Volume 12, Numbers 1 and 2 ISSN 1524-7252

ACADEMY OF INFORMATION AND MANAGEMENT SCIENCES JOURNAL

Sharad K. Maheshwari
Editor
Hampton University

The *Academy of Information and Management Sciences Journal* is owned and published by the DreamCatchers Group, LLC, and printed by Whitney Press, Inc. Editorial content is under the control of the Allied Academies, Inc., a non-profit association of scholars, whose purpose is to support and encourage research and the sharing and exchange of ideas and insights throughout the world.

Whitney Press, Inc.

Printed by Whitney Press, Inc.
PO Box 1064, Cullowhee, NC 28723
www.whitneypress.com

Authors provide the Academy with a publication permission agreement. Neither the Academy of Information and Management Sciences, the Allied Academies, the DreamCatchers Group nor Whitney Press is responsible for the content of the individual manuscripts. Any omissions or errors are the sole responsibility of the individual authors. The Editorial Board is responsible for the selection of manuscripts for publication from among those submitted for consideration. The Editors accept final manuscripts in digital form and the Publishers make adjustments solely for the purposes of pagination and organization.

The *Academy of Information and Management Sciences Journal* is owned and published by the DreamCatchers Group, LLC, PO Box 2689, Cullowhee, NC 28723, USA, (828) 293-9151. Those interested in subscribing to the *Journal*, advertising in the *Journal*, submitting manuscripts to the *Journal,* or otherwise communicating with the *Journal*, should contact the Allied Academies' Executive Director at info@.alliedacademies.org.

ACADEMY OF INFORMATION AND MANAGEMENT SCIENCES JOURNAL EDITORIAL BOARD MEMBERS

ACADEMY OF INFORMATION AND MANAGEMENT SCIENCES JOURNAL

CONTENTS

ACADEMY OF INFORMATION AND MANAGEMENT SCIENCES JOURNAL

CONTENTS

LETTER FROM THE EDITOR

Welcome to the *Academy of Information and Management Sciences Journal,* the official journal of the Academy of Information and Management Sciences. The Academy is one of several academies which collectively comprise the Allied Academies. Allied Academies, Incorporated is a non-profit association of scholars whose purpose is to encourage and support the advancement and exchange of knowledge.

The editorial mission of the *AIMSJ* is to publish empirical and theoretical manuscripts which advance the disciplines of Information Systems and Management Science. All manuscripts are double blind refereed with an acceptance rate of approximately 25%. Manuscripts which address the academic, the practitioner, or the educator within our disciplines are welcome. And, diversity of thought will always be welcome.

Please visit the Allied Academies website at www.alliedacademies.org for continuing information about the Academy and the *Journal*, and about all the Allied Academies and our collective conferences. Submission instructions and guidelines for publication are also provided on the website.

Sharad K. Maheshwari
Hampton University

www.alliedacademies.org

This is a combined edition containing both Volume 12, Number 1, and Volume 12, Number 2

Articles for Volume 12, Number 1

IT WORKERS ON OUTSOURCING: WHAT ABOUT ME? WHAT ABOUT THE PROFESSION?

Nita G. Brooks, Middle Tennessee State University
Robert E. Miller, Ashland University
Melinda L. Korzaan, Middle Tennessee State University

ABSTRACT

To date, there have been numerous studies examining the impact of outsourcing on the IT function within and across organizations. Practitioners and researchers have focused tremendous efforts on understanding how to enter into and maintain relationships with outsourcing vendors. There have been, however, very few studies focusing on the impact outsourcing is having on the IT worker. In order to understand how outsourcing is impacting the IT workforce, this paper empirically examines the attitudes of IT workers toward IT outsourcing. Perceptions of how individuals perceive the impact of outsourcing on them personally and on the profession are explored along with whether these individuals perceive the impact as having already occurred or will occur. Findings from this study of over 450 IT workers from various organizations in the U.S. indicate that individuals do have different perceptions of how outsourcing impacts them versus how outsourcing impacts the IT profession. Analysis of how these perceptions impact outcomes shown to be important to IT workers was also conducted. Key outcomes including satisfaction and turnover intention were found to be significantly and negatively related to IT workers' perceptions of outsourcing's impact. Implications of these findings are provided along with direction for future research.

