic upswings circa 2004. Despite the similarity, the patterns of term use in the two clusters (as measured by symmetrized KL divergence) do not converge at the next level of aggregation. This interesting finding seems to suggest the substantial diversity of web 2.0 technologies.

Lastly, Cluster 4 has nine concepts all related to mobile or wireless technologies. Some, such as bluetooth and Wi-Fi, are the underlying mobile technologies. Others, such as TabletPC and iPod, are the devices enabled by the wireless/mobile technologies. Figure 2.7 shows that the rising popularity of iPhone coincided with the dwindling popularity of iPod, suggesting, once again, that the new replaces the old.

2.5 Summary

The results show that similar IT concepts can be identified in the analysis. Also, the popularity curves of a pair of similar IT concepts seem to suggest a certain relationship among the pair, which can further be investigated. In addition, the results from the KL-divergence and clustering analysis are consistent with our *a priori* knowledge about the relationship among these IT concepts. Such consistency provides

reasonable confidence in the validity of the study's computational approach to understanding IT concept relationships.

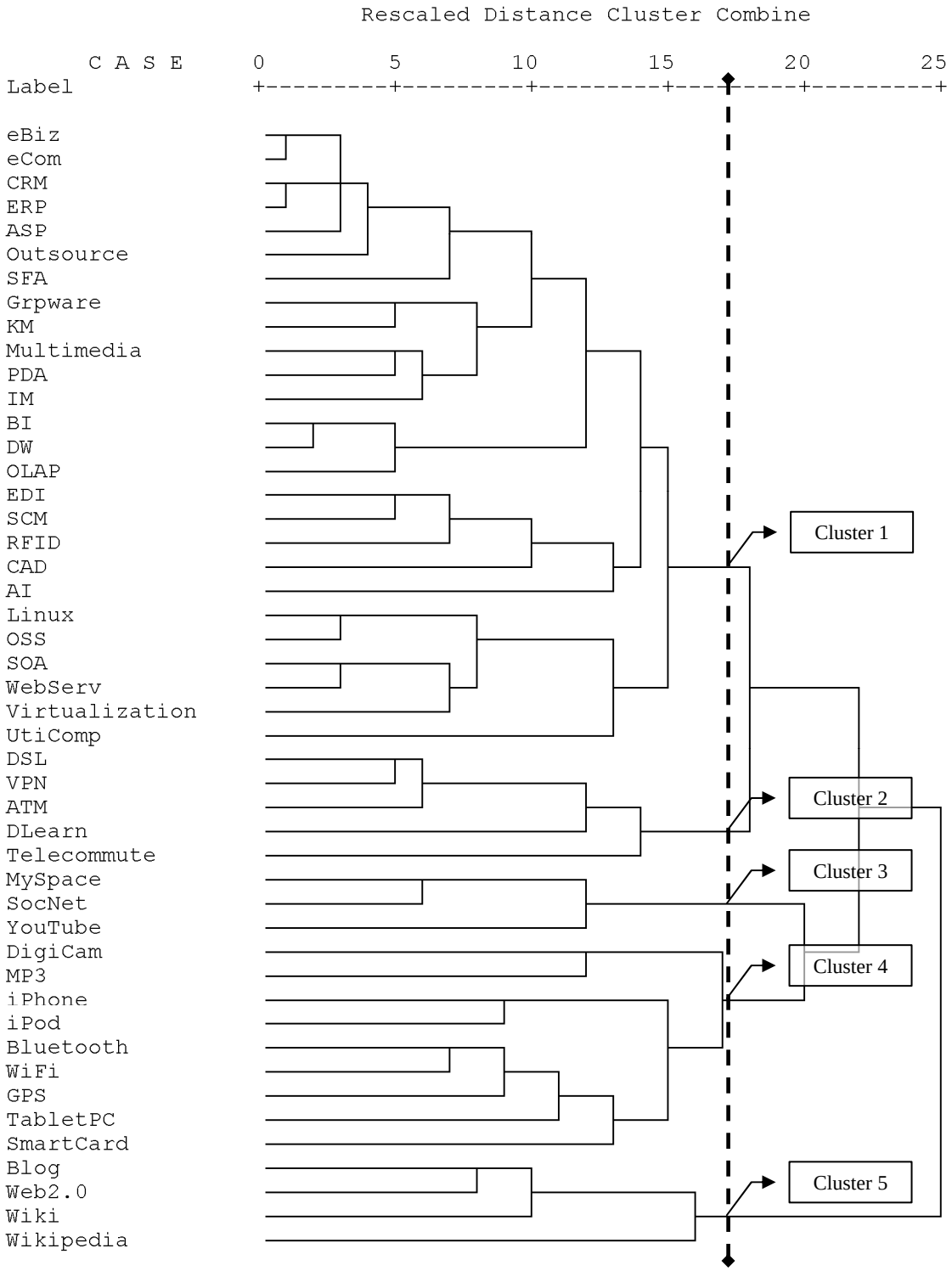

Figure 2.1: Hierarchical Clustering Result on the KL Divergence Matrix

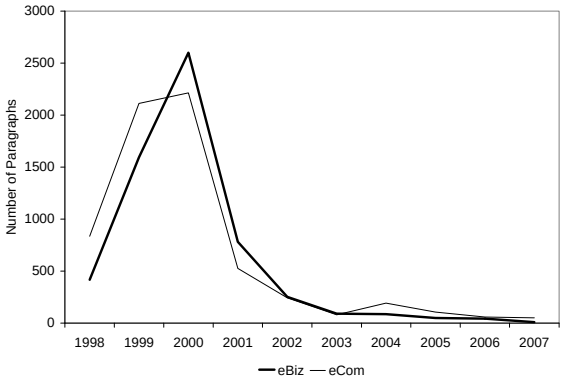

Figure 2.2: Popularity of e-business and e-commerce

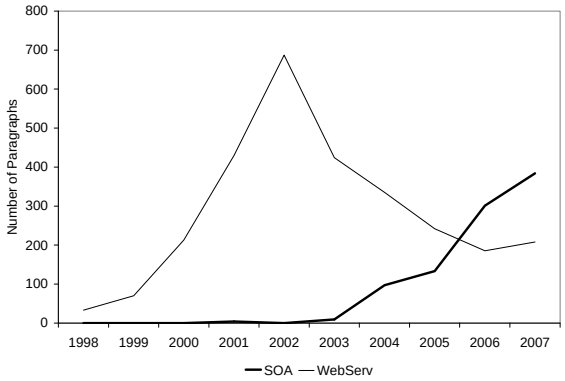

Figure 2.3: Popularity of SOA and Web Services

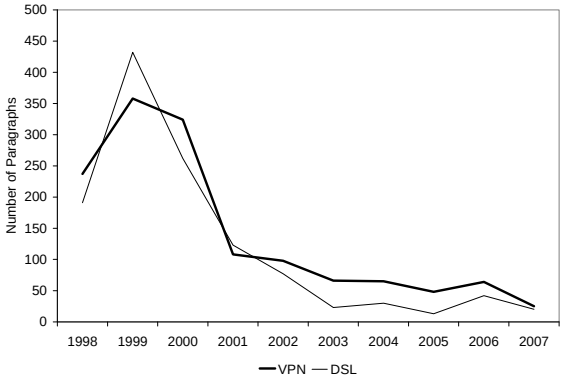

Figure 2.4: Popularity of DSL and VPN

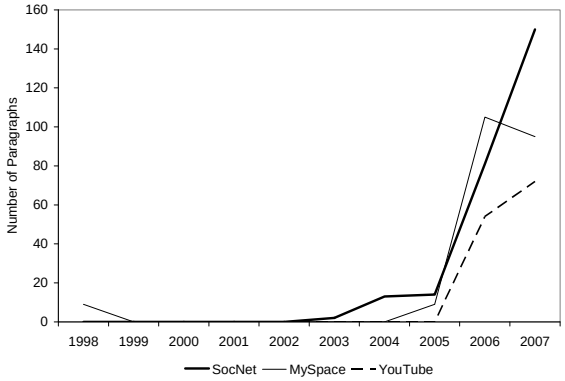

Figure 2.5: Popularity of Social Networking Concepts

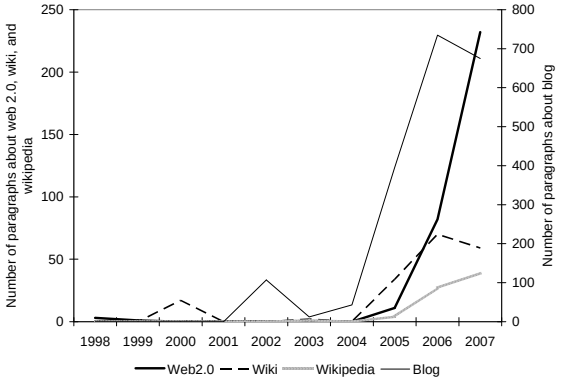

Figure 2.6: Popularity of Web 2.0 Concepts with User Generated Contents

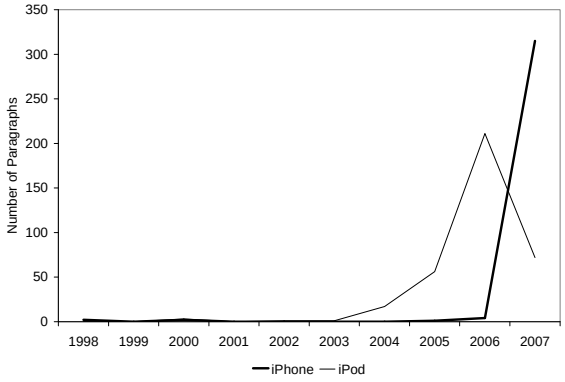

Figure 2.7: Popularity of iPhone and iPod

Chapter 3: Building an IT Taxonomy with Co-occurrence Analysis, Hierarchical Clustering, and Multidimensional Scaling

This chapter makes use of co-occurrence analysis as a substitute for KL divergence in Chapter 2. According to the literature, a co-occurrence matrix carries categorical information. Therefore, it would be interesting to compare the results from the co-occurrence matrix with those from KL divergence. Co-occurrence of terms (or co-word analysis) has also been used in various fields such as computational linguistics and information retrieval to study the relationships among terms. Instead of one trade magazine in Chapter 2, six magazines are used for better representation. The results are interesting in that many groups identified are very similar to those found in Chapter 2. The results also suggest that the method could help build an IT taxonomy.

3.1 Introduction

The proliferation of information technologies (ITs) has been accompanied by the proliferation of information in recent decades. Opportunities emerge from such proliferation of information and technologies, making the iField an increasingly prominent and vibrant area for research and practice. At the same time, just as the explosion of information presents serious challenges in information management, the seemingly everlasting propagation of numerous ITs poses challenges in IT management. The bewildering amount of IT confronting IT practitioners and researchers renders it a challenging task to make sense of the technologies, in order to effectively manage or productively study them. In practice, IT management has been traditionally undertaken

along functional lines such as hardware, software, networking, and services. Streams in IT management research, on the other hand, have mapped well onto traditional categories in practice, drawing insights from various reference disciplines such as computer science, psychology, economics, and sociology. However, recent technological and managerial advances have blurred the boundaries of traditional categories. For example, software and service have converged under the rubric of "software as a service" (SaaS). Moreover, because different types of IT may entail different cost structures, work processes, and potential returns, different ITs may require different management practices and different research methodologies. Hence, contemporary IT management practices (such as IT portfolio management) and the increasing emphasis on interdisciplinary research call for rigorous and up-to-date classifications, or taxonomies, of IT.

3.1.1 Taxonomy for Information Management

Taxonomy is a classification scheme (often hierarchical) of information components (for example, terms, concepts, graphics, sounds) and their interrelationships (Harris, Caldwell, Linden, Knox, & Logan, 2003). Taxonomy creation is usually a "top-down process" by which domain experts provide an overview of the domain, list categories and features of each category, and finally classify categories into broader classes according to how similar the features of the categories are (Logan, 2009). Categories that do not match current classes are put aside until enough categories with sufficiently similar features appear to justify the creation of new classes (Lambe, 2007). It has been recommended that analysts use and customize pre-populated taxonomies whenever available (Jagerman, 2006).

3.1.2 Limitations of Extant Approaches

To varying degrees, extant methods for creating taxonomies in general and specifically for IT rely on experts. While expert opinions are valuable in grounding the taxonomy in specific domains and detecting subtleties in the relationships among categories, current approaches have several limitations.

First, the structures of extant taxonomies represent a relatively narrow set of views from only a few experts. For instance, the choice of features (such as attributes and functions of IT) for classification depends on the specific opinions or background knowledge of the experts who participate in the study. Second, taxonomies built by this approach seem static, fixed at the time when experts created them. Efforts to update existing taxonomies are few and far in between. For example, the ACM Computing Classification System currently being used was created in 1998. As another example, the official Keyword Classification Scheme for Information System Research was last updated in 1993 (Barki, Rivard, & Talbot, 1993). Finally, such scant efforts to update existing taxonomies may be due to another limitation – methods relying on experts are not scalable, lending themselves poorly to automation. As the number of ITs increases, the effort by human experts to describe each technology according to its attributes and functions increases, and the reliability of that classification work may decrease. Addressing these limitations of the extant approaches to IT taxonomy creation, the dissertation tries to develop a methodology that allows wider representations of opinions, dynamic updating at multiple points of times, and large-scale automated analysis of a large number of technologies.

3.2 Data Collection

This chapter focused on two IT trade magazines (InformationWeek and Computerworld), two business magazines (BusinessWeek and The Economist), and two news magazines (Newsweek and US News & World Report). All articles published during a ten-year period (1998-2007) in the six magazines were downloaded from the Lexis/Nexis online database. In total, there are about 220,000 articles in the data (Table 3.1). The scale of the data collected from the six magazines is large enough to demonstrate the scalability of the approach. In addition to the scale, the data is also diverse, representing a wide range of views on IT and broader topics.

Table 3.1: Number of Articles for the Six Magazines from 1998 to 2007

Magazine	Number of Articles
InformationWeek	30,432
Computerworld	28,535
BusinessWeek	52,033
The Economist	45,597
Newsweek	41,152
US News & World Report	21,419
Total	219,168

Meanwhile, a list of 50 IT concepts similar to but not exactly the same as those in the previous chapter were compiled (Table 3.2). Admittedly, this list is *ad hoc*, but it serves the illustration purpose because the list covers a broad range of technologies in the examination period. Then, all paragraphs that contain any of the technologies on the list were extracted from the articles. In doing so, plural forms and acronyms unique to the technology were considered for each technology. For example, in extracting paragraphs containing “digital subscriber line,” paragraphs mentioning “digital subscriber lines” and “DSL” were also included. In total, 105,400 paragraphs containing at least one

technology on the list were extracted from the full text of the articles published in the six magazines.

3.3 Data Analysis

The approach in the chapter is to infer similarity of technologies from their co-occurrences in a paragraph. Both hierarchical clustering analysis and multidimensional scaling are used for classifying.

Table 3.2: IT Concepts Examined in the Chapter

AI	Artificial intelligence	**Multimedia**	Multimedia
ASP	Application service provider	**MP3**	MP3 player
BI	Business intelligence	**MySpace**	MySpace
Blog	Blog	**NeuralNet**	Neural net
Bluetooth	Bluetooth	**OLAP**	Online analytical processing
BizProReen	Business process reengineering	**OSS**	Open source software
CloudCom	Cloud computing	**Outsource**	Outsourcing
CRM	Customer relationship management	**PDA**	Personal digital assistant
DigiCam	Digital camera	**RFID**	Radio frequency identification
DLearn	Distance learning	**SmartCard**	Smart card
DSL	Digital subscriber line	**SCM**	Supply chain management
DecisionSS	Decision support system	**SFA**	Salesforce automation
DW	Data warehouse	**SocNet**	Social networking
eBiz	Electronic business	**SOA**	Service oriented architecture
eCom	Electronic commerce	**Telecommute**	Telecommuting
EDI	Electronic data interchange	**TabletPC**	Tablet PC
ERP	Enterprise resource planning	**UtiComp**	Utility computing
ExpertSys	Expert system	**Virtualization**	Virtualization
GPS	Global positioning system	**VPN**	Virtual private network
Grpware	Groupware	**Web2**	Web 2.0
IM	Instant messaging	**WebServ**	Web services
iPhone	iPhone	**WiFi**	Wi-Fi
iPod	iPod	**Wiki**	Wiki
KM	Knowledge management	**Wikipedia**	Wikipedia
Linux	Linux	**YouTube**	YouTube

3.3.1 Co-occurrence Analysis

Co-occurrence of terms has been used in various fields such as computational linguistics (Burgess & Lund, 1997a, 1997b) and information retrieval (Smadja, 1993) to study the relationships among terms. For example, Spence and Owens (1990) used co-occurrence to evaluate the strength of word association. They found that related pairs of nouns co-occur considerably more often than unrelated pairs. Their finding suggests that co-occurrence frequency indicate the strength of word association.

Analysis of co-occurrence should define a proper size of the window where terms co-occur. A window size can be a certain number of words or characters (Spence & Owens, 1990) or a logical division of an input text (Schvaneveldt, 1990). The paragraph was chosen as the window size in the study because it sufficiently captures the context for describing related technologies.

To measure co-occurrence at the paragraph level, from the 105,400 paragraphs initially extracted, paragraphs containing two or more IT concepts in the list were further selected. This filtering process returned approximately 12,000 paragraphs. Then a co-occurrence matrix of 50 by 50 was created with each row or column representing an IT concept on the list. Each cell in the matrix displays the frequency of one IT concept co-occurring with another concept in paragraphs. The co-occurrence matrix can be considered as a similarity matrix. The matrix was transformed to a dissimilarity matrix with the formula: $1/(x+0.1)$.