INTRODUCTION

IT outsourcing is not by any means a new phenomenon. Outsourcing of IT functions can be traced back to the early 1960's when EDS took over the data processing functions of two large companies: Frito-Lay and Blue Cross (Lacity & Hirchheim, 1993). Loh and Venkatramen (1992) coined the phrase "Kodak effect" to refer to the impact of Kodak's arrangement to turn over the majority of their IT operations to IBM, Businessland, and DEC on IT outsourcing in general. Since that time, IT outsourcing has grown and is considered a viable and strategic option for organizations. Organizations use different sourcing strategies ranging from total outsourcing to selective outsourcing. Firms continue to outsource "an increasingly large range of and depth of services"

(Barthelemy & Geyer, 2004, p. 91), this proliferation of IT outsourcing is not anticipated to slowdown in the near future (Computer Economics, 2006).

A current trend is also emerging in which medium and small-sized companies are jumping on the IT outsourcing bandwagon. While large organizations have historically outsourced their IT functions, today small firms are progressively farming out IT operations to companies that provide a range of options for the smaller scale business. “Overall, small and medium-sized businesses are forecast to make up two-thirds of the outsourced help-desk market in 2011” (Tam, 2007, p. B.5). Therefore, the effects of outsourcing are touching an increasingly wider range of companies and industries.

Economic indicators and uncertainty have further perpetuated the outsourcing trend (Business Wire, 2008). Many companies have come to believe that outsourcing is good for the bottom-line. Chrysler, a company which has struggled for years, has taken this idea to heart by incorporating additional IT outsourcing as part of its plan for recovery (Overby, 2008). While many companies are outsourcing to domestic firms, a growing number have decided specifically to outsource offshore. Companies are driven to offshore outsourcing because they have become convinced it is necessary in order to be competitive in the global economy (HR Focus, 2008). The demand for offshore outsourcing is so great that IBM “announced plans to invest $6 billion in India to expand its workforce there” (Songini, 2007, p. 1).

Given the significance of outsourcing, it is not surprising that the phenomenon has been widely investigated by IT researchers. Studies have examined why organizations outsource and the motivations behind outsourcing arrangements (Lacity & Hirschheim, 1995); the relationships between client firms and vendors (Kim & Chung, 2003); and how the technology acceptance model can be used to examine intentions to outsource (Benamati & Fajkumar, 2002). Within this body of research, there has, however, been a noticeable lack of attention given to the IT worker. This is somewhat surprising considering the need for organizations to recruit and retain talented and creative IT workers.

As indicated by recent research, organizations of all sizes intend to increase their IT workforce by the year 2008 (Zweig, 2006). At the same time organizations are adding to the IT workforce pool, as just discussed, they are increasing their use of outsourcing as a means to reduce costs. While these two ideas seem to be in conflict, they are not. In a recent IT workforce trends study, median firms indicated that they expect to source approximately 17% of their IT workforce in *certain positions* and to increase their IT workforce in other areas (Zweig, 2006). While not all positions in the IT profession are moving to outsourcing vendors off-shore, it is important to note the impact these changes in strategy can have on the profession as a whole. For example, educational programs in information technology and information systems have experienced lower and lower enrollments. It could be that a misperception about the impact of outsourcing on the profession is a cause (Panko, 2008).

As stated by Barthelemy (2003), one of the “deadliest sins” of management when engaging in outsourcing ventures is to overlook workforce issues. It is imperative that organizations recognize and acknowledge the IT worker when engaging in sourcing arrangements. Only one study was found that examined IT worker perceptions of outsourcing and how those perceptions impacted outcomes (Kennedy, Holt, Ward, & Rheg, 2002).

In order to more thoroughly understand how IT workers perceive outsourcing and how those perceptions impact outcomes, a research framework was developed. This framework was then used as a guide for a field study targeting IT workers from various organizations and industries. A complete discussion of the framework, field study, and study results are provided. Implications for managers and organizations are discussed as well along with directions for future research.