3.3.2 Hierarchical Clustering

As described in Chapter 2.

3.3.3 Multidimensional Scaling

Previous research has found that applying multidimensional scaling (MDS) and clustering separately to the same proximity data results in greater insight into the structure underlying the data and can detect more subtle and complex relationships than either method used alone (Kruskal, 1977; Napior, 1972; Shepard & Arabie, 1979). Both clustering and MDS are visualization techniques. The key difference between the two techniques is that MDS provides a spatial representation of the data, while clustering provides a tree representation (Kruskal, 1977).

Based upon a matrix of item-item similarities or dissimilarities, an MDS algorithm assigns a location to each item in a multidimensional space such that the distances between the items correspond as closely as possible to the measured dissimilarities between the items. In other words, the proximity of items to each other in the space indicates how similar they are. The MDS procedure based on the ALSCAL or alternating least squares scaling (Takane, Young, & de Leeuw, 1977), a popular MDS algorithm, was used. For easy interpretation of the result, the MDS solutions in a two-dimensional scatter plot were presented.

3.4 Results

The clustering analysis of the transformed co-occurrence matrix generated a hierarchical structure of 50 technologies in a dendrogram (Figure 3.1), where vertical lines show joined clusters and the position of the lines on the scale from 1 to 25 indicates the distance at which clusters are merged. By inspecting the dendrogram, eight clusters can be identified. These eight clusters are indicated by the intersections between the dendrogram and the vertical dotted line in Figure 3.1. Table 3.3 summarizes the

membership of each cluster. In Figure 3.2, the 50 ITs were depicted in a two-dimensional MDS plot. Following Shepard and Arabie's (1979) suggestion, different colors were used to represent the eight clusters in the plot. Generally, most of the technologies in the same cluster are located close to each other in the MDS plot. Several clusters are described in more details below.

Cluster 1 includes twelve IT concepts. All of them are enterprise IT applications except outsourcing, which is a strategy for managing enterprise IT. Business process reengineering (BPR) was the last to join the cluster, suggesting that it is the least similar to the others in the cluster. This may explain why BPR looks like an outlier in the cluster in the MDS plot (Figure 3.2). Cluster 5 includes four IT concepts. Among them, digital subscriber line and virtual private network are both telecommunication technologies, which may be employed in the other two IT applications (telecommuting and distance learning). Cluster 6 has ten IT concepts, all related to mobile or wireless technologies. Some, such as bluetooth and Wi-Fi, are the underlying mobile technologies. Others, such as TabletPC and PDA, are the devices enabled by the wireless/mobile technologies. Cluster 7 has eight IT concepts. They are the so-called Web 2.0 technologies that have become highly popular in recent years. Lastly, Cluster 8 includes seven IT concepts of similar type such as utility computing, Web service, and cloud computing.

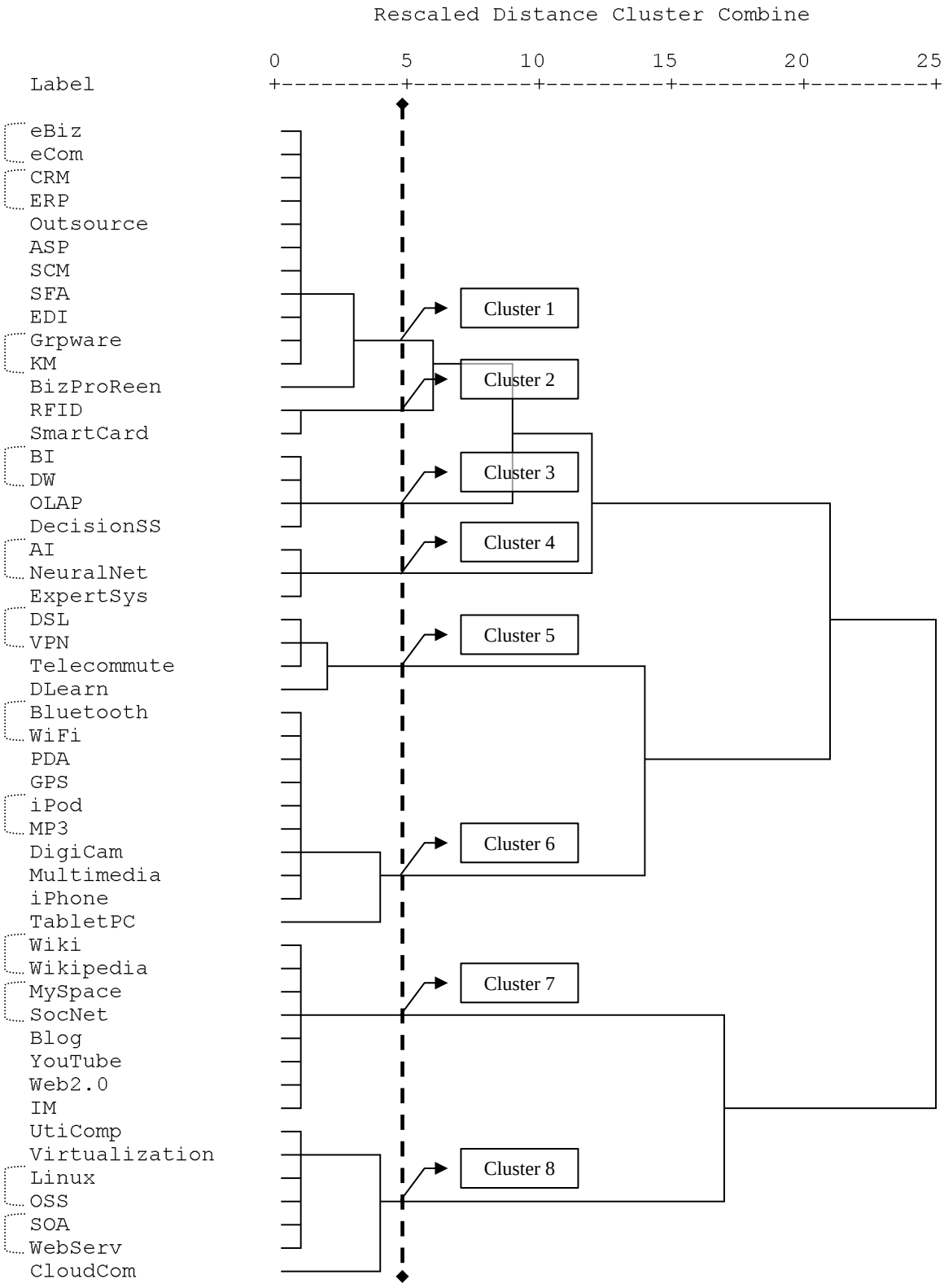

Figure 3.1: Hierarchical Clustering Result on the Co-occurrence Matrix

Table 3.3: Membership of the Clusters

Cluster	Labels of Information Technologies
1	eBiz, eCom, CRM, ERP, Outsource, ASP, SCM, SFA, EDI, Grpware, KM, BizProReen
2	RFID, SmartCard
3	BI, DW, OLAP, DecisionSS
4	AI, NeuralNet, ExpertSys
5	DSL, VPN, Telecommute, DLearn
6	Bluetooth, WiFi, PDA, GPS, iPod, MP3, DigiCam, Multimedia, iPhone, TabletPC
7	Wiki, Wikipedia, MySpace, SocNet, Blog, YouTube, Web2.0, IM
8	UtiComp, Virtualization, Linux, OSS, SOA, WebServ, CloudCom

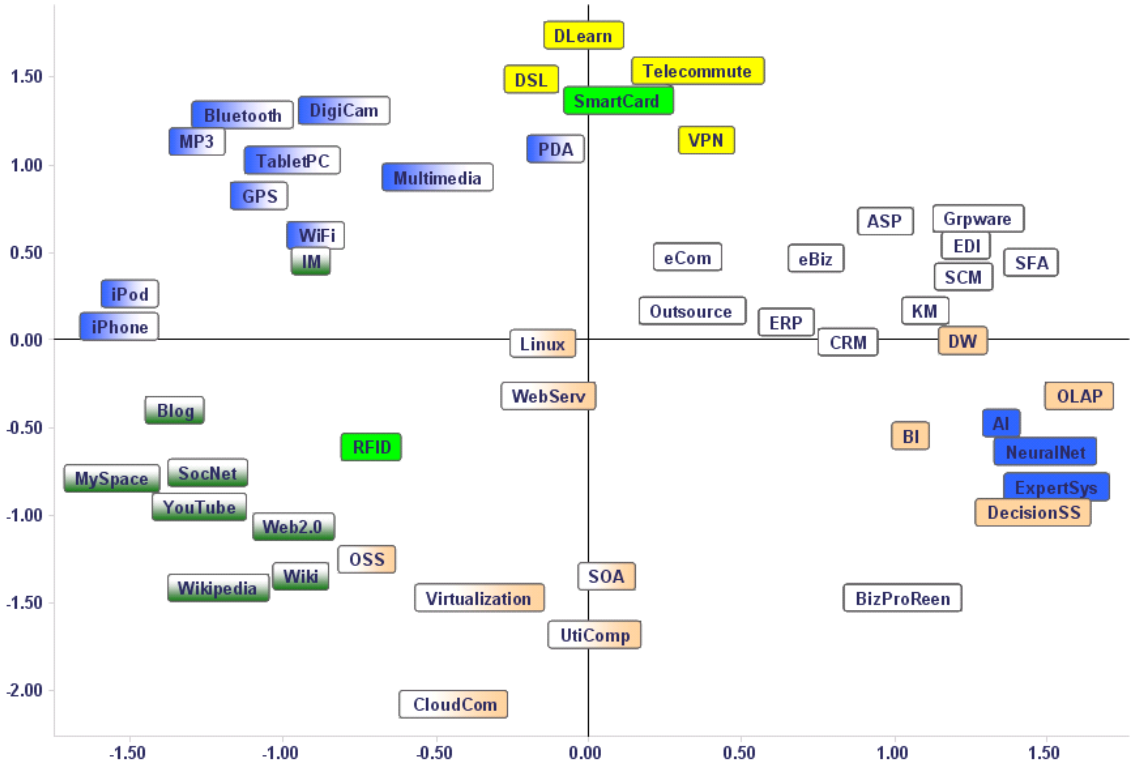

Figure 3.2: Multidimensional Scaling Result on the Co-occurrence Matrix

According to the agglomeration schedule, a series of steps during clustering, twelve pairs of IT concepts can be identified as most similar to each other in the list (see Table 3.4). The pairs include, for example, e-business and e-commerce, iPod and MP3,

and artificial intelligence and neural net. These pairs are compatible with even rudimentary understanding of these technologies.

Table 3.4: Pairs of Most Similar IT Concepts

Pair	IT Concepts	Pair	IT concepts
1	eBiz, eCom	7	Bluetooth, WiFi
2	CRM, ERP	8	iPod, MP3
3	Linux, OSS	9	DSL, VPN
4	BI, DW	10	Grpware, KM
5	SOA, WebServ	11	AI, NeuralNet
6	MySpace, SocNet	12	Wiki, Wikipedia

3.5 Summary

The results illustrate that co-occurrence analysis can be utilized for IT classification. Co-occurrence analysis, supplemented by the two classification techniques, has yielded results that can be interpreted fairly easily, even without the presence of sophisticated expert knowledge of the various domains that the list of IT concepts covers. The face validity in this illustration study provides reasonable confidence in applying the methodology to other circumstances where *a priori* knowledge is unavailable, such as the cases of new or unknown technologies.

Chapter 4: Evaluating the Two Methods of Classifying IT Concepts with Help from an Existing Thesaurus

In this chapter, I compare co-occurrence analysis and KL divergence to a ground truth thesaurus. The F-measure is used as a systematic way to assess the similarity between each of the two automatic classifications and the ground truth thesaurus. The results indicate that co-occurrence analysis outperforms KL divergence in agreeing with the ground truth thesaurus.

4.1 Introduction

The previous work in Chapter 2 and Chapter 3 has demonstrated automatic IT classification without human experts by employing either KL divergence or co-occurrence analysis. However, the results from the work lack the presence of *ground truth* for comparison. In this chapter, I use the ProQuest classification as the ground truth and compare results from the two classification methods with the ground truth. I illustrate the approach with an empirical study of 35 IT concepts in six magazines over ten years.

4.2 Data Collection

My institution subscribes to the Lexis/Nexis Academic database. The database indexes full-text articles of a wide variety of publications in a format that is easy to convert to machine-readable. Therefore we downloaded from Lexis/Nexis Academic all articles published during a ten-year period (1998-2007) in six magazines, including IT trade magazines (InformationWeek and Computerworld), business magazines

(BusinessWeek and The Economist), and news magazines (Newsweek and US News & World Report).

To select the IT concepts to study, we reviewed the indexes of textbooks on Information Systems or IT management published during the same ten-year period (1998-2007). Then we selected IT concepts that appeared in the indexes of more than two textbooks. From this pool, we furthered selected 18 IT concepts that appeared as main entries in the ProQuest Controlled Vocabulary of Subject Terms (ProQuest classification hereafter). The 18 IT concepts were used as seeds for subsequent analysis. They are artificial intelligence, blog, business process reengineering, customer relationship management, data warehouse, electronic commerce, enterprise resource planning, global positioning system, instant messaging, mp3 player, object oriented programming, open source software, personal digital assistant, radio frequency identification, service oriented architecture, supply chain management, utility computing, and virtualization.

The ProQuest classification is chosen as the ground truth because it is a high-quality thesaurus that ProQuest content analysts manually produce (Wang, 2009). Then, each of the 18 seed IT concepts was joined by its "related terms" according to the ProQuest classification to form a group. As a result, 120 IT concepts in 18 groups were identified.

Next, the paragraph was chosen as the unit of analysis in the study since a paragraph sufficiently captures the context for describing related IT concepts. I then counted the number of paragraphs that mention each of the 120 IT concepts identified above. The frequency of each concept ranges from zero (e.g., intelligent vehicle highway system) to over eighty-five thousand (e.g., software). Figure 4.1 shows the number of

paragraphs in log scale for the 120 IT concepts. To avoid data sparseness in later analysis, I focused on IT concepts with frequencies above the median (251.50), leaving 60 IT concepts.

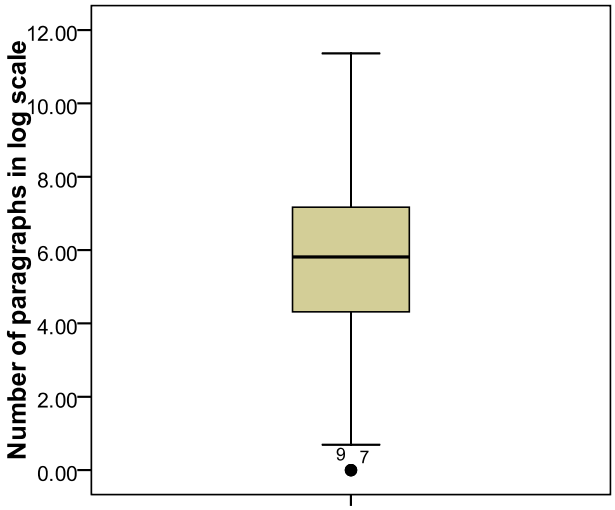

Figure 4.1: Number of Paragraphs in Log Scale for the 120 IT Concepts

As the list is too diversified, I further refined it with the following three steps. First, I deleted five non-IT terms. Second, I deleted sixteen terms which are too general or abstract in meaning. At last, four singletons were removed. Table 4.1 lists the terms I removed during the refinement. As a result, 35 IT concepts are left (Table 4.2). Table 4.3 shows the 35 IT concepts in 14 groups according to the ProQuest classification. For definitions of the 35 IT concepts, please refer to Appendix A.

Table 4.1: Terms Removed during the Refinement

Non-IT terms	benchmark, customer satisfaction, diary, logistics, real time
General/abstract terms	data processing, distribution channel, internet access, information management, internet, information system, information technology, operating system, server, software, software service, supply chain, systems management, systems development, user interface, web site
Singletons	java, personal digital assistant, radio frequency identification, utility computing

Table 4.2: The 35 IT Concepts and their Labels

AI	artificial intelligence	**OLAP**	online analytical processing
ASP	application service provider	**OLAdvertising**	online advertising
ATC	air traffic control	**OLSales**	online sales
Aviation	aviation	**OSS**	open source software
BIS	business intelligence software	**PKI**	public key infrastructure
Blog	blog	**QualityCtrl**	quality control
ChatRoom	chat room	**Robot**	robot
CRM	customer relationship management	**RSS**	rss technology
DataMining	data mining	**SCM**	supply chain management
DigitalMusic	digital music	**SFA**	salesforce automation
DW	data warehouse	**6Sigma**	six sigma
eCom	electronic commerce	**SocNet**	social networking
ERP	enterprise resource planning	**SOA**	service oriented architecture
GPS	global positioning system	**Virtualization**	virtualization
IM	instant messaging	**VPN**	virtual private network
InvenManage	inventory management	**WebServ**	web service
Linux	linux	**WWW**	world wide web
MP3	mp3 player		

Table 4.3: IT Concepts in the ProQuest Classification

No.	IT concepts
1	air traffic control, aviation, global positioning system
2	application service provider, enterprise resource planning
3	artificial intelligence, robot
4	blog, rss technology, social networking
5	business intelligence software, customer relationship management, data mining, salesforce automation
6	chat room, instant messaging
7	data warehouse, online analytical processing
8	digital music, mp3 player
9	electronic commerce, online advertising, online sales, public key infrastructure, world wide web
10	inventory management, supply chain management
11	linux, open source software
12	quality control, six sigma
13	service oriented architecture, web service
14	virtual private network, virtualization

4.3 Data Analysis

I first used co-occurrence analysis and KL divergence in parallel to automatically construct two separate proximity matrices of the 35 IT concepts. The co-occurrence matrix can be considered as a matrix of similarity while the KL divergence matrix is considered as a matrix of dissimilarity. In order to have the same comparison basis for the two methods, I transformed the co-occurrence matrix into a dissimilarity matrix with the formula: $1/(x+0.1)$. Then I applied hierarchical clustering to the two matrices to classify the IT concepts. The complete link standard was used as an aggregation criterion in hierarchical clustering because it produces small and tight clusters (Manning & Schütze, 1999; van Rijsbergen, 1979). Hierarchical clustering can generate specific numbers of clusters. I obtained 14 clusters as there are 14 groups in the ground truth thesaurus. In

addition to hierarchical clustering, I applied multidimensional scaling to the two matrices. Finally, I used the F-measure to compare between each automatic classification with the ProQuest classification. Figure 4.2 shows the flow chart of the data analysis.

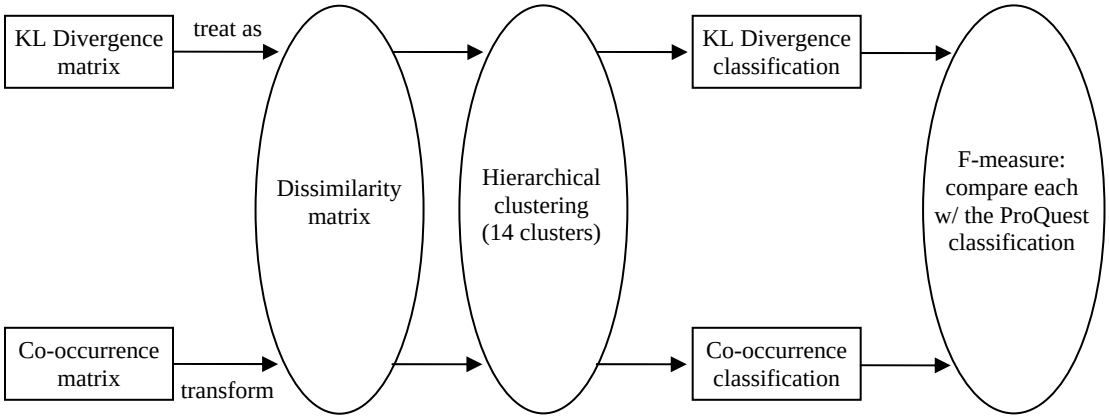

Figure 4.2: Flow Chart of the Data Analysis

Below, I describe the F-measure in detail. For other components of the data analysis such as hierarchical clustering and multidimensional scaling, please refer to the previous chapters.

4.3.1 F-measure

Using the ProQuest classification as the ground truth, Precision, Recall, and F-measure were computed. In the context of classification, the terms true positive (*tp*), true negative (*tn*), false positive (*fp*), and false negative (*fn*) are used to compare the *obtained* classification of an object (the class assigned by a classifier) with the *correct* classification (the class to which an object truly belongs). The four terms are illustrated in Table 4.4. In the study, the classification in the ground truth thesaurus is treated as correct while the clustering result from co-occurrence analysis or KL divergence is treated as

obtained. Besides, the result is *positive* if a pair of IT concepts is in the same group while the result is *negative* if a pair is not in the same group. Depending on correct and obtained results for a pair of concepts, the count for one of the four terms will be added one. For example, if both correct and obtained results are positive, the count for true positive will be added one. As a result, with all pairs of IT concepts and their correct and obtained results, *tp*, *tn*, *fp*, and *fn* are available and Precision and Recall can be computed accordingly ($\text{Precision} = \frac{tp}{tp + fp}$; $\text{Recall} = \frac{tp}{tp + fn}$). F-measure (van Rijsbergen, 1979) is a measure that considers both Precision and Recall. It can be calculated as a harmonic mean of Precision and Recall ($\text{F-measure} = \frac{2 \times precision \times recall}{precision + recall}$). F-measure is a similarity measure between two classifications. It equals one when the two classifications are in fact identical. The F-measure is used in the study as a systematic way to assess the similarity between the ProQuest classification and one of the two automatic classifications.

Table 4.4: Illustration of the Terms tp, tn, fp, and fn

		Correct result or classification	
		Positive	Negative
Obtained result or classification	Positive	true positive (*tp*)	false positive (*fp*)
	Negative	false negative (*fn*)	true negative (*tn*)

4.4 Results

The results of hierarchical clustering and multidimensional scaling on the co-occurrence matrix are in Figure 4.3 and Figure 4.4 respectively. The multidimensional scaling result is shown only for complement as the classification (see Table 4.5) is generated by obtaining 14 clusters in hierarchical clustering. Comparing this

classification with the ProQuest classification (Table 4.3) results in an F-measure of 0.729. Besides, 6 out of 14 groups are identical to those in the ProQuest classification.

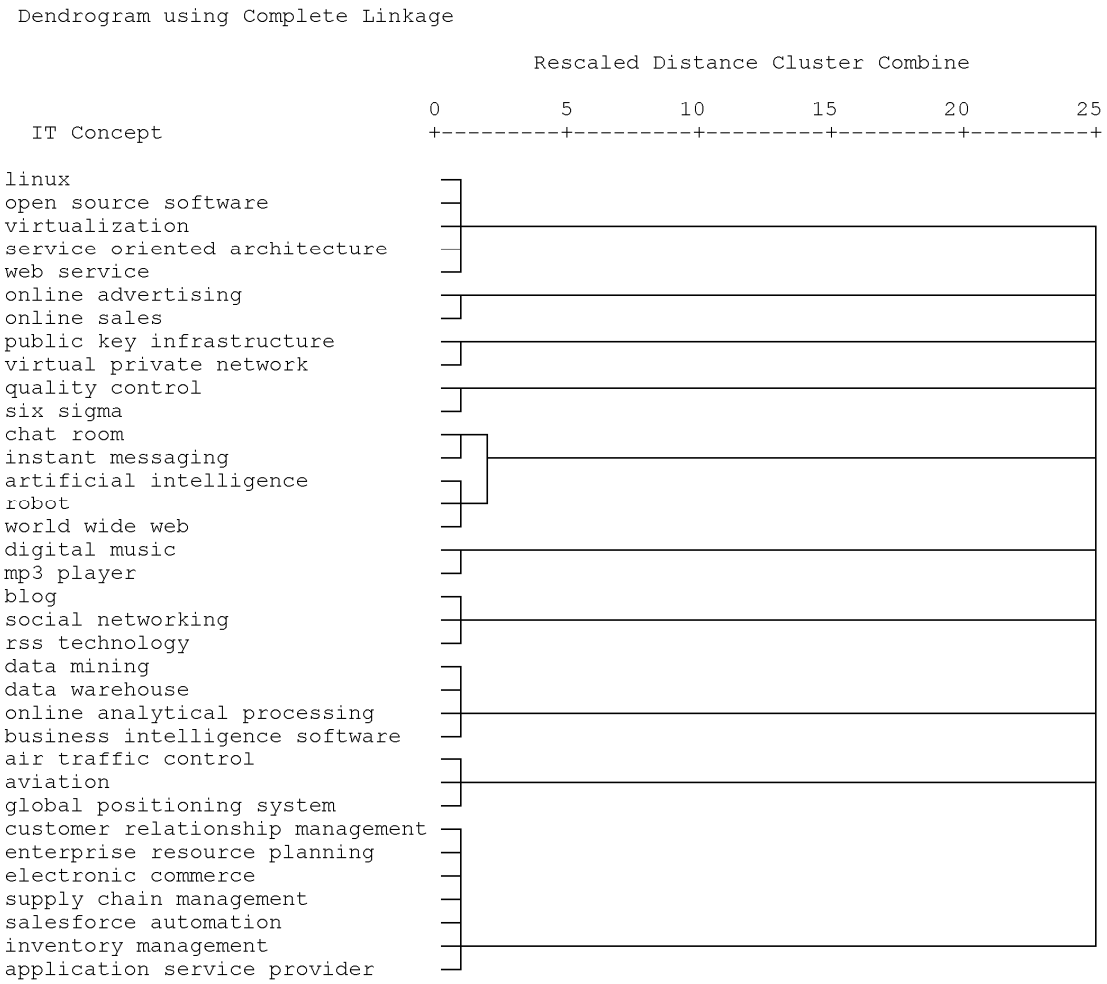

Figure 4.3: Hierarchical Clustering Result on the Co-occurrence Matrix

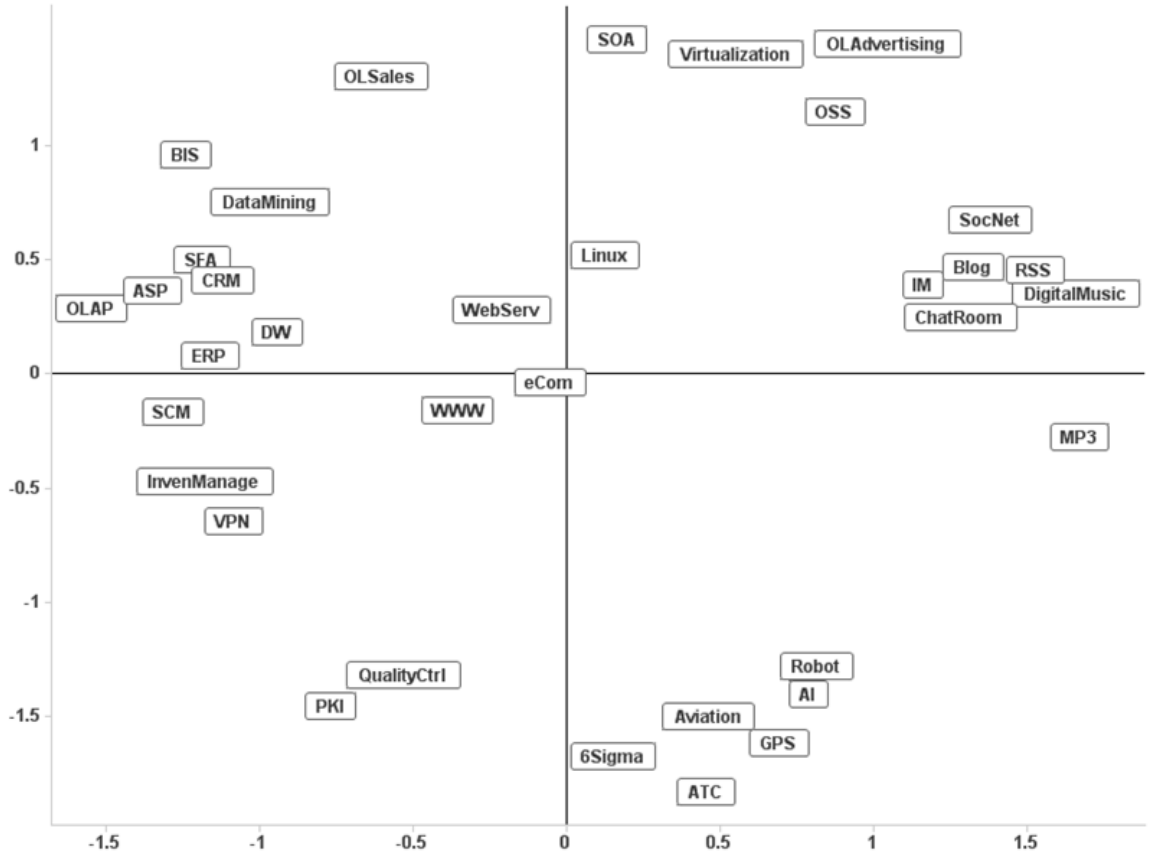

Figure 4.4: Multidimensional Scaling Result on the Co-occurrence Matrix

On the other hand, the results of hierarchical clustering and multidimensional scaling on the KL divergence matrix are in Figure 4.5 and Figure 4.6 respectively. The multidimensional scaling result was shown only for complement as the classification (see Table 4.6) is generated by obtaining 14 clusters in hierarchical clustering. Comparing this classification with the ProQuest classification (Table 4.3) results in an F-measure of 0.615. Besides, 3 out of 14 groups are identical to those in the ProQuest classification.

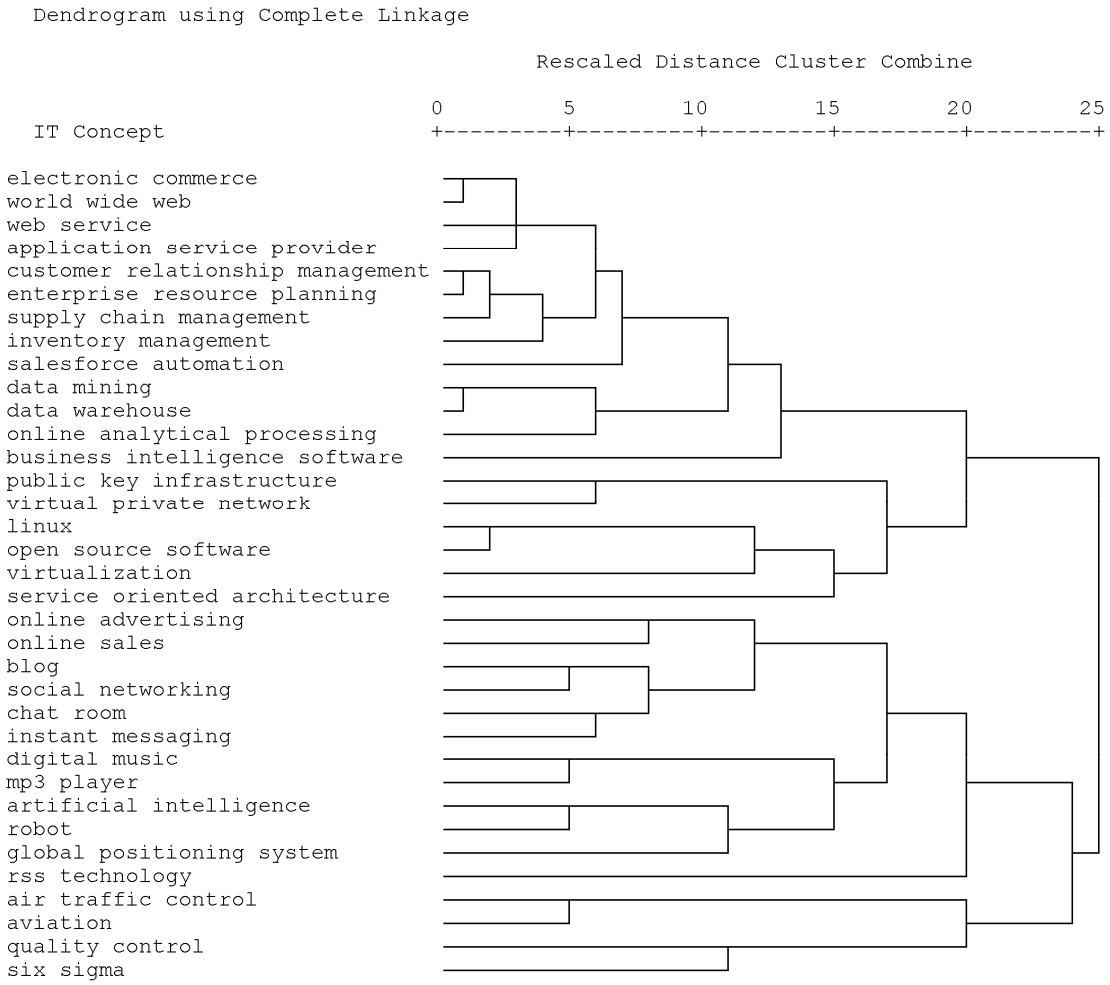

Figure 4.5: Hierarchical Clustering Result on the KL Divergence Matrix

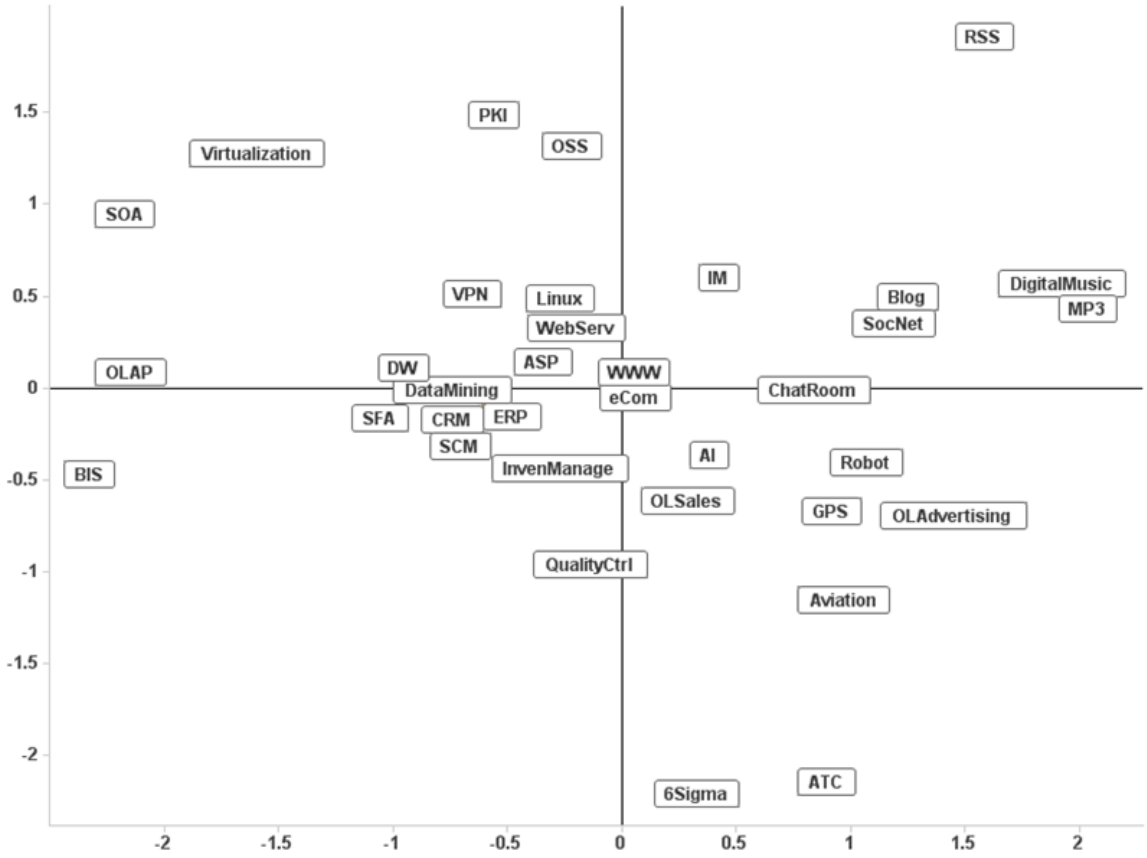

Figure 4.6: Multidimensional Scaling Result on the KL Divergence Matrix

Table 4.3: IT Concepts in the ProQuest Classification (Repeated for Comparison)