RESEARCH FRAMEWORK

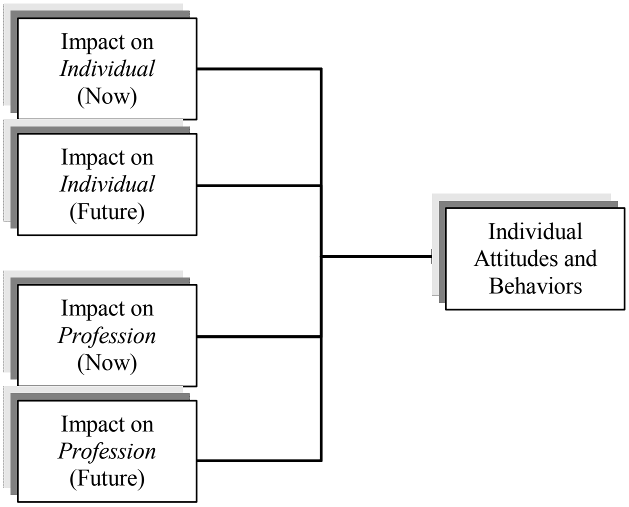
In order to more accurately and thoroughly understand how IT workers perceive the impact of outsourcing, it was first necessary to develop a research framework to serve as a guide. As stated previously, only one study was found that empirically examined how IT workers perceived outsourcing (Kennedy, et al., 2002). The sample for this study was taken from Air Force engineering managers and focused only on how the individuals in the study perceived outsourcing would, *in the future*, impact their job functions.

There are actually many ways in which outsourcing can impact individuals in relation to recruitment and retention. For example, if an individual perceives that outsourcing *will* negatively impact their job or place in the profession, it would be more likely that the individual would seek out other career or job opportunities. If the individual perceives that outsourcing currently *has* impacted their job, the outcomes could be different. It is possible that both perceptions of outsourcing's impact could inherently influence the individual's perception that the profession is breaching a perceived or psychological contract. This could prove to be very significant as it relates to key outcomes such as recruitment and retention.

Another perspective taken in this study relates to how individuals view outsourcing's impact on the profession as a whole and not solely on them personally. For example, if an individual feels that outsourcing will negatively impact the IT profession, they would be less likely to join and more likely to leave the profession. The temporal aspect is also important here. If the individual feels that outsourcing *has* negatively impacted the profession, would they be more inclined to leave? These different perceptions were taken into account in creating the research framework presented and used here.

The framework, provided in Figure 1 as mentioned, was developed to serve as a guide to examining individual attitudes towards IT outsourcing. The framework is unique in that it considers two main factors: 1) the target of the perceptions (individual or IT profession) and 2) the temporal component (present and future). By focusing on these specific areas, it may be possible to better understand how the individual's attitudes could impact factors such as entry into the profession, job performance, career decisions, satisfaction and commitment to the profession, and intention to turnover from the profession.

Figure 1: Proposed Framework of Worker Perceptions of Outsourcing


Impact on *Individual* (Now)
Impact on *Individual* (Future)
Impact on *Profession* (Now)
Impact on *Profession* (Future)
Individual Attitudes and Behaviors

In order to explore the importance of the perceptions highlighted in Figure 1, it is necessary to relate them to outcomes. Within the body of research examining the IT workforce, there have been several key outcomes noted as important to retention. Two primary outcomes include satisfaction and intention to turnover. The majority of work on the IT worker has focused on attitudes and behaviors within an organizational setting. Within this body of work, job satisfaction and career satisfaction have been shown to impact organizational commitment and intention to turnover from the organization (Igbaria & Greenhaus, 1992; Gupta, Guimaraes, & Raghunathan, 1992; and Igbaria & Siegel, 1993), and intention to turnover from the organization has been shown to impact actual turnover (Thatcher, Stepina, & Boyle, 2003).

This study provides a broader examination of IT worker perceptions that extend beyond the current organization. Perceptions are studied related to how outsourcing has/will personally impact the individual and how outsourcing has/will impact the profession. Outcomes studied relate to the individual's place in the profession and include overall career satisfaction, general satisfaction with the profession, and intention to turnover from the profession. The research method and measures used in this study are discussed in the following section. Results of the analysis are also provided.

RESEARCH METHOD

Individuals currently working in the IT profession were targeted from several large organizations in the U.S. Each person was sent an email explaining the study and requesting his or her participation. A link to a web survey was provided in the email. All individuals were assured anonymity. Eight-hundred and twelve (812) individuals were contacted requesting their participation in this study. Usable responses were received from 454 individuals representing an overall response rate of 56%.

Demographics for the sample are provided in Table 1. The sample consisted of 61.7% male respondents. The average age of the participants was 39.7. Despite efforts to receive adequate representation from different minority groups, approximately 87.4% of sample participants were white. The majority of the participants worked full time (93%); had an undergraduate degree or higher (71%); were married or living with a partner (71.8%); and represented a variety of IT positions.

The average number of years spent in the IT profession for this sample was 14, and the average number of organizations worked for as an IT professional was 2.85. Thirty-eight different companies are represented in the sample. The majority of respondents (approximately 90%) were from seven organizations providing representation from state government, healthcare, information systems, transportation, and energy industries.

In order to measure how individual IT workers perceive the impact of outsourcing on them and on the profession (now and in the future), items were developed to reflect each component of the framework. The instrument provided by Kennedy, et al. (2002) was used a guide for creating the new items. The Kennedy, et al. (2002) items had to be modified because the study focused solely on how outsourcing would impact the individual's *job* in the *future* and did not consider the impact outsourcing has had on the individual or the individual's place in the profession.

Measures for outcome variables (general satisfaction with the profession, overall career satisfaction, and intention to turnover from the profession) were based on existing instruments. General satisfaction with the profession was measured using two items provided by Hackman and Oldham (1976). Career satisfaction was measured using five items from Greenhaus, Parasuraman, and Wormley (1990), and intention to turnover from the profession was measured with four items from Meyer, Allen, and Smith (1993).

In order to determine the validity of the perception factors presented in the framework in Figure 1, a factor analysis was conducted using orthogonal rotation. Items and factor loadings are provided in Table 2. After reviewing the factor matrix, it was apparent that there was no distinction made by individuals in this sample related to the temporal aspect provided in the research framework. Individuals did not differentiate between outsourcing's impact on them *now* versus its impact on them in the *future*.

Table 1: Respondent Demographics	
Gender	
Male	61.7%
Female	37.7%
Missing	0.7%
Ethnicity	
White	87.4%
African-American	6.2%
Hispanic	1.5%
Other	4.2%
Missing	0.7%
Education	
Some College	10.3%
Associates Degree	8.1%
Undergraduate Degree	57.2%
Masters Degree	14.0%
Other	1.8%
Missing	8.6%
Marital Status	
Never Married	15.0%
Married/Living with Partner	71.8%
Separated/Divorced	9.3%
Widowed	0.9%
Missing	3.1%
Age	
20-29	17.2%
30-39	31.3%
40-49	31.5%
50-59	13.6%
60+	2.6%
Missing	3.7%

Table 2: Factor Loadings for Outsourcing Perceptions

	Factor Loadings	
	1	2
Impact on Individual		
Outsourcing …		
has negatively influenced **my** IT career. **(out1)**	**.807**	.258
will, in the future, negatively influence **my** IT career. **(out2)**	**.815**	.389
has negatively impacted **my** mobility in the IT profession. **(out3)**	**.849**	.278
will, in the future, negatively impact **my** mobility in the IT profession. **(out4)**	**.830**	.372
has caused **my** IT career to become less secure. **(out5)**	**.772**	*.444*
will, in the future, cause **my** IT career to become less secure. **(out6)**	**.757**	*.492*
has greatly reduced **my** opportunities for advancement in the IT profession. **(out7)**	**.830**	.303
will, in the future, greatly reduce **my** opportunities for advancement in the IT profession. **(out8)**	**.794**	.395
Impact on Profession		
Outsourcing …		
has caused jobs across the IT profession to become less secure. **(out9)**	.357	**.833**
will, in the future, cause jobs across the IT profession to become less secure. **(out10).**	.336	**.854**
has negatively impacted the mobility of individuals across the IT profession. **(out11)**	.344	**.864**
will, in the future, negatively impact the mobility of individuals across the IT profession. **(out12)**	.335	**.875**
has negatively influenced career advancement opportunities for individual in the IT profession. **(out13)**	.386	**.828**
will, in the future, negatively influence career advancement opportunities for individuals in the IT profession. **(out14)**	.384	**.840**

Factors did emerge for constructs related to the individual and the profession as expected. Slight cross loadings were revealed between two items (out5 and out6). These items measure career security. Individuals in the sample may have seen career security as transcending the individual/profession line if they perceived career security as being tied to their current place in IT. However, since both items are below the .50 threshold recommended for practical significance (Hair, et al., 2006) it was determined to leave them for this analysis.