No.	IT concepts
1	air traffic control, aviation, global positioning system
2	application service provider, enterprise resource planning
3	artificial intelligence, robot
4	blog, rss technology, social networking
5	business intelligence software, customer relationship management, data mining, salesforce automation
6	chat room, instant messaging
7	data warehouse, online analytical processing
8	digital music, mp3 player
9	electronic commerce, online advertising, online sales, public key infrastructure, world wide web
10	inventory management, supply chain management
11	linux, open source software
12	quality control, six sigma
13	service oriented architecture, web service
14	virtual private network, virtualization

Table 4.5: Automatic Classification by Co-occurrence Analysis

No.	IT concepts
1	air traffic control, aviation, global positioning system
2	application service provider
3	artificial intelligence, robot, world wide web
4	blog, rss technology, social networking
5	business intelligence software, data mining, data warehouse, online analytical processing
6	chat room, instant messaging
7	customer relationship management, electronic commerce, enterprise resource planning, inventory management, salesforce automation, supply chain management
8	digital music, mp3 player
9	linux, open source software, virtualization
10	online advertising
11	online sales
12	public key infrastructure, virtual private network
13	quality control, six sigma
14	service oriented architecture, web service

Note: **6** out of 14 groups are identical (F=.729)

Table 4.6: Automatic Classification by KL Divergence

No.	IT concepts
1	air traffic control, aviation
2	application service provider, customer relationship management, electronic commerce, enterprise resource planning, inventory management, salesforce automation, supply chain management, web service, world wide web
3	artificial intelligence, global positioning system, robot
4	blog, chat room, instant messaging, social networking
5	business intelligence software
6	data mining, data warehouse, online analytical processing
7	digital music, mp3 player
8	linux, open source software
9	online advertising, online sales
10	public key infrastructure, virtual private network
11	quality control, six sigma
12	rss technology
13	service oriented architecture
14	virtualization

Note: **3** out of 14 groups are identical (F=.615)

4.5 Discussion

The comparison of the three classifications leads to interesting findings. First, co-occurrence analysis has a better F-measure (0.729) than KL divergence does (0.615). This difference suggests that the co-occurrence analysis provides a classification more similar to the ProQuest classification than KL divergence does. While there is no rule of thumb for F-measure, an F-measure of 0.729 should be considered high similarity. In addition, there seems to be a mix of strengths and weaknesses in all three classifications, although the ProQuest and co-occurrence-based classifications appear to have sounder groupings than the KL divergence-based classification.

Second, when comparing the 35 IT concepts across the three classifications, I found some IT concepts consistently appear in the same group in the three classifications. Table 4.7 lists 9 pairs of these overlapping concepts. This observation suggests that there is commonality across the three classification methods. Furthermore, these 9 pairs can be easily observed in the two MDS plots and each concept is placed close to the other from the same pair (Figure 4.7 and Figure 4.8). Interestingly, the 9 pairs are in the same quadrant when comparing the two figures, except for only one pair (Linux and OSS).

Table 4.7: IT Concepts Grouped Together in the Three Classifications

1	air traffic control, aviation
2	artificial intelligence, robot
3	blog, social networking
4	chat room, instant messaging
5	data warehouse, online analytical processing
6	digital music, mp3 player
7	inventory management, supply chain management
8	linux, open source software
9	quality control, six sigma

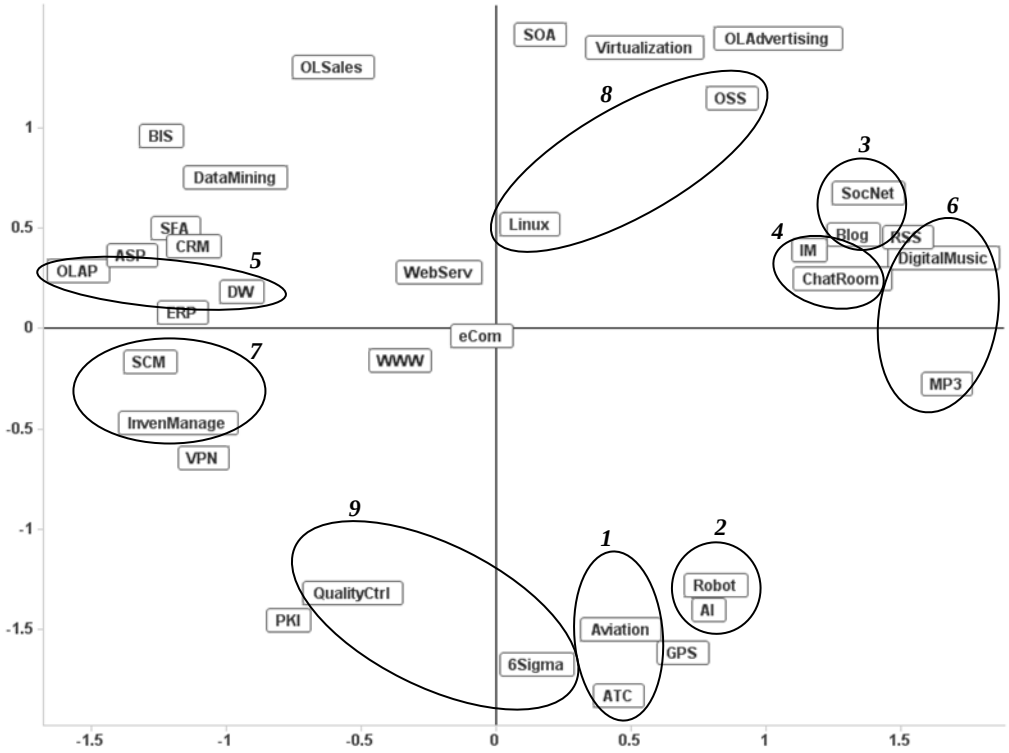

Figure 4.7: The 9 pairs of IT Concepts on the Co-occurrence-based MDS Plot

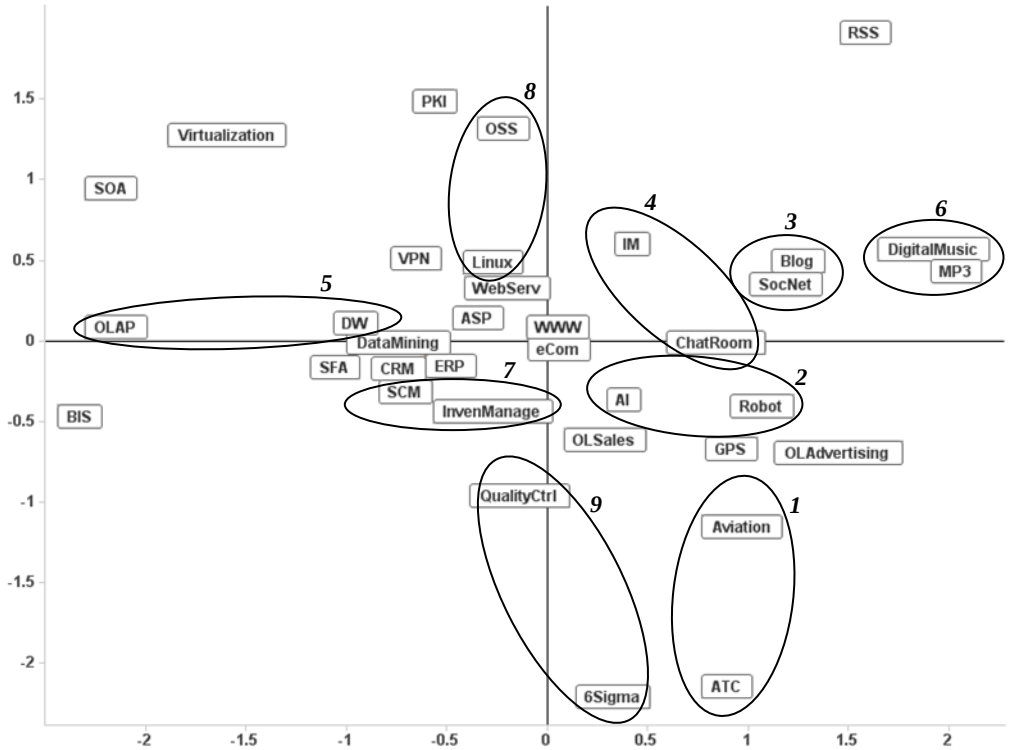

Figure 4.8: The 9 pairs of IT Concepts on the KL Divergence-based MDS Plot

Third, I found that some IT concepts are grouped together consistently in the automatic classifications but not in the ground truth. These IT concepts belong to two sets: public key infrastructure and virtual private network; data mining, data warehouse, and online analytical processing. In the ProQuest classification, public key infrastructure is grouped with electronic commerce, online advertising, online sales, and World Wide Web. According to definitions of the IT concepts in Appendix A, public key infrastructure is an important online security infrastructure for electronic commerce. Therefore, this group in the ProQuest classification is more of electronic commerce-related. However, in both co-occurrence-based and KL divergence-based classifications, public key infrastructure is together with virtual private network to form a group which is more of online security measure. As for the other set, data mining is grouped with business intelligence software, customer relationship management, and salesforce automation in the ProQuest classification. This group in the ProQuest classification is about enterprise ITs which may provide data-mining facilities. On the other hand, in both automatic classifications, data mining is with data warehouse and online analytical processing to form a group which is more of online data processing. In these cases, both automatic classifications, in my opinion, seem to have produced more relevant results than those in the ground truth. This finding suggests that the automatic approach can help update existing classifications in some way.

Fourth, I found that both of the automatic classifications tend to put general enterprise ITs in one large group. In the co-occurrence-based classification, group number 7 includes customer relationship management, electronic commerce, enterprise resource planning, inventory management, salesforce automation, and supply chain

management. However, the six IT concepts are from four different groups in the ProQuest classification. On the other hand, group number 2 in the KL divergence-based classification consists of application service provider, customer relationship management, electronic commerce, enterprise resource planning, inventory management, salesforce automation, supply chain management, web service, and world wide web. However, the 9 IT concepts are from five different groups in the ProQuest classification. In these cases, both automatic classifications seem to be able to identify "general" enterprise ITs and put them in a large group but fail to further distinguish subtle difference among them.

Fifth, both of the automatic classifications seem to have problems classifying certain IT concepts which include online advertising, online sales, virtualization, and world wide web. Take world wide web as an example, it is grouped with electronic commerce, online advertising, online sales, and public key infrastructure in the ProQuest classification. However, world wide web is grouped with artificial intelligence and robot in the co-occurrence-based classification, which looks odd. In addition, there are some singletons in both automatic classifications. In the co-occurrence-based classification, there are three singletons. On the other hand, there are four singletons in the KL divergence-based classification. The singletons in the automatic classifications suggest that they are so unique that they cannot be grouped with others when considering 14 clusters.

4.6 Summary

The empirical study of 35 IT concepts over ten years illustrates that both co-occurrence analysis and KL divergence can be utilized for classifying IT concepts. According to the F-measures, the automatic classification based on co-occurrence

analysis outperforms the classification based on KL divergence. In the next chapter, I further evaluate the effectiveness of the automatic classifications and conduct an online survey to determine which classification information professionals prefer. The long-term goal is to discover and develop effective new ways (that complement existing methods) to make sense of the dynamic, complex relationships among IT concepts.

Chapter 5: Evaluating the Two Automatic Classifications of IT Concepts with a Survey

In this chapter, I conduct a survey to have information professionals evaluate the three classifications of the same 35 IT concepts. One classification is based on the ProQuest Controlled Vocabulary of Subject Terms, which is considered as ground truth in the study. The other two are the automatic classifications based on co-occurrence analysis and KL divergence respectively. As the survey involves human subjects, an IRB application was submitted and approved. See Appendix B for the IRB protocol approval.

5.1 Introduction

The targeted research subjects of the survey are the subscribers to the AIS World mailing list. AIS stands for Association for Information Systems, which aims to create and maintain a professional identity for educators, researchers, and professionals in the Information Systems. As a result, the prospective subjects could be IS researchers, or professionals, including faculty and (likely doctoral) students as well as practitioners in industry, non-profit organizations, and government. I sent a request email to the list and invited the subscribers to participate voluntarily in the survey. The email (see Appendix C) includes a link to the informed consent form which explains the study to potential subjects, tells them what they are expected to do, and informs them that they can withdraw from the survey at any time without penalty. The email also includes a link to the online survey, which begins with the essence of the informed consent form.

In the request email as well as the survey, the subjects were asked to evaluate three IT classifications. However, they were not notified that one of the three classifications is the ProQuest classification, which is considered as the ground truth in the study; only the other two classifications are automatically generated based on either co-occurrence analysis or KL divergence. When taking the survey, the subjects should assume they were evaluating three classifications. We did this because we were also evaluating the ProQuest classification's role as the ground truth.

5.2 Pilot Study

Before disseminating the survey to the AIS World mailing list, I did a pilot study with my fellow iSchool doctoral students. The pilot study was conducted in the end of July 2010. The pilot study helped me redesign the survey and present the three classifications for comparison more effectively. Originally in the pilot study, I presented the three classifications all together on one screen and asked subjects to pick the best and the worst one in a single question. This caused various difficulties in evaluating the three classifications especially when the survey was conducted on the Internet via a subject's computer. Below are three responses I received regarding the presentation of the three classifications in the pilot study: "*The way the three classifications were displayed made it hard to keep track of my thinking as I compared the three columns. I found myself really wanting to print these out and make notes…*", "*Impossible to assess. It is VERY hard to tell from the presentation, however, since the categories are not lined up for readability so it is difficult to compare across the classification groups…*", *and* "*The classification comparison required more cognitive effort than most people are willing to spend in a survey. As a researcher I would worry about the quality of the responses. I*

would search other formats to simplify the task to the participant, i.e. partition of the classifications in several pages, comparison of groups in pairs."

According to the pilot study, I redesigned the survey and compared the three classifications in pairwise so that the evaluation can be done in three steps: A versus B, B versus C, and A versus C. As a result, instead of picking the best and the worst among the three classifications in a single step, subjects were asked to pick a better one in each of the three pairwise comparisons. Because only two classifications appeared in each comparison, I was able to remove groups with exactly the same IT concepts from both classifications. This made each comparison easier. From the results of three pairwise comparisons, I was also able to identify which classification was the best and which was the worst from subjects' perspective. See Appendix D for the final version of the survey.

5.3 Survey Results

The survey was distributed to the AIS World mailing list in the end of January, 2011. It lasted for about one week until the last response was received. In the end, twenty-two responses were collected while twenty-one were considered as complete.

5.3.1 Time Spent on the Survey

For the 21 completed responses, the time spent by the subjects on the survey was calculated. There is one subject who spent extremely long time (over 58 hours) on the survey. It seems that the subject started the survey at some point but didn't finish, and came back several days later to submit it. As a result, such long time is not included in the following descriptive statistics. Table 5.1 indicates the statistics of the time spent on the survey from the 20 responses. In average, a subject spent 12.5 minutes on the survey

with a standard deviation of 7.5 minutes. Overall, the respondents spent enough time on the survey for quality responses.

Table 5.1: Statistics of the Time Spent on the Survey

Statistics	Time
Average	12:33
Standard deviation	7:38
Minimum	2:41
Maximum	30:41

5.3.2 Survey Respondents

In a section of the survey, the subjects were asked for their demographic information. Although the section was optional, nine subjects had provided their information. According to their responses, most respondents were professors. As for their area, most were in Information System. Table 5.2 and Table 5.3 show demographics of the respondents about their current position and area of degree respectively. In general, the respondents have adequate knowledge about the IT concepts in the survey.

5.3.3 Evaluation of the Three Classifications

The survey compares the classifications in pairwise. As a result, the evaluation among the three classifications was done in three steps: A versus B, B versus C, and A versus C. In each pairwise comparison, subjects were asked to not only pick a better one but explain the logic behind the choice. Below, I first summarize pros and cons for each classification from the responses. Then I apply nonparametric statistics to the data to identify which classification is the most or least preferred overall in the survey.