RESEARCH RESULTS

The factor analysis provided in the previous section indicates that individuals do perceive a difference in how outsourcing impacts them versus how it impacts the profession. Descriptive statistics, correlations, and reliabilities are provided in Table 3 for the two factor solution and the outcomes included in the study. Reliabilities for all constructs were well above the recommended .70 cutoff (Hair, et al., 2006).

Table 3: Descriptive Statistics, Reliabilities, and Correlations among Study Variables

Measure	M (SD)	1	2	3	4	5
1. Outsourcing (Impact on Individual)	3.9559 (1.57)	**(.96)**				
2. Outsourcing (Impact on Profession)	4.9284 (1.43)	.733**	**(.97)**			
3. Career Satisfaction	4.7605 (1.25)	-.209**	-.112*	**(.90)**		
4. General Satisfaction	5.2467 (1.18)	-.167**	-.047	.574**	**(.91)**	
5. Turnover Intention	3.7124 (.921)	.184**	.171**	-.338**	-.499**	**(.88)**

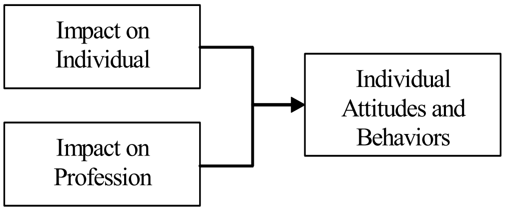
**=p<.01

Findings from the correlation matrix reveal that perceptions of outsourcing are significantly related to key outcomes. Perceptions of outsourcing's impact on the individual were found to be significantly and negatively related to career satisfaction and general satisfaction with the profession, and positively related to intention to turnover from the profession. Perceptions of outsourcing's impact on the profession as a whole were found to be negatively related to career satisfaction and positively related to intention to turnover from the profession. Strong and positive relationships exist, as would be expected, between how individuals perceive the impact of outsourcing on themselves and the impact of outsourcing on the profession.

Although the two perception measures may be correlated, a comparison of the means indicates that, in general, individuals perceived outsourcing's impact on the profession be to more negative than its impact on them personally. The implications of this finding, as well as those presented previously will be discussed in the following section.

Additional analysis was conducted to determine if the perceptions differed based on key demographic factors: gender and age. For this sample, there were no significant differences found in the perceptions of outsourcing's impact on the individual or the profession when comparing men and women. These results are consistent with previous findings indicating that men and women tend to have similar perceptions of career experiences (Sumner & Niederman, 2003/2004). For this sample, men tended to have slightly more negative perceptions of outsourcing's impact overall. To examine if there were any differences based on age, a one-way ANOVA was used. As with gender, there were no significant differences indicated among the various age groups examined.

Figure 2: Final Model of Worker Perceptions of Outsourcing



IMPLICATIONS

Findings from this study revealed two important issues: 1) individuals in the IT profession perceive outsourcing's impact on the profession differently from its impact on them personally and 2) the IT worker's perceptions of outsourcing's impact have implications for outcomes such as satisfaction and turnover intention. This study shows that IT professionals are aware of outsourcing's impact on themselves and on their profession. The results indicate that, both personally and professionally, outsourcing was perceived to have a negative impact.

Regardless of which positions are actually being outsourced and which are in higher demand locally, individuals in the IT profession view outsourcing as having a negative impact. Also, the fact that IT professionals actually differentiate in their perceptions of outsourcing's impact on themselves as compared to the profession is quite interesting. Instead of portending a future of doom and gloom it appears that IT professionals can acknowledge the negative impact of outsourcing on the profession without necessarily internalizing the fact. Based on this rationalization, IT professionals may not be "driven off" from the profession as some have suggested. The ability to differentiate may also mean that IT recruiting is not a "lost cause". It does, however, provide some evidence that recruitment and retention efforts should work to attenuate these negative feelings.