Table 5.2: Demographics of Respondents – Current Position

Current position	Number of people
Professor	6
Doctoral student	2
Research fellow	1

Table 5.3: Demographics of Respondents – Area of Degree

Area of Degree	Number of people
Information System	4
Management Information System	2
Computer Science	1
Internet	1
Business & IT	1

The ProQuest classification is treated as the ground truth in the study. However, it has its pros and cons according to the survey. For the cons, some respondents do not like the combination of application service provider and enterprise resource planning in ground number 2 of the classification. One states that: "*ASP and ERP are two very distinct topics*". Another one thinks ERP should belong to the other enterprise applications such as CRM, salesforce automation, and SCM, as it appears in the co-occurrence-based classification. As for the pros, some respondents like the group number 9 which includes electronic commerce, online advertising, online sales, public key infrastructure, and world wide web. However, in the other two automatic classifications, electronic commerce, online advertising, online sales, and world wide web do not appear together.

The most obvious con in the co-occurrence-based classification is probably in group number 3 in which world wide web is put together with artificial intelligence and

robot. Almost half of the respondents note this as a disadvantage of the classification. Another widely perceived con is in group number 9 where virtualization is placed together with linux and open source software. Many respondents do not understand the relationship between virtualization and the other two concepts in that group. As for the pros, some respondents like group number 5 and group number 7 in view of the other classifications. In the group number 5, business intelligence software and data mining are together with data warehouse and online analytical processing while in the ProQuest classification business intelligence software and data mining are with other enterprise applications such as customer relationship management and salesforce automation. As for the group number 7, it is mainly about enterprise IT applications including customer relationship management, enterprise resource planning, and supply chain management, etc. One respondent states that the group is good as "*it has a clearer theme*".

The most obvious con in the KL divergence-based classification is in group number 3 in which global positioning system is put together with artificial intelligence and robot. Four respondents note this as a disadvantage of the classification. The other con is in group number 4 where social networking is placed together with instant messaging and other concepts. One respondent states that "*social networking has nothing to do with IM, really*". Another con is in group number 2 where many unrelated IT concepts form a large conglomerate. For example, respondents state "*why is web service and www in that group?*" and "*web service doesn't belong with inventory management*". On the other hand, there is no obvious pro for the KL divergence-based classification except that one respondent prefers the concept virtualization to be on its own in the classification.

According to the summary above, it seems that the respondents prefer the ProQuest classification the most and the KL divergence-based classification the least. However, this conclusion is better supported with a statistics test. Given the nature of the data, I first applied a coding schema to the three classifications according to each respondent's evaluation in the three-step pairwise comparisons. If a classification is ranked the best, "1" is assigned to the classification. For the second place and the worst, "2" and "3" are assigned accordingly. If there is a tie, say between the second and the third place, "2.5" is assigned for both classifications. Therefore, the total of the three numbers assigned to the three classifications remains "6" for each response. After coding, I ran Friedman's test to identify the order of preference for the three classifications and to see if the difference of preference among the three is statistically significant. The test results are shown in Table 5.4 and Table 5.5 below.

Table 5.4: Mean Rank of the Three Classifications

IT Classification	Mean Rank
ProQuest classification	1.55
Co-occurrence-based classification	2.00
KL divergence-based classification	2.45

Table 5.5: Friedman's Test of the Evaluation of the Three Classifications

Friedman Test Statistics	
N	21
Chi-Square	11.108
df	2
Asymp. Sig.	0.004

According to the mean rank (Table 5.4), the ProQuest classification (1.55) is the most preferred while the KL divergence-based classification (2.45) is the least preferred

by the respondents. The Chi-square of 11.108 with a *p*-value of .004 (Table 5.5) indicates that the mean rank difference among the three classifications is statistically significant. In other words, in a statistically significant way, the ProQuest classification is the best while the KL divergence-based classification is the worst among the three classifications. A post-hoc analysis for the Friedman's test of mean ranks (Conover, 1980) shows a critical rank difference of 0.48, which suggests that the mean rank difference between the ProQuest classification (the best) and the KL divergence-based classification (the worst) mainly contributes to the significance in the Friedman's test.

5.4 Discussion

An open-ended question was asked in the survey to share thoughts on how IT concepts may be usefully classified. Some respondents state that there could be more than one classification system that is appropriate. They think that a good classification depends on the purpose or context. For example, one respondent wrote: "*cannot be stated without knowing the context that the classification will be used in. A useful classification always depends on who is supposed to use this classification. There are several 'correct' ways to classify*". In addition, there is one respondent who prefers a tagging system to a classification system as "*classification pigeon-holes items into a hierarchy when in fact, they related in a multi-dimensional matrix fashion*".

Based on their input, I have the following comments. First, given the nature of the two automatic classification methods, I could only state that our automatic classifications are *generic*. In the co-occurrence-based classification, relatedness between a pair of IT concepts is inferred from the frequency of co-occurrence in paragraphs. If a pair of IT concepts co-occurs more often in paragraphs, they tend to be grouped together in the

classification. As for the KL divergence-based classification, each IT concept is first represented by concatenating all of the paragraphs mentioning the concept. Then, divergence between a pair of IT concepts is calculated based on the different use of language between two sets of paragraphs, each set representing one of the two IT concepts. If a pair of IT concepts shares similar use of language between their own sets of paragraphs, they tend to be put together in the classification. As a result, the two automatic classifications of IT concepts are generic by their nature. They were not designed to be used in any particular context in the first place. However, they both provide a quick and easy way to help make sense of a given list of IT concepts by classifying so that "related" terms can be grouped together. When the two automatic classifications were evaluated together with the ProQuest classification, the ground truth, in the survey, all three classifications have their own pros and cons, suggesting that they can help improve the ground truth. Although the automatic classifications are generic, they can further be tweaked for a particular context given the objects in the list are representative.

Second, regarding the comment about a tagging system, while hierarchical clustering is mainly used in forming the automatic classifications, multidimensional scaling is provided as a complement throughout the study. Multidimensional scaling displays a map, usually in two dimensions, where relatedness of the objects can be inferred from their proximity on the map. Although the multidimensional scaling map is not truly in multi-dimensional matrix fashion and there is usually some distortion on the map as the dimensions are reduced to only two on the map, it provides a spatial representation of the data where objects are not related in hierarchy.

5.5 Summary

According to the Friedman's test of mean ranks, the ProQuest classification is the best among the three classifications in a statistically significant way. This justifies the use of the ProQuest classification as the ground truth in the study. In addition, the mean ranks of the three classifications correspond to the F-measures calculated in the previous chapter. Table 5.6 shows both mean rank and F-measure for the three classifications. A Pearson correlation coefficient calculated between mean rank and F-measure is -.973. This highly negative correlation coefficient indicates that the more similar to the ProQuest classification (the higher the F-measure), the better the classification (the less the mean rank). In this aspect, the co-occurrence-based classification which comes with a higher F-measure than the KL divergence-based classification is also a better classification according to the survey. Besides, the results also indicate that F-measure is a reliable and sensible measure to assess the similarity between the ground truth and an automatic classification.

Table 5.6: Mean Rank and F-measure of the Three Classifications

IT Classification	Mean Rank [a]	F-measure [b]
ProQuest classification	1.55	1.000
Co-occurrence-based classification	2.00	.729
KL divergence-based classification	2.45	.615

a. Rank ranges from 1 to 3 where 1 indicates the best among the three.

b. F-measure assesses the similarity between the ProQuest classification and one of the other two classifications. It equals 1 when two classifications are identical.

Chapter 6: Further Exploration of the Co-occurrence Analysis

In the previous chapters, the co-occurrence-based classification is better than the KL divergence-based classification according to either F-measure or the survey. In this chapter, I further explore the co-occurrence analysis and try a different treatment for the co-occurrence matrix before classifying. Following a procedure in ACA literature, I first generate a matrix of Pearson correlations from the co-occurrence matrix. Applying hierarchical clustering and factor analysis in parallel to the matrix of Pearson correlations, I create two co-occurrence-based classifications. The F-measure indicates that both the new co-occurrence-based classifications are better in agreeing with the ProQuest classification than the original co-occurrence-based classification in Chapter 4. Besides, the factor analysis shows some interesting results compared to the hierarchical clustering in classifying.

6.1 Introduction

According to McCain (1990), "Author cocitation analysis (ACA) is a set of data gathering, analytical, and graphic display techniques that can be used to produce empirical maps of prominent authors in various areas of scholarship" (p. 443). Its purpose is to recognize influential authors and show their interrelationships from the citation record (White & McCain, 1998).

ACA applies a series of techniques that are quite similar to those in the study. These are co-occurrence matrix, followed by multivariate analysis. For multivariate analysis, in addition to multidimensional scaling and hierarchical clustering, ACA

includes factor analysis. Between the two different applications with similar techniques, ACA observes co-occurrence of first authors in reference of a paper in field publication while the study examines co-occurrence of IT concepts in a paragraph of articles from the six magazines. Both applications employ multivariate analysis for classification purpose. However, this study aims to classify IT concepts within a list while ACA classifies authors to identify specialties within a discipline. For example, McCain (1990) identifies specialties in macroeconomics; White and McCain (1998) identifies specialties in information science.

Another difference between the two applications is treatment of a co-occurrence matrix. In ACA, second-order co-occurrence is always considered while the previous chapters use first-order co-occurrence. In the first-order co-occurrence, two terms are considered similar if they co-occur frequently directly with each other. However, in the second-order co-occurrence, two terms are considered similar if they have similar co-occurrence pattern with all other terms in the list. For example, in the second-order co-occurrence, A and B are considered similar when they both co-occur with C but seldom co-occur with others in the matrix regardless of how often they co-occur with each other.

In practice, many ACA researchers convert a co-occurrence matrix into a matrix of Pearson correlations when considering the second-order co-occurrence (McCain, 1990). According to McCain (1990), the creation of a correlation matrix has an advantage that the correlation coefficient removes differences in “scale” between objects which occur more frequently and those which have similar profiles but occur less frequently.

6.2 Method

Figure 6.1 shows the flow chart of the data analysis. In Chapter 4, the co-occurrence matrix is processed directly as a proximity matrix (first-order co-occurrence). In this chapter, it is treated as raw data and converted into a matrix of Pearson correlations (second-order co-occurrence). Then, I apply hierarchical clustering and factor analysis in parallel to the matrix of Pearson correlations to create two classifications. At last, I use F-measure to assess the similarity between each of the two automatic co-occurrence-based classifications and the ProQuest classification.

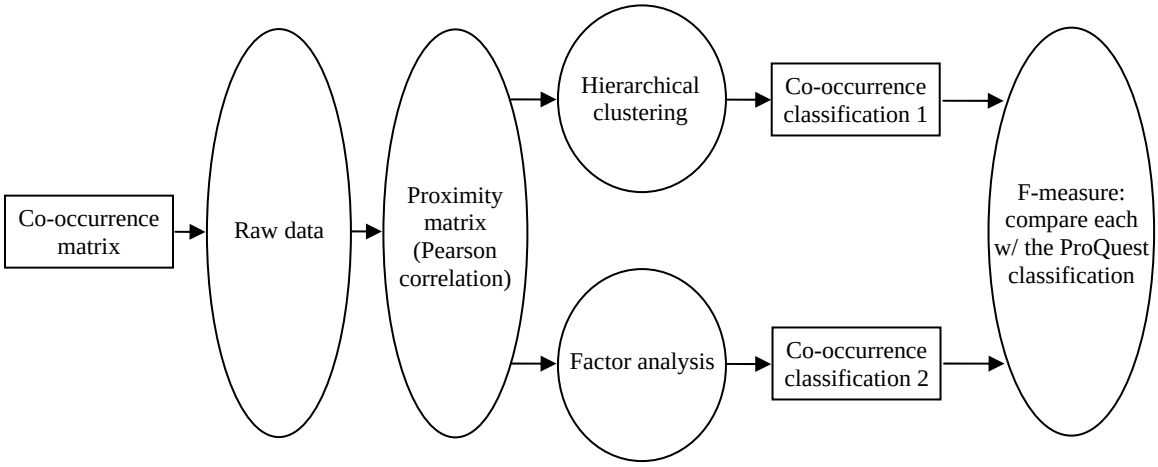

Figure 6.1: Flow Chart of the Data Analysis

Like hierarchical clustering and multidimensional scaling, factor analysis is a classification technique (Sokal, 1974). It attempts to describe interrelationships among observed variables through the creation of a lower number of derived variables or factors. In the study, factors are extracted by principal components analysis with varimax rotation as this produces factors that are uncorrelated, with most objects having high loadings on only one factor.

6.3 Results

The hierarchical clustering result on the matrix of Pearson correlations is in Figure 6.2. The classification of IT concepts is generated by obtaining 14 clusters from the result (Table 6.2). Comparing this classification with the ProQuest classification (Table 6.1) results in an F-measure of 0.830. Besides, 7 out of 14 groups are identical to those in the ProQuest classification.

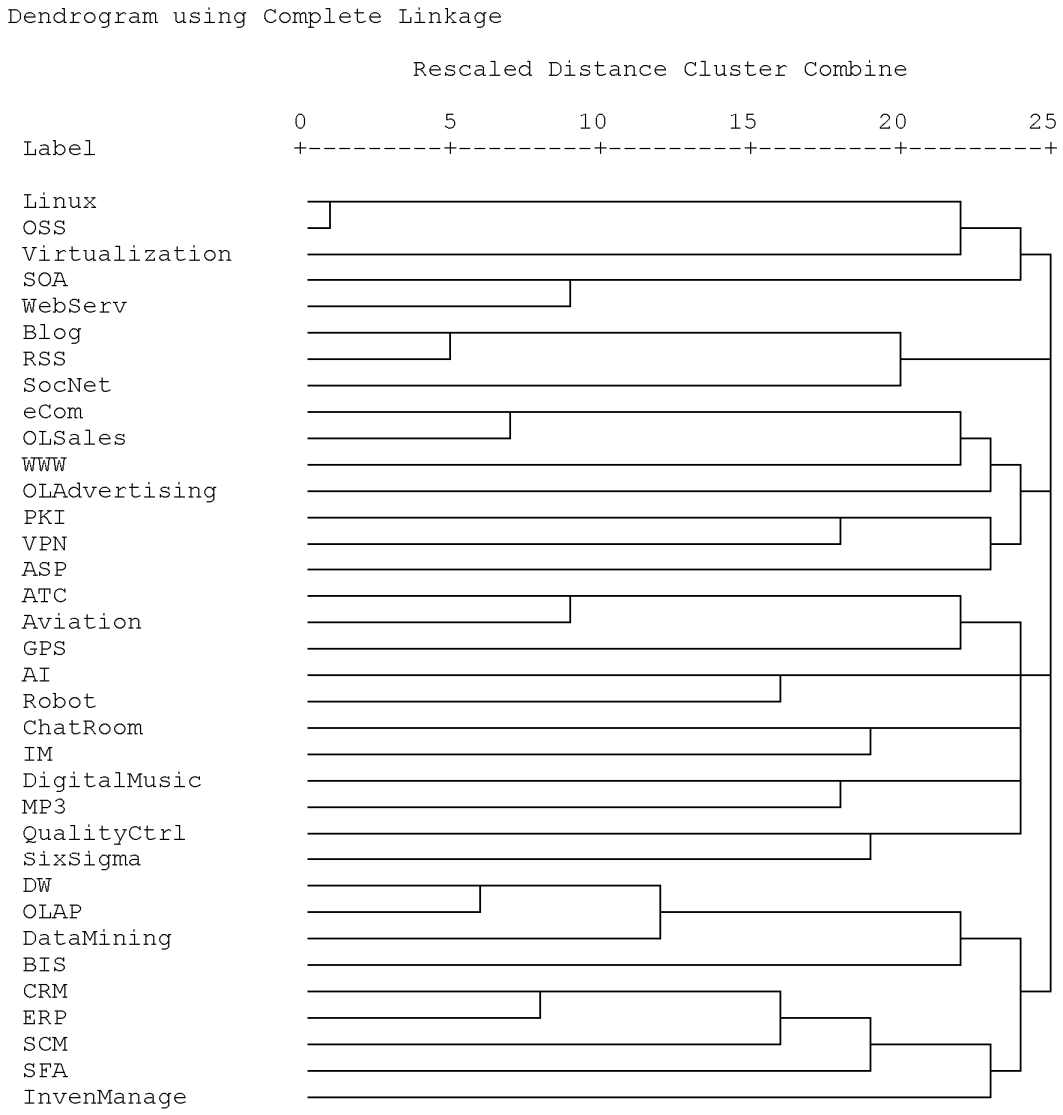

Figure 6.2: Hierarchical Clustering Result on the Matrix of Pearson Correlations

The multidimensional scaling result on the matrix of Pearson correlations is in Figure 6.3. It is shown only as a complement as there is no clear demarcation of groups for classification. The R^2 goodness-of-fit value for two dimensions is very high (0.990). However, the plot is not considered good as many related IT concepts are not close to each other. For example, the 9 pairs in Table 4.7 can be easily located on both the co-occurrence-based and KL divergence-based MDS plots (see Figure 4.7 and Figure 4.8). However, most of the pairs cannot be easily found on the plot.

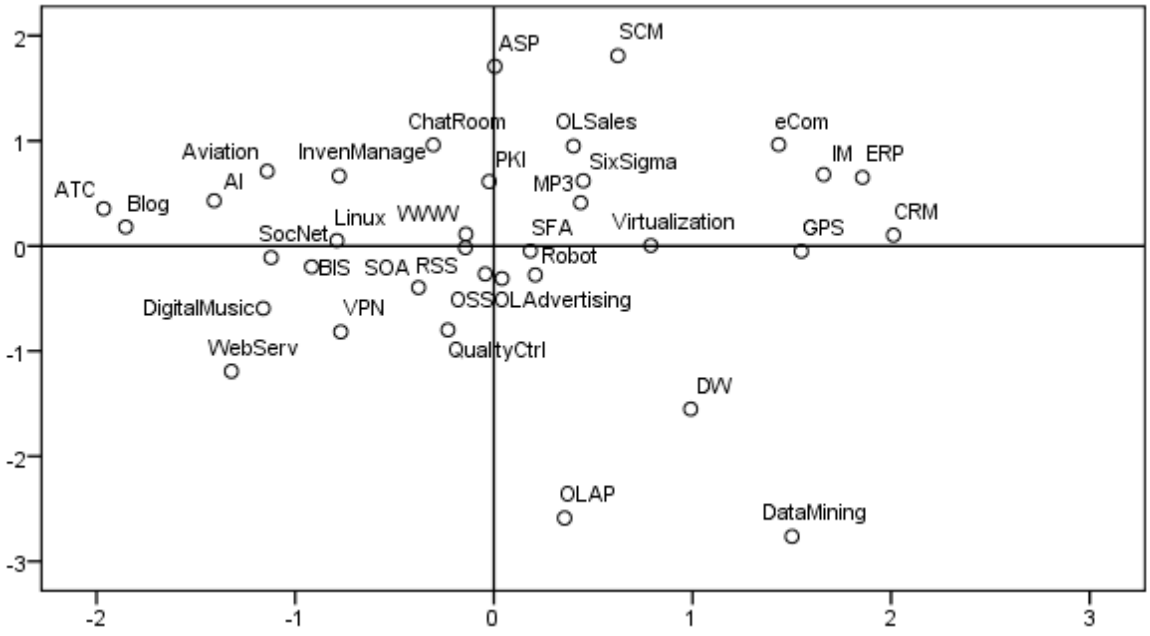

Figure 6.3: Multidimensional Scaling Result on the Matrix of Pearson Correlations

As for the factor analysis, Figure 6.4 shows the scree plot. For the number of factors to extract, I used Cattell scree test to determine 11 factors. Varimax rotation was used after extraction.

Table 6.3 shows the factor analysis result. The conventional reporting threshold of loading is 0.40. However, for classification purpose here, when loadings of an IT concept

across the factors are below the threshold, the highest loading of the IT concept is reported (in gray). If this is done, the factors in Table 6.3 show “simple structure”, in which each IT concept loads on only one factor. “Simple structure” is ideal for classification (non-overlapping) as the relationship between objects and factors is unambiguous.

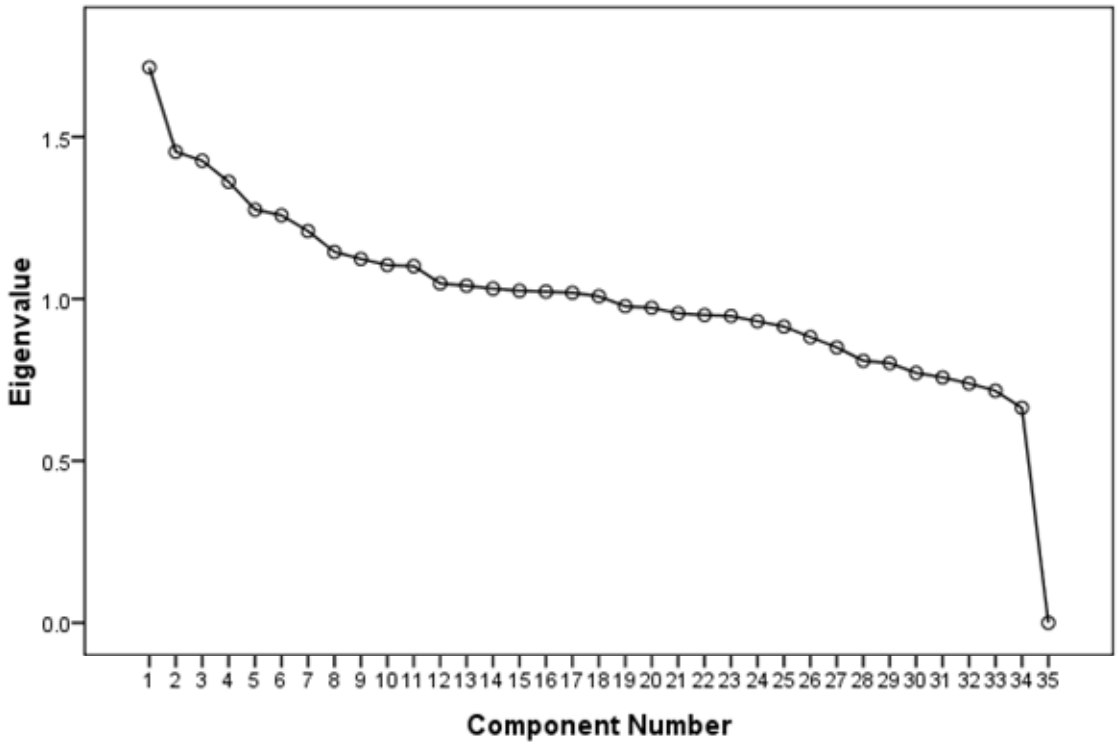

Figure 6.4: Scree Plot

Among the eleven factors, there is one factor with which some IT concepts load positively while others load negatively. As a result, 12 groups are identified from the analysis. Table 6.4 shows the resulting classification in 12 classes. In the classification, loadings are displayed next to IT concepts. Unlike other classifications in which IT concepts are sorted alphabetically, the IT concepts in the classification are sorted by their

loadings in descending order. In addition, those IT concepts with loadings below the threshold are marked gray.

Comparing this classification with the ProQuest classification (Table 6.1) results in an F-measure of 0.797. Besides, 7 out of 12 groups are identical to those in the ProQuest classification.

Table 6.1: IT Concepts in the ProQuest Classification

No.	IT concepts
1	AI, Robot
2	ASP, ERP
3	ATC, Aviation, GPS
4	BIS, CRM, DataMining, SFA
5	Blog, RSS, SocNet
6	ChatRoom, IM
7	DigitalMusic, MP3
8	DW, OLAP
9	eCom, OLAdvertising, OLSales, PKI, WWW
10	InvenManage, SCM
11	Linux, OSS
12	QualityCtrl, SixSigma
13	SOA, WebServ
14	Virtualization, VPN

Table 6.2: The Second-order Co-occurrence-based Classification by Clustering

No.	IT concepts
1	AI, Robot
2	ASP
3	ATC, Aviation, GPS
4	BIS, DataMining, DW, OLAP
5	Blog, RSS, SocNet
6	ChatRoom, IM
7	CRM, ERP, SCM, SFA
8	DigitalMusic, MP3
9	eCom, OLAdvertising, OLSales, WWW
10	InvenManage
11	Linux, OSS, Virtualization
12	PKI, VPN
13	QualityCtrl, SixSigma
14	SOA, WebServ

Note: **7** out of 14 groups are identical (F=.830)

Table 6.4: The Second-order Co-occurrence-based Classification by Factor Analysis

No.	IT concepts (loading)
1	AI (-0.715), Robot (-0.679)
2	ATC (-0.742), Aviation (-0.718), GPS (-0.111)
3	OLAP (0.708), DW (0.656), DataMining (0.598), BIS (0.241)
4	Blog (0.745), RSS (0.677), SocNet (0.466)
5	ChatRoom (-0.675), IM (-0.651)
6	CRM (0.644), ERP (0.627), SCM (0.496), SFA (0.475), InvenManage (0.121)
7	MP3 (0.283), DigitalMusic (0.263)
8	OLSales (0.715), eCom (0.712), WWW (0.266), OLAdvertising (0.191), ASP (0.167)
9	Linux (0.785), OSS (0.782), Virtualization (0.252)
10	VPN (0.717), PKI (0.669)
11	SixSigma (-0.666), QualityCtrl (-0.661)
12	SOA (0.763), WebServ (0.735)

Note: **7** out of 12 groups are identical (F=.797)

Table 6.3: Factor Analysis of the 35 IT Concepts

IT Concept	Component										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
CRM	.644										
ERP	.627										
SCM	.496										
SFA	.475										
InvenManage	.121										
OLAP		.708									
DW		.656									
DataMining		.598									
BIS		.241									
Linux			.785								
OSS			.782								
Virtualization			.252								
Blog				.745							
RSS				.677							
SocNet				.466							
OLSales					.715						
eCom					.712						
WWW					.266						
OLAdvertising					.191						
ASP					.167						
ATC						-.742					
Aviation						-.718					
GPS						-.111					
SOA							.763				
WebServ							.735				
AI								-.715			
Robot								-.679			
VPN									.717		
PKI									.667		
SixSigma										-.666	
QualityCtrl										-.661	
DigitalMusic										.263	
MP3										.283	
ChatRoom											-.675
IM											-.651
Eigenvalues	1.51	1.50	1.41	1.39	1.31	1.26	1.25	1.16	1.14	1.13	1.11
% Variance	4.32	4.29	4.02	3.97	3.75	3.60	3.58	3.31	3.26	3.22	3.17

Table 6.5: The First-order Co-occurrence-based Classification

No.	IT concepts
1	AI, Robot, WWW
2	ASP
3	ATC, Aviation, GPS
4	BIS, DataMining, DW, OLAP
5	Blog, RSS, SocNet
6	ChatRoom, IM
7	CRM, eCom, ERP, InvenManage, SCM, SFA
8	DigitalMusic, MP3
9	Linux, OSS, Virtualization
10	OLAdvertising
11	OLSales
12	PKI, VPN
13	QualityCtrl, SixSigma
14	SOA, WebServ

Note: **6** out of 14 groups are identical (F=.729)

Table 6.2: The Second-order Co-occurrence-based Classification by Clustering (Repeated for Comparison)

No.	IT concepts
1	AI, Robot
2	ASP
3	ATC, Aviation, GPS
4	BIS, DataMining, DW, OLAP
5	Blog, RSS, SocNet
6	ChatRoom, IM
7	CRM, ERP, SCM, SFA
8	DigitalMusic, MP3
9	eCom, OLAdvertising, OLSales, WWW
10	InvenManage
11	Linux, OSS, Virtualization
12	PKI, VPN
13	QualityCtrl, SixSigma
14	SOA, WebServ

Note: **7** out of 14 groups are identical (F=.830)

Table 6.4: The Second-order Co-occurrence-based Classification by Factor Analysis (Repeated for Comparison)

No.	IT concepts (loading)
1	AI (-0.715), Robot (-0.679)
2	ATC (-0.742), Aviation (-0.718), GPS (-0.111)
3	OLAP (0.708), DW (0.656), DataMining (0.598), BIS (0.241)
4	Blog (0.745), RSS (0.677), SocNet (0.466)
5	ChatRoom (-0.675), IM (-0.651)
6	CRM (0.644), ERP (0.627), SCM (0.496), SFA (0.475), InvenManage (0.121)
7	MP3 (0.283), DigitalMusic (0.263)
8	OLSales (0.715), eCom (0.712), WWW (0.266), OLAdvertising (0.191), ASP (0.167)
9	Linux (0.785), OSS (0.782), Virtualization (0.252)
10	VPN (0.717), PKI (0.669)
11	SixSigma (-0.666), QualityCtrl (-0.661)
12	SOA (0.763), WebServ (0.735)

Note: **7** out of 12 groups are identical (F=.797)

6.4 Discussion

First, when comparing the two second-order co-occurrence-based classifications with the ProQuest classification (see Table 6.1, 6.2, and 6.4), both automatic classifications have the same 7 groups which are identical to those in the ProQuest classification. The second-order co-occurrence-based classification by clustering has a better F-measure (0.830) than the other classification by factor analysis (0.797).

Second, when comparing the two second-order co-occurrence-based classifications with the first-order co-occurrence-based classification in chapter 4 (see Table 6.2, 6.4, and 6.5), both the second-order co-occurrence-based classifications have a better F-measure than the first-order co-occurrence-based classification (0.729). When comparing the first-order co-occurrence-based classification (Table 6.5) and the second-order co-occurrence-based classification by clustering (Table 6.2), the difference is the five IT concepts: world wide web, electronic commerce, inventory management, online advertising, and online sales. The second-order co-occurrence-based classification by clustering places world wide web, electronic commerce, online advertising, and online sales together in a group, and inventory management alone in another group. When comparing the second-order co-occurrence-based classification by clustering (Table 6.2) and the second-order co-occurrence-based classification by factor analysis (Table 6.4), the difference is only the two IT concepts: inventory management and application service provider. The second-order co-occurrence-based classification by factor analysis groups inventory management with the CRM group, and application service provider with the eCom group. On the other hand, the two IT concepts are singletons in the second-order co-occurrence-based classification by clustering.

Interestingly, compared to the first-order co-occurrence-based classification, both the two second-order co-occurrence-based classifications put the four IT concepts together: world wide web, electronic commerce, online advertising, and online sales. This is why the two second-order co-occurrence-based classifications have a higher F-measure than the first-order co-occurrence-based classification as the four IT concepts are in fact in the same group in the ProQuest classification. Table 6.6 shows the co-occurrence matrix of the four IT concepts. The co-occurrence counts among online advertising, online sales, and world wide web are low. The co-occurrence counts between electronic commerce and the three concepts are high but the number of occurrence of electronic commerce is extremely high (12,311). As a result, it explains why the first-order co-occurrence-based classification fails to group the four concepts together. On the other hand, with information of each IT concept's co-occurrence "pattern" with other concepts in the list, the second-order co-occurrence-based classifications are able to group them together.

Table 6.6: Co-occurrence Matrix of the eCom Group

	eCom	**OLAdvertising**	**OLSales**	**WWW**
eCom	12311	28	191	141
OLAdvertising	28	473	3	5
OLSales	191	3	742	6
WWW	141	5	6	1411

Factor analysis, like multidimensional scaling, is an ordination technique (Sokal, 1974). Although the factors in Table 6.3 display "simple structure", in which each IT concept loads on only one factor. It is not usually the case in factor analysis; some objects could load on more than one factor while others with loadings below the threshold are not

reported. Generally, factor analysis is to interpret extracted factors with those objects loading on them. However, like other multivariate analysis in the study, factor analysis is used as a classification technique to classify objects. It provides a middle ground between hierarchical clustering and multidimensional scaling that one gives a hierarchic classification and the other offers a spatial representation of objects without a clear demarcation of groups.

In the factor analysis of the 35 IT concepts (Table 6.3), none of them loads on more than one factor. On the other hand, there are 9 IT concepts with loadings below the threshold. They are still reported for classification but marked gray in the analysis (Table 6.3) as well as in the resulting classification (Table 6.4). These 9 IT concepts and their loadings are: GPS (-0.111), BIS (0.241), InvenManage (0.121), MP3 (0.283), DigitalMusic (0.263), WWW (0.266), OLAdvertising (0.191), ASP (0.167), and Virtualization (0.252). The membership of the concepts is in gray area and it is very likely that in other classifications they are put in a different group. For example, application service provider is in the group of electronic commerce in the factor analysis but it is a singleton both in the first-order co-occurrence-based classification (Table 6.5) and in the second-order co-occurrence-based classification by clustering (Table 6.2). Another example is virtualization, which is grouped with linux and open source software across the three co-occurrence-based classifications (Table 6.2, 6.4, and 6.5). However, in the survey (Chapter 5), virtualization's membership with linux and open source software is considered as a major disadvantage.

Finally, factor analysis complements with other multivariate analysis including hierarchical clustering and multidimensional scaling. Comparing factor analysis and

hierarchical clustering, the loadings of objects in factor analysis correspond to the agglomeration schedule in hierarchical clustering. Take the group of electronic commerce (Group number 8 in Table 6.4) as an example. The loadings and dendrogram section for the group are in Table 6.7 and Figure 6.5 respectively. In Figure 6.5, online sales and electronic commerce, with the highest two loadings in the group, are first joined. World wide web, with the third highest loading, is later joined followed by online advertising with the fourth highest loading. As for application service provider with the lowest loading, it is not in the dendrogram section. In the classification based on the dendrogram (see Table 6.2), it is a singleton.

Table 6.7: The Loadings for the eCom Group

8	OLSales (0.715), eCom (0.712), WWW (0.266), OLAdvertising (0.191), ASP (0.167)

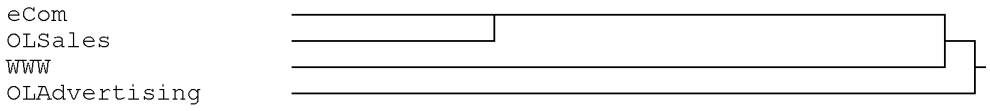

Figure 6.5: The Dendrogram Section for the eCom Group

Comparing factor analysis and multidimensional scaling, the loadings of IT concepts in factor analysis also correspond to their location on the plot. As noted in the multidimensional scaling result of this chapter, the plot based on second-order co-occurrence is not good because many related IT concepts are not close to each other on the plot. Therefore, I use the plot based on first-order co-occurrence (see Figure 4.4) to compare with factor analysis. Interestingly, the 9 IT concepts with loadings below the threshold in factor analysis appear as an outlier on the plot except that world wide web

(0.266) is located close to the center of the plot. This demonstrates that the three multivariate analysis techniques correspond to each other in some ways and they should complement with each other in exploring the relationships of IT concepts.