That said, the study's findings are hardly all positive. In fact, it is clear from the results that perceptions of outsourcing's impact are significantly related to negative workforce outcomes. These findings support Barthelemy's (2003) assertion that management must consider workforce issues when engaging in outsourcing. An organization's plan to start outsourcing, or to expand current outsourcing levels, can lead to serious workforce issues. If an organization's foray into outsourcing is to be successful, management must give all due consideration to the organization's existing workforce. Failure to address current employee issues can lead to a drop in worker satisfaction and ultimately higher turnover.

Along with the previous findings, it is interesting to note that individuals did not perceive a difference in outsourcing's impact in relation to time. There were no distinctions made by IT professionals in terms of outsourcing's current versus future impact. This appears to indicate that IT professionals do not perceive outsourcing's impact will increase with time. Said another way, IT professionals may believe outsourcing has already had its major impact. Individuals may view it as one of many possible strategies for management to use to meet organizational needs.

CONCLUSION

While the study presented herein has examined a number of issues related to IT professionals and their perceptions of outsourcing, other issues remain to be addressed. As an example, the ability of IT professionals to differentiate between outsourcing's impact on the profession and its impact on them personally should be investigated further. It would be interesting to examine the influence of personality traits and how they might relate to the ability to differentiate impacts. In particular, researchers could investigate the role of locus of control in the perception of impacts on the individual versus the profession.

Research shows that the use of IT outsourcing shows no sign of stopping. Whether these outsourced IT jobs go to domestic or offshore firms, the impact on the IT workforce will continue to be significant. As such, furthering our understanding of outsourcing's impact on current and potential IT workers is a must.

REFERENCES

Barthelemy, J. (2003). The Seven Deadly Sins of Outsourcing. *Academy of Management Executive, 17*(2), 87-98.

Barthelemy, J., & Geyer, D. (2004). the Determinants of Total IT Outsourcing: An Empirical Investigation of French and German Firms. *Journal of Computer Information Systems, 44*(3), 91-97.

Benamati, J., & Fajkumar, T. (2002). The Application Development Outsourcing Decision: An Application of the Technology Acceptance Model. *Journal of Computer Information Systems, 42*(4), 35-43.

Greenhaus, J., Parasuraman, S., & Wormley, W. (1990). Effects of Race on Organizational Experiences, Job Performance Evaluations, and Career Outcomes. *Academy of Management Journal, 33*(1), 64-86.

Growth of IT Outsourcing: No End in Sight. (2008, August). Retrieved July 14, 2008, from Computer Economics: http://www.computereconomics.com/article.cfm?id=1161

Gupta, Y., Guimaraes, T., & Raghunathan, T. (1992). Attitudes and Intentions of Information Center Personnel. *Information and Management, 22* (3), 151-160.

Hackman, J., & Oldham, G. (1976). Motivation through the Design of Work: Test of a Theory. *Organizational Behavior and Human Performance, 16*, 250-279.

Hair, J., Black, W., Babin, B., Anderson, R., & Tatham, R. (2006). *Multivariate Data Analysis.* Upper Saddle River, New Jersey: Prentice Hall.

Igbaria, M., & Greenhaus, J. (1992). Determinants of MIS Employees' Turnover Intentions. *Communications of the ACM, 35* (2), 35-49.

Igbaria, M., & Siegel, S. (1993). The Career Decision of IS People. *Information and Management, 24*(1), 23-32.

Kennedy, J., Holt, D., Ward, M., & Rheg, M. (2002). The Influence of Outsourcing on Job Satisfaction and Turnover Intentions of Technical Managers. *Human Resource Planning, 24*(1), 23-31.

Kim, S., & Chung, Y. (2003). Critical Success Factors for IS Outsourcing Implementation from an Interorganizational Relationship Perspective. *Journal of Computer Information Systems, 43*(4), 81-90.

Lacity, M., & Hirschheim, R. (1995). *Information Systems Outsourcing Bandwagon: The Insourcing Response.* Chichester, England: John Wiley.

Lacity, M., & Hirschheim, R. (1993). *Information Systems Outsourcing: Myths, Metaphors, and Realities.* Chichester, England: John Wiley and Sons, Ltd.

Loh, L., & Venkatraman, N. (1992). Diffusion of Information Technology Outsourcing: Influence Sources and the Kodak Effect. *Information Systems Research, 3*(4), 334-358.