6.5 Summary

In the chapter, I apply second-order co-occurrence by converting the co-occurrence matrix into a matrix of Pearson correlations. Then, I use hierarchical clustering and factor analysis in parallel to create the two second-order co-occurrence-based classifications. Compared to the first-order co-occurrence-based classification in Chapter 4, both the classifications are able to group electronic commerce, online advertising, online sales, and world wide web together, the way the ProQuest classification does. As a result, both have a higher F-measure than the first-order co-occurrence-based classification. Besides, factor analysis has shown some advantages in classifying. The loadings in factor analysis are related to not only the agglomeration schedule in hierarchical clustering but also the location of IT concepts on the multidimensional scaling plot. The results show that the three multivariate analysis techniques should complement with each other in classifying IT concepts. Lastly, classifications based on a matrix of Pearson correlations may not always be better. For example, the multidimensional scaling plot in Figure 6.3 is not as good as that based on first-order co-occurrence (see Figure 4.4).

Chapter 7: Conclusions

7.1 Introduction

In the current status of the IT concept literature, most studies employ single-concept research designs, leaving the relationships among IT concepts underexplored. On the other hand, the few multi-concept studies have had to rely on domain experts to evaluate IT concept relationships. Such expert evaluations are difficult to replicate, to generalize to other IT concepts, or to scale up to examine the relationships among a large number of IT concepts.

In the dissertation, classification is used to study the relationships of IT concepts. The first step in classification is to construct a proximity matrix which consists of similarity or dissimilarity measures between a pair of objects. Generally, a proximity matrix can be generated in two ways. One is by derived measures which can be inferred from raw data coded in a matrix of objects by attributes. The other is by direct measures which can be pairwise comparisons of objects in terms of similarity or dissimilarity by human judges. Both ways are not scalable because they have to rely on domain experts. Therefore, the dissertation applies either KL divergence or co-occurrence analysis as a computational approach to construct a proximity matrix automatically for classification.

Classification techniques include hierarchical clustering and multidimensional scaling. The two techniques are used regularly throughout the study to analyze a proximity matrix and to classify IT concepts. Through a series of empirical studies, the

dissertation aims to provide a computational approach to study multiple IT concepts in terms of their similarity and relationships.

7.2 Summary of the Empirical Studies and Findings

The dissertation contains five empirical studies, unified under one overall research question: How can the relationships among multiple IT concepts be described and analyzed in a representative, dynamic, and scalable way?

In the first empirical study, I employ KL divergence to compare the semantic similarity of forty-seven IT concepts discussed in a trade magazine during a ten-year period. Using hierarchical clustering, I found that the similarity of the concepts can be mapped in a hierarchy and similar IT concepts demonstrate similar discourses. The results show that KL divergence can be utilized to construct a proximity matrix of IT concepts for classification.

In the second study, I employ co-occurrence analysis to explore the relationships among fifty IT concepts discussed in six magazines over the same ten-year period. Using hierarchical clustering and multidimensional scaling, I am able to identify general patterns similar to those found in the first study, but with interesting nuances. The results show that co-occurrence analysis can also be used to construct a proximity matrix of IT concepts for classification.

The findings from the two studies imply reasonable validity of the computational approach. In the third study, to validate and evaluate the effectiveness between KL divergence and co-occurrence analysis in classifying IT concepts I make use of the ProQuest classification as the ground truth and compare the two automatic IT classifications against it. In practice, I use an F-measure to assess the similarity between

two classifications of the same IT concepts. The F-measures show that the co-occurrence-based classification (F=.729) outperforms the KL divergence-based classification (F=.615) in agreeing with the ground truth.

In the fourth study, I conduct a survey to compare the three classifications: the two automatic classifications and the ground truth. The results show that the difference among the three classifications is statistically significant. The ProQuest classification is the best, and the co-occurrence-based classification is better than the KL divergence-based classification. The results correspond to those in the third study. They also justify the use of the ProQuest classification as the ground truth, and the use of F-measure as a similarity measure between two classifications (the higher the F-measure is, the better the classification is).

In the fifth study, I further explore the co-occurrence analysis with second-order co-occurrence. In practice, I construct a matrix of Pearson correlations from the co-occurrence matrix and use it as a proximity matrix for IT classification. Both hierarchical clustering and factor analysis produce an IT classification with a better F-measure than the first-order co-occurrence-based classification generated in Chapter 4. However, a classification based on second-order co-occurrence may not always show improvement. For example, the multidimensional scaling plot based on second-order co-occurrence is not as good as that based on first-order co-occurrence.

7.3 Enriched Research Method

After a series of empirical studies, an enriched research method can be shown in Figure 7.1. The enriched research method aims to provide a solid and sound approach in studying the relationships of IT concepts in the future. It is described in detail below.

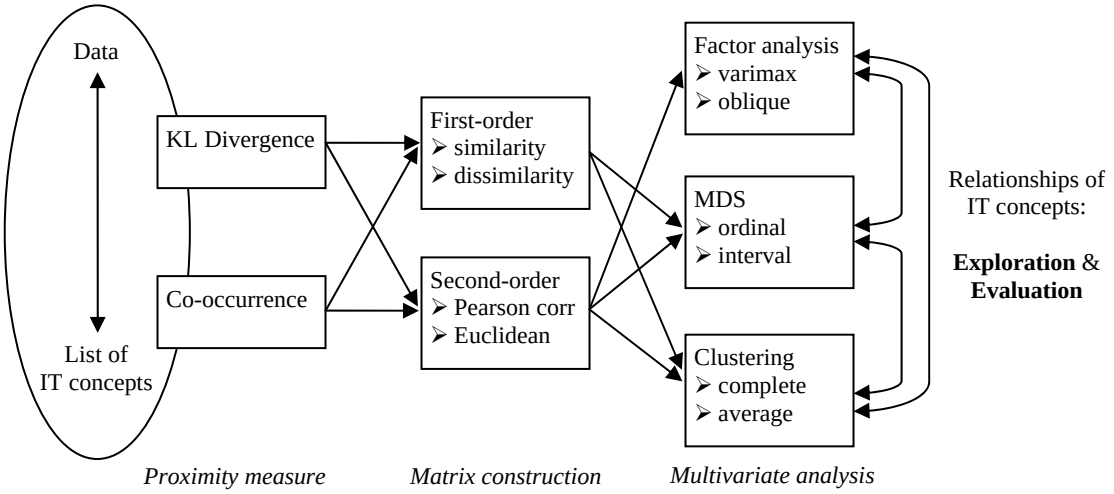

Figure 7.1: Enriched Research Method

First, it starts with a list of IT concepts and data collection. For a better representation, the data should be IT-related discourse, over a long period of time, and involving different aspects of topics. For example, the data collection in the dissertation includes two IT trade magazines, two business magazines, and two news magazines from 1998 to 2007. As the results indicate, the data collection should be representative enough in studying the relationships of IT concepts.

Second, either KL divergence or co-occurrence can be applied as a proximity measure for a pair of IT concepts. KL divergence compares the use of language between two sets of paragraphs while co-occurrence is about the frequency of paragraphs where two IT concepts co-occur. According to the results, the co-occurrence-based classification seems to be better than the KL divergence-based classification when considering first-order matrix construction (see next step).

Third, as either KL divergence or co-occurrence can serve as a proximity measure of IT concepts, there are two ways to construct a proximity matrix: first-order and

second-order. When considering the first-order matrix construction, the matrix of KL divergence or co-occurrence is treated as a proximity matrix. The similarity or dissimilarity between a pair of IT concepts is directly inferred from either proximity measure. However, in the second-order matrix construction, the proximity matrix is created based on a matrix of either proximity measure. The similarity or dissimilarity between a pair of IT concepts is inferred from respective proximity-measure profiles, each with all other concepts. In Chapter 6, Pearson correlation is applied as a second-order matrix construction while there are many other ways. According to the results, the second-order co-occurrence-based classifications in Chapter 6 are better than the first-order co-occurrence-based classification developed in Chapter 4.

Fourth, multivariate analysis is used to analyze the proximity matrix from the last step. Factor analysis, multidimensional scaling, and hierarchical clustering are multivariate analysis, and they are also classification techniques. Each technique has its own specifications. With a different specification, a technique generates similar classification results, but with some nuances. For example, complete link and average link are the commonly used aggregation criteria in hierarchical clustering. IT concepts placed differently by these two criteria may in fact have both strong links to two different clusters. According to the results, the three techniques complement with each other and can all be applied to a same proximity matrix. Among the three, factor analysis seems to show some advantages with loadings of IT concepts. The loading information in factor analysis mainly helps identify ambiguous relationship between IT concepts and factors. There are two types of ambiguous relationship factor analysis can identify. The first type is when IT concepts load (above the threshold) on more than one factor. On a

multidimensional scaling plot, these concepts tend to be placed near the center of the plot in order to relate to more than one factor. The second type is when concepts do not load on any factor. This is the type of ambiguous relationship that is observed for some concepts in Chapter 6. As noted, these concepts are more likely to be placed near the boundary on a multidimensional scaling plot. In addition, these concepts tend to be joined in later phases of agglomeration schedule in hierarchical clustering. They are also the main cause of nuances between different aggregation criteria in hierarchical clustering. This demonstrates that factor analysis plays a crucial role among the three classification techniques in studying the relationships of IT concepts.

Finally, as there are two proximity measures, two matrix construction methods, and three classification techniques in the enriched research method and each component has more than one specification, various classification results can be created. In one aspect, some are better than others. For example, Chapter 4 and Chapter 5 show that the co-occurrence-based classification is better than the KL divergence-based classification, and Chapter 6 shows that the second-order co-occurrence-based classifications are better than the first-order co-occurrence-based classification. In another aspect, they are usually meaningful in some ways and can be complement with each other. For example, Chapter 6 demonstrates that the multidimensional scaling plot based on first-order co-occurrence (Figure 4.4) can complement with the factor analysis based on second-order co-occurrence (Table 6.4) in identifying ambiguous relationship between IT concepts and factors. Therefore, the final step in the enriched research method focuses on exploration and evaluation of various classification results. The goal is to explore the relationships of

IT concepts and to evaluate and find a classification that best represents their relationships.

7.4 Contributions

The scalable computational approach demonstrated in the empirical studies is useful to help expand IT research along three dimensions: concepts, data sources, and time. First, the approach can help IT researchers overcome the limitation of single-concept designs by concurrently exploring multiple concepts. By facilitating the study of multiple concepts and their relationships over time, the approach enables researchers to develop theories of IT research on a more realistic footing.

Second, the approach is not limited to a specific source. Any source may be biased by its own specifics and thus it would be useful to collect data from multiple sources and apply the approach. On one hand, a study may discover a more objective representation of the concept relationships by pooling the data from multiple sources in proportions that may represent key constituencies of concept communities. On the other hand, researchers may analyze the data collected from each source and compare the results, revealing similarities and differences among various segments of the concept communities that different sources represent. For example, a previous study (Tsui, Wang, Fleischmann, Oard, & Sayeed, 2009) found that the IT concept relationships discovered in InformationWeek are very similar to those found in ComputerWorld. Comparing the results using the same approach across different sources will enrich understanding of IT concepts as well as the communities underlying these sources.

Third, due to the dynamic nature of concept relationships, it would be useful to conduct the approach at multiple times. The evolving hierarchical structure of IT

concepts would reveal what is going on, as concepts with changing meanings might cluster at different times with different concepts. To illustrate the utility of the methodology in multi-period analysis, I sliced the data by year and performed the same analysis on each year's data. The result shows that the hierarchical structure of concepts did change over time. However, some concepts such as e-business and e-commerce are clustered together throughout the years.

In sum, this scalable computational methodology enables multi-concept, multi-source, and multi-period studies, potentially advancing knowledge about the dynamic relationships of IT concepts.

7.5 Limitations of the Study

As the computational approach for IT classification is based on data, different data would result in a different classification under the same approach. Therefore, the data should be representative for the IT concepts under study. For example, the data in the study includes multiple discourses in which various IT concepts are discussed, described, and related to each other from various aspects. Besides, the scale of the data should be large enough to avoid data sparseness problem. In the empirical studies in which either KL divergence or co-occurrence analysis is applied, data sparseness could be an issue. For example, in the co-occurrence analysis, to fully recognize an IT concept and relate it with other concepts for their relationships, enough occurrences of the concept should be observed to have a valid co-occurrence profile with other concepts. The occurrence of an IT concept is measured by the number of paragraphs mentioning the concept. In Chapter 4, I try to avoid data sparseness by excluding those IT concepts with too few occurrences in paragraphs. Out of 120 IT concepts, I use their media

(251.50) as a threshold and pick the upper 60 IT concepts. The threshold is used in the study. However, it could be generalized to other similar studies to avoid data sparseness. Finally, to avoid data sparseness regardless of the threshold, a common way for this is always to collect more data. With more relevant data, it would be more possible that a proximity-measure profile of an IT concept with other concepts is meaningful.

7.6 Avenues for Future Study

The dissertation offers some insights to future studies that aim to develop a more robust way in studying the relationships of IT concepts. First, as discussed in the enriched research method, there are multiple paths in the procedure of classifying IT concepts. Each component in the method has more than one specification. A future study may try different paths or different specifications that haven't been used in the study. For example, one can try second-order KL divergence-based classification and see if it makes any improvement in terms of F-measure as in the co-occurrence analysis.

Second, in light of the limitations stated above regarding data sparseness, a future study could include additional trade magazines and other sources such as popular press, newspapers, academic journals, and informal media. With more data, the list of IT concepts can be expanded as well, to make the list more diversified and more representative in terms of the IT field. The three classification techniques together should be able to identify not only obvious but subtle relationships among such list of IT concepts.

7.7 Concluding Remarks

Due to their complication and varieties, IT concepts are related in complex ways. This dissertation provides a computational approach to study the relationships of IT concepts. It includes a series of empirical studies to explore IT relationships from different aspects such as on a plot or in a hierarchy, and to evaluate various IT classifications generated in the approach. The results have shown that the approach is capable of producing an automatic IT classification which is very much similar to the ground truth. Different methods are also compared with each other to gain insights regarding the relationships of IT concepts. However, the computational approach in the study does not aim to get rid of experts. Expert input is not scalable, but the approach can never replace experts. The values of the approach lie in the possibilities that the results from the approach may help experts in useful ways: saving time, finding patterns more easily, and seeing other alternatives, etc. The ultimate goal is to combine experts and computers in the IT-concept classification work so that manual and computational analyses complement each other to produce the best results more efficiently and effectively. Overall, the dissertation establishes a good foundation for studying the relationships of IT concepts in a representative, dynamic, and scalable way.

Appendix A: Definition of the 35 IT concepts

All definitions below are retrieved from Dictionary.com with various sources.

IT concept	Definition	Source
artificial intelligence	the capacity of a computer to perform operations analogous to learning and decision making in humans, as by an expert system, a program for CAD or CAM, or a program for the perception and recognition of shapes in computer vision systems. Abbreviation: AI, A.I. Origin: 1965–70	Dictionary.com Unabridged
application service provider	(ASP) A service (usually a business) that provides remote access to an application program across a network protocol, typically HTTP. A common example is a website that other websites use for accepting payment by credit card as part of their online ordering systems. As this term is complex-sounding but vague, it is widely used by marketroids who want to avoid being specific and clear at all costs.	The Free On-line Dictionary of Computing
air traffic control	a government service that facilitates the safe and orderly movement of aircraft within and between airports by receiving and processing data from radar and devices that monitor local weather conditions and by maintaining radio contact with pilots. Origin: 1930–35	Dictionary.com Unabridged
aviation	the design, development, production, operation, and use of aircraft, esp. heavier-than-air aircraft. Origin: 1865–70	Dictionary.com Unabridged
business intelligence	the process of gathering information about a business or industry matter; a broad range of applications and technologies for gathering, storing, analyzing, and providing access to data to help make business	Dictionary.com's 21st Century Lexicon

software	decisions; abbr. BI	
blog	Full name: weblog a journal written on-line and accessible to users of the internet	Collins English Dictionary - Complete & Unabridged 10th Edition
chat room	a branch of a computer network in which participants can engage in real-time discussions with one another.	Dictionary.com Unabridged
customer relationship management	(CRM, CIS, Customer Information Systems, Customer Interaction Software, TERM, Technology Enabled Relationship Manager) Enterprise-wide software applications that allow companies to manage every aspect of their relationship with a customer. The aim of these systems is to assist in building lasting customer relationships - to turn customer satisfaction into customer loyalty. Customer information acquired from sales, marketing, customer service, and support is captured and stored in a centralised database. The system may provide data-mining facilities that support an opportunity management system. It may also be integrated with other systems such as accounting and manufacturing for a truly enterprise-wide system with thousands of users.	The Free On-line Dictionary of Computing
data mining	Analysis of data in a database using tools which look for trends or anomalies without knowledge of the meaning of the data. Data mining was invented by IBM who hold some related patents. Data mining may well be done on a data warehouse.	The Free On-line Dictionary of Computing
digital music	*Not available*	
data warehouse	(Or corporate data warehouse, CDW) Any system for storing, retrieving and managing large amounts of data. Data warehouse software often includes sophisticated compression and hashing techniques for fast searches, as well as advanced filtering. A data warehouse is often a relational database containing a recent snapshot of corporate data and optimised for searching. Planners and researchers can use this database without worrying about slowing down day-to-day operations of the production database. The latter can be optimised for transaction processing (inserts and updates).	The Free On-line Dictionary of Computing

electronic commerce	(EC) The conducting of business communication and transactions over networks and through computers. As most restrictively defined, electronic commerce is the buying and selling of goods and services, and the transfer of funds, through digital communications. However EC also includes all inter-company and intra-company functions (such as marketing, finance, manufacturing, selling, and negotiation) that enable commerce and use electronic mail, EDI, file transfer, fax, video conferencing, workflow, or interaction with a remote computer. Electronic commerce also includes buying and selling over the World-Wide Web and the Internet, electronic funds transfer, smart cards, digital cash (e.g. Mondex), and all other ways of doing business over digital networks.	The Free On-line Dictionary of Computing
enterprise resource planning	(ERP) Any software system designed to support and automate the business processes of medium and large businesses. This may include manufacturing, distribution, personnel, project management, payroll, and financials. ERP systems are accounting-oriented information systems for identifying and planning the enterprise-wide resources needed to take, make, distribute, and account for customer orders. ERP systems were originally extensions of MRP II systems, but have since widened their scope. An ERP system also differs from the typical MRP II system in technical requirements such as relational database, use of object oriented programming language, computer aided software engineering tools in development, client/server architecture, and open system portability.	The Free On-line Dictionary of Computing
global positioning system	(GPS) A system for determining postion on the Earth's surface by comparing radio signals from several satellites. When completed the system will consist of 24 satellites equipped with radio transmitters and atomic clocks. Depending on your geographic location, the GPS receiver samples data from up to six satellites, it then calculates the time taken for each satellite signal to reach the GPS receiver, and from the difference in time of reception, determines your location.	The Free On-line Dictionary of Computing

instant messaging	*IM* the online facility that allows the instant exchange of written messages between two or more people using different computers or mobile phones	Collins English Dictionary - Complete & Unabridged 10th Edition
inventory management	*Not available*	
linux	a nonproprietary computer operating system suitable for use on personal computers	Collins English Dictionary - Complete & Unabridged 10th Edition
mp3 player	a small portable digital audio player capable of storing MP3 files downloaded from the internet or transferred from a CD	Collins English Dictionary - Complete & Unabridged 10th Edition
online analytical processing	(OLAP) A category of database software which provides an interface such that users can transform or limit raw data according to user-defined or pre-defined functions, and quickly and interactively examine the results in various dimensions of the data. OLAP primarily involves aggregating large amounts of diverse data. OLAP can involve millions of data items with complex relationships. Its objective is to analyze these relationships and look for patterns, trends, and exceptions.	The Free On-line Dictionary of Computing
online advertising	*Not available*	
online sales	*Not available*	
open source software	A method and philosophy for software licensing and distribution designed to encourage use and improvement of software written by volunteers by ensuring that anyone can copy the source code and modify it freely.	The Free On-line Dictionary of Computing
public key infrastructure	(PKI) A system of public key encryption using digital certificates from Certificate Authorities and other registration authorities that verify and authenticate the validity of each party involved in an electronic	The Free On-line Dictionary of Computing

	transaction. PKIs are currently evolving and there is no single PKI nor even a single agreed-upon standard for setting up a PKI. However, nearly everyone agrees that reliable PKIs are necessary before electronic commerce can become widespread.	
quality control	a system for verifying and maintaining a desired level of quality in a product or process by careful planning, use of proper equipment, continued inspection, and corrective action as required. Origin: 1930–35	Dictionary.com Unabridged
robot	A machine designed to replace human beings in performing a variety of tasks, either on command or by being programmed in advance.	The American Heritage® Science Dictionary
rss technology	Really Simple Syndication: a way of allowing web users to receive syndicated newsletters and email alerts	Collins English Dictionary - Complete & Unabridged 10th Edition
supply chain management	*Not available*	
salesforce automation	(Sales Automation, SFA, SFFA, Sales & Field Force Automation) Software to support sales reps. The software gives sales representatives access to contacts, appointments and e-mail. It is likely to be integrated with Customer Relationship Management systems and Opportunity Management Systems.	The Free On-line Dictionary of Computing
six sigma	*trademark* a business management strategy that uses statistical methods to identify defects and improve performance	Collins English Dictionary - Complete & Unabridged 10th Edition
social networking	the use of a website to connect with people who share personal or professional interests, place of origin, education at a particular school, etc.	Dictionary.com's 21st Century Lexicon
service oriented architecture	(SOA) Systems built from loosely-coupled software modules deployed as services, typically communicating via a network. This allows different modules to be implemented and deployed in	The Free On-line Dictionary of Computing

	different ways, e.g. owned by different organisations, developed by different teams, written in different programming languages, running on different hardware and operating systems. The key to making it work is interoperability and standards so that modules can exchange data. SOAs often support service discovery, allowing a service to be changed without having to explicitly reconnect all its clients. Many different frameworks have been developed for SOA, including SOAP, REST, RPC, DCOM, CORBA, web services and WCF.	
virtualization	*Not available*	
virtual private network	(VPN) The use of encryption in the lower protocol layers to provide a secure connection through an otherwise insecure network, typically the Internet. VPNs are generally cheaper than real private networks using private lines but rely on having the same encryption system at both ends. The encryption may be performed by firewall software or possibly by routers.	The Free On-line Dictionary of Computing
web service	A family of standards promoted by the W3C for working with other business, developers and programs through open protocols, languages and APIs, including XML, Simple Object Access Protocol, WSDL and UDDI.	The Free On-line Dictionary of Computing
world wide web	The complete set of electronic documents stored on computers that are connected over the Internet and are made available by the protocol known as HTTP. The World Wide Web makes up a large part of the Internet.	The American Heritage® Science Dictionary

Appendix B: IRB Protocol Approval

UNIVERSITY OF MARYLAND
INSTITUTIONAL REVIEW BOARD

Initial Application Approval Notification

To: Principal Investigator, Ping Wang, College of Information Studies
Student, Chia-jung Tsui, College of Information Studies

From: James M. Hagberg
IRB Co-Chair
University of Maryland College Park

Re: IRB Protocol: 10-0391 - Scalable Computational Analysis of the Diffusion of Technological Concepts

Approval Date: July 16, 2010

Expiration Date: July 16, 2013

Application: Initial

Review Path: Exempt

The University of Maryland, College Park Institutional Review Board (IRB) Office approved your Initial IRB Application. This transaction was approved in accordance with the University's IRB policies and procedures and 45 CFR 46, the Federal Policy for the Protection of Human Subjects. Please reference the above-cited IRB Protocol number in any future communications with our office regarding this research.

Recruitment/Consent: For research requiring written informed consent, the IRB-approved and stamped informed consent document will be sent via mail. The IRB approval expiration date has been stamped on the informed consent document. Please note that research participants must sign a stamped version of the informed consent form and receive a copy.

Continuing Review: If you intend to continue to collect data from human subjects or to analyze private, identifiable data collected from human subjects, beyond the expiration date of this protocol, you must submit a Renewal Application to the IRB Office 45 days

prior to the expiration date. If IRB Approval of your protocol expires, all human subject research activities including enrollment of new subjects, data collection and analysis of identifiable, private information must cease until the Renewal Application is approved. If work on the human subject portion of your project is complete and you wish to close the protocol, please submit a Closure Report to irb@umd.edu.

Modifications: Any changes to the approved protocol must be approved by the IRB before the change is implemented, except when a change is necessary to eliminate an apparent immediate hazard to the subjects. If you would like to modify an approved protocol, please submit an Addendum request to the IRB Office.

Unanticipated Problems Involving Risks: You must promptly report any unanticipated problems involving risks to subjects or others to the IRB Manager at 301-405-0678 or jsmith@umresearch.umd.edu

Additional Information: Please contact the IRB Office at 301-405-4212 if you have any IRB-related questions or concerns. Email: irb@umd.edu

The UMCP IRB is organized and operated according to guidelines of the United States Office for Human Research Protections and the United States Code of Federal Regulations and operates under Federal Wide Assurance No. FWA00005856.

0101 Lee Building
College Park, MD 20742-5125
TEL 301.405.4212
FAX 301.314.1475
irb@umd.edu
http://www.umresearch.umd.edu/IRB

Appendix C: Survey Request Email

To: AISWorld mailing list <aisworld@lists.aisnet.org>
Subject: Help Needed for IT Classification Study

Dear Colleague,

At the University of Maryland, we are developing automatic methods for classifying IT concepts and we invite you to take a short online survey to help us evaluate our methods. Our goal is to improve our ability to classify IT concepts automatically. If you are familiar with a broad range of IT concepts, please help us by completing this short survey here:
http://www.surveymonkey.com/s/XSHHZWX

Your response will be kept confidential and only aggregated results will be reported. For more information about the study, please refer to the Consent Form at this URL:
http://terpconnect.umd.edu/~ctsui/ConsentForm.html

Thank you very much for your time and we hope you might be able to help us by participating in this study.

Sincerely,

Chia-jung Tsui
The PopIT Research Team
College of Information Studies
University of Maryland
http://terpconnect.umd.edu/~pwang/PopIT/

Appendix D: Survey

Page 1

Welcome to this survey to study IT classification!

You must be 18 years of age or older to participate and your participation is voluntary. Your response will be confidential. You will NOT be identified under any circumstance and only aggregated results will be reported. By clicking “Next”, you agree to participate in this study.

Page 2

Our goal of this research is to classify various information technology (IT) concepts into meaningful categories. Classification is the arrangement of objects into groups based on their relationships. The purpose of a classification is to simplify these relationships so that general statements can be made about groups of objects. Current manual classification methods are labor-intensive and time-consuming. We are trying to develop automatic ways to classify IT concepts. We invite you to help us evaluate the outcomes of our automatic classification.

Below are three different classifications (A, B, and C) of the same 35 IT concepts. Each classification has 14 groups. The 14 groups in each classification and the IT concepts within each group are in alphabetical order.

Please review and compare the three classifications in pairwise and answer the following questions.
(Note that in the following pairwise comparisons, groups with exactly the same IT concepts are removed from both classifications.)

The first comparison: A and B

Classification A
application service provider
artificial intelligence, robot, world wide web
business intelligence software, data mining, data warehouse, online analytical processing
customer relationship management, electronic commerce, enterprise resource planning, inventory management, salesforce automation, supply chain management
linux, open source software, virtualization
online advertising
online sales
public key infrastructure, virtual private network

Classification B
application service provider, enterprise resource planning
artificial intelligence, robot
business intelligence software, customer relationship management, data mining, salesforce automation
data warehouse, online analytical processing
electronic commerce, online advertising, online sales, public key infrastructure, world wide web
inventory management, supply chain management
linux, open source software
virtual private network, virtualization

1. In your opinion, which classification, A or B, is a better one? Please explain.

Page 3

The second comparison: B and C

Classification B
air traffic control, aviation, global positioning system
application service provider, enterprise resource planning
artificial intelligence, robot
blog, rss technology, social networking
business intelligence software, customer relationship management, data mining, salesforce automation
chat room, instant messaging
data warehouse, online analytical processing
electronic commerce, online advertising, online sales, public key infrastructure, world wide web
inventory management, supply chain management
service oriented architecture, web service
virtual private network, virtualization

Classification C
air traffic control, aviation
application service provider, customer relationship management, electronic commerce, enterprise resource planning, inventory management, salesforce automation, supply chain management, web service, world wide web
artificial intelligence, global positioning system, robot
blog, chat room, instant messaging, social networking
business intelligence software
data mining, data warehouse, online analytical processing
online advertising, online sales
public key infrastructure, virtual private network
rss technology
service oriented architecture
virtualization

2. In your opinion, which classification, B or C, is a better one? Please explain.

Page 4

The last comparison: A and C

Classification A
air traffic control, aviation, global positioning system
application service provider
artificial intelligence, robot, world wide web
blog, rss technology, social networking
business intelligence software, data mining, data warehouse, online analytical processing
chat room, instant messaging
customer relationship management, electronic commerce, enterprise resource planning, inventory management, salesforce automation, supply chain management
linux, open source software, virtualization
online advertising
online sales
public key infrastructure, virtual private network
service oriented architecture, web service

Classification C
air traffic control, aviation
application service provider, customer relationship management, electronic commerce, enterprise resource planning, inventory management, salesforce automation, supply chain management, web service, world wide web
artificial intelligence, global positioning system, robot
blog, chat room, instant messaging, social networking
business intelligence software
data mining, data warehouse, online analytical processing
linux, open source software
online advertising, online sales
public key infrastructure, virtual private network
rss technology
service oriented architecture
virtualization

3. In your opinion, which classification, A or C, is a better one? Please explain.

Page 5

In the space below, please share with us your thoughts on how IT concepts may be usefully classified or categorized. For example, what kinds of categories/classes would be useful or meaningful, for what purposes? And how to identify those categories/classes efficiently?

Page 6

Please let us know who you are to help us better understand your response. Although the questions below are optional, please be sure to click "Done" in the end of this page to submit the survey.

Name:

Email address:

Current Position:

Current Organization:

Highest Degree Obtained:

Area of Degree:

This is the end of the survey. Please click "Done" below to submit it.
Thank you very much for your participation! :-)

References

Barki, H., Rivard, S., & Talbot, J. (1993). A keyword classification scheme for IS research literature: An update. *MIS Quarterly, 17*(2), 209-226.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: Worlds, sentences, discourse. *Discourse Processes, 25*(2&3), 211-257.

Burgess, C., & Lund, K. (1997a). Parsing constraints and high-dimensional semantic space. *Language and Cognitive Processes, 12*, 177-210.

Burgess, C., & Lund, K. (1997b). *Representing abstract words and emotional connotation in high-dimensional memory space.* Paper presented at the 19th Annual Conference of the Cognitive Science Society, Mahwah, NJ.

Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In E. Dietrich & A. B. Markman (Eds.), *Cognitive dynamics: Conceptual and representational change in humans and machines* (pp. 117-156). Mahwah, NJ: L. Erlbaum Associates.

Carroll, J. D., Arabie, P., Chaturvedi, A., & Hubert, L. (2004). Multidimensional scaling and clustering in marketing: Paul green's role. In P. E. Green & Y. Wind (Eds.), *Marketing research and modeling: Progress and prospects: A tribute to paul e. Green* (pp. 71-100). Boston: Kluwer Academic Publishers.

Conover, W. J. (1980). *Practical nonparametric statistics*. New York: Wiley.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.

Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management, 37*(6), 817-842.

Ein-Dor, P., & Segev, E. (1993). A classification of information systems: Analysis and interpretation. *Information Systems Research, 4*(2), 166-204.

Everitt, B., Landau, S., & Leese, M. (2001). *Cluster analysis* (four ed.). New York: Oxford University Press.

Fichman, R. G. (2004). Going beyond the dominant paradigm for information technology innovation research: Emerging concepts and methods. *Journal of the Association for Information Systems, 5*(8), 314-355.

Hansen, P., & Jaumard, B. (1997). Cluster analysis and mathematical programming. *Mathematical programming, 79*(1-3), 191-215.

Harris, K., Caldwell, F., Linden, A., Knox, R., & Logan, D. (2003). *Taxonomy creation: Bringing order to complexity* (No. QA-20-8719): Gartner, Inc.

Jagerman, E. J. (2006). *Creating, maintaining and applying quality taxonomies*. Zoetermeer: Lulu.com.

Kojadinovic, I. (2004). Agglomerative hierarchical clustering of continuous variables based on mutual information. *Computational Statistics & Data Analysis, 46*(2), 269-294.

Kraskov, A., Stogbauer, H., Andrzejak, R. G., & Grassberger, P. (2005). Hierarchical clustering using mutual information. *Europhysics Letters, 70*(2), 278-284.

Kruskal, J. B. (1977). The relationship between multidimensional scaling and clustering. In J. Van Ryzin (Ed.), *Classification and clustering* (pp. 17-44). New York: Academic Press.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79-86.

Lambe, P. (2007). *Organising knowledge: Taxonomies, knowledge and organisational effectiveness*. Oxford: Chandos.

Logan, D. (2009). *Best practices for taxonomy creation* (No. G00167683): Gartner, Inc.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers, 28*(2), 203-208.

Lund, K., Burgess, C., & Atchley, R. A. (1995). *Semantic and associative priming in high-dimensional semantic space.* Paper presented at the Seventeenth Annual Conference of the Cognitive Science Society, Hillsdale, NJ.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Mayr, E. (1942). *Systematics and the origin of species, from the viewpoint of a zoologist*. New York: Columbia Univ. Press.

McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science, 41*(6), 433-443.

Napior, D. (1972). Nonmetric multidimensional techniques for summated ratings. In R. N. Shepard, A. K. Romney & S. B. Nerlove (Eds.), *Multidimensional scaling:*

Theory and applications in the behavioral sciences (pp. 157-178). New York: Seminar Press.

Schvaneveldt, R. W. (1990). *Pathfinder associative networks: Studies in knowledge organizations*. Norwood, NJ: Ablex Pub. Corp.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review, 86*(2), 87-123.

Shepard, R. N., Romney, A. K., & Nerlove, S. B. (Eds.). (1972). *Multidimensional scaling: Theory and applications in the behavioral sciences*. New York: Seminar Press.

Sireci, S. G., & Geisinger, K. F. (1992). Analyzing test content using cluster analysis and multidimensional scaling. *Applied Psychological Measurement, 16*(1), 17-31.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics, 19*, 143-177.

Sokal, R. R. (1974). Classification: Purposes, principles, progress, prospects. *Science, 185*(4157), 1115-1123.

Spence, D. P., & Owens, K. C. (1990). Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research, 19*(5), 317-330.

Strang, D., & Soule, S. A. (1998). Diffusion in organizations and social movements: From hybrid corn to poison pills. *Annual Review of Sociology, 24*, 265-290.

Takane, Y., Young, F. W., & de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika, 42*(1), 7-67.

Tsui, C.-j., Wang, P., Fleischmann, K. R., Oard, D. W., & Sayeed, A. B. (2009). *Understanding IT innovations through computational analysis of discourse.* Paper presented at the 30th International Conference on Information Systems (ICIS), Phoenix, AZ.

van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). London: Butterworths.

Wang, P. (2009). Popular concepts beyond organizations: Exploring new dimensions of information technology innovations. *Journal of the Association for Information Systems, 10*(1), 1-30.

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science, 49*(4), 327-355.