Meyer, J., Allen, N., & Smith, C. (1993). Commitment to Organizations and Occupations: Extension and Test of a Three Component Conceptualization. *Journal of Applied Psychology, 78*, 538-551.

Overby, S. (2008, April 9). *Chrysler Stakes Turnaround on IT Outsourcing*. Retrieved July 14, 2008, from CIO: http://advice.cio.com/stephanie_overby/chrysler_stakes_turnaround_on_it_outsourcing

Panko, R.R. (2008). IT Employment Prospects: Beyond the Dotcom Bubble. *MIS Quarterly Executive, 17*, 182-197.

Songini, M. L. (1007, April 2). *Circuit City Awards $775M IT Outsourcing Contract to IBM*. Retrieved July 14, 2008, from Computerworld: http://www.computerworld.com/action/article.do?command=viewArticleBase&articleid=9015341

Sumner, M., & Niederman, F. (2003/2004). The Impact of Gender Differences on Job Satisfaction, Job Turnover, and Career Experiences of Information Systems Professionals. *Journal of Computer Information Systems, 44*(2), 29-39.

Tam, P. (2007, April 17). Business Technology: Outsourcing Finds New Niche; More Small Firms Farm out Tech Work to Tap Experts, Pare Costs. *Wall Street Journal* , p. B5.

Thatcher, J., Stepina, L., & Boyle, R. (2003). Turnover of IT Workers: Examining Empirically the Influence of Attitudes, Job Characteristics, and External Markets. *Journal of Management Information systems, 19*(3), 231-261.

Will More Offshoring Be a Result of Economic Uncertainties. (2008). *HR Focus, 85*(6), 9.

Worldwide Outsourcing Industry Rebounding, According to EquaTerra's 4Q07 Pulse Survey. (2008, January 18). Retrieved July 14, 2008, from Business Wire: http://www.reuters.com/article/pressRelease/idUS155349+18-Jan-2008+BW20080118

THE IMPACT OF STOPPING RULES ON HIERARCHICAL CAPACITATED CLUSTERING IN LOCATION ROUTING PROBLEMS

Marco Lam, York College of Pennsylvania
John Mittenthal, The University of Alabama
Brian Gray, The University of Alabama

ABSTRACT

The objective of the multiple depot location routing problems (MDLRP) is to minimize warehousing and transportation costs by selecting warehouses out of a set of possible warehouses and to assign customers to routes serviced from these warehouses. The customer locations are fixed and customer demand is known. The constraints are that the demand for each customer must be met and facility and vehicle capacities cannot be exceeded. Each customer has to be placed on a single route serviced by a vehicle from a warehouse. Multiple routes may originate from each warehouse.

Because the MDLRP has been shown to be NP-hard, the extant literature has developed various heuristics, including clustering based heuristics, to solve the MDLRP. The underlying assumption of the clustering based heuristic approach is that clustering customers based on proximity is a reasonable approach to minimizing routing costs. Because hierarchical clustering does not require a priori assumptions about the number of clusters, we propose two stopping criteria: minimum number of clusters required and change in within cluster variation. Our results indicate that significant savings can be achieved by considering multiple stopping rules.

INTRODUCTION

The objective of clustering in a location routing setting is to identify the underlying structure (i.e., distribution of customers over the area). Mojena (1977) provides evidence that hierarchical clustering has been used successfully for this purpose. The underlying assumption of this approach is that clustering customers based on proximity is a reasonable approach to minimizing routing costs.

In general, clustering methods are used when a multi-dimensional space with a relatively high-density of points is separated from other regions with high densities of points by regions with relatively low densities of points (Jain & Dubes, 1988). Because customer locations often fit this requirement (e.g., Daskin, 1995), clustering approaches have been used in solving vehicle routing problems and location routing problems (e.g., Min, 1987; Barreto, Ferreira, Paixão, & Sousa Santos, 2007).

In practice, clusters might have to satisfy certain constraints. For instance, in a vehicle routing context, one might want to consider clustering customers based on proximity while the total demand of the customers in a cluster cannot exceed the vehicle or warehouse capacity, i.e., capacitated clustering. The vehicle capacity constraint could be weight, volume or both. Alternatively, when minimizing the route length, a constraint could be placed on the number of customers in a cluster.

Agglomerative hierarchical clustering techniques start with as many groups as observations. At each step, the groups that are most similar are merged. In a location routing context, similarity equates to distance; customers that are located closer to one another are geographically more similar. At each step, more distant observations are merged and thus the similarity measure decreases. Eventually, only one cluster remains with all observations.

For hierarchical cluster analysis no assumptions about the number of clusters or the underlying structure is required (Johnson & Wichern, 2002). Therefore, various stopping rules have been proposed in the prior literature to determine the number of clusters present (e.g., Milligan & Cooper, 1985). We contribute to this literature in the following ways. First, we extend hierarchical clustering approaches to include multiple side constraints. Second, we propose and analyze the performance of three stopping rules for capacitated hierarchical clustering. Third, we apply the stopping rules in a location routing setting.

The remainder of the chapter is organized as follows. In Section 2, we describe clustering and hierarchical clustering methods. In Section 3, we provide an overview of the clustering and capacitated clustering literature. In Section 4, we discuss how hierarchical clustering approaches can be modified to allow for clustering with constraints and discuss possible stopping criteria. We also provide an example. In Section 5, we provide an application of our approach in a vehicle routing context. Finally, in Section 6 we provide our conclusions.

CLUSTER ANALYSIS

The objective of cluster analysis is to partition data into subsets such that observations in a subset are similar while observations in different subsets are dissimilar. Each observation has to be assigned to exactly one cluster. To determine which observations to cluster, clustering requires a "closeness" or "similarity" measure. At each stage, observations that are similar or close are clustered. A distance metric is often used as a measure of similarity (Johnson & Wichern, 2002). In a location routing context, Euclidean distances have been used as the closeness or similarity measure (see e.g., Min, 1996; Barreto et al., 2007).

Agglomerative hierarchical clustering techniques sequentially merge subgroups into groups. The process starts with one subgroup for each observation. At each stage, the subgroups that have the highest level of similarity, or smallest distance, are clustered. As the subgroups are clustered into fewer remaining clusters, the similarity within the clusters or subsets decreases. Commonly used hierarchical clustering techniques are single linkage, complete linkage, average linkage, and Ward's method (e.g., Barreto et al., 2007). Single linkage clusters the subgroups based on the smallest distance where distance is defined as the distance of the nearest pair of items, one from each subgroup. Complete linkage clusters based on the smallest distance where distance is defined as the distance of the most distant pair of items, one from each subgroup. Average linkage also merges subgroups based on smallest distance but defines distance as the average distance of all pairs of items consisting of one item of each subgroup. Ward's method merges subgroups that minimize the increase in within cluster variation. An extensive description of these heuristics is provided in Johnson and Wichern (2002).

CAPACITATED CLUSTERING LITERATURE REVIEW

Hierarchical clustering approaches can be extended to include constraints. A special case of clustering with side constraints is the capacitated clustering problem (CCP). As in regular clustering, i.e., uncapacitated clustering, the problem is to partition a set of observations into a set of clusters such that each observation is assigned to exactly one cluster. The CPP has the additional requirement that the "size" of each cluster may not exceed a given capacity. For instance, one might want to cluster customers based on proximity. A constraint could be that the total of customers' demand within a cluster cannot exceed vehicle capacity or warehouse capacity.

Min (1996) provides a case in point. In this research, customers are clustered based on proximity. The clusters used are those that satisfy the vehicle capacity constraint. The main difference between the approach we propose and the approach Min (1996) used is the role of vehicle capacity as a stopping criterion for the clustering approach. Specifically, when at a given stage the best agglomeration of clusters violates the vehicle capacity, Min (1996) stops the agglomeration of clusters. In other words, they use the traditional hierarchical clustering approach and 'cut' the tree diagram at the level of the vehicle capacity. On the other hand, we use vehicle capacity as a constraint that limits the creation of certain clusters. When the best possible agglomeration is infeasible because of the capacity constraint, the next best feasible agglomeration, if one exists, is chosen. The conceptual difference between using vehicle capacity in capacitated clustering and using vehicle capacity as a stopping value is explained and demonstrated in the example and shown in the tree-diagram below (Figure 1).

Figure 1. Tree diagram for constraint clustering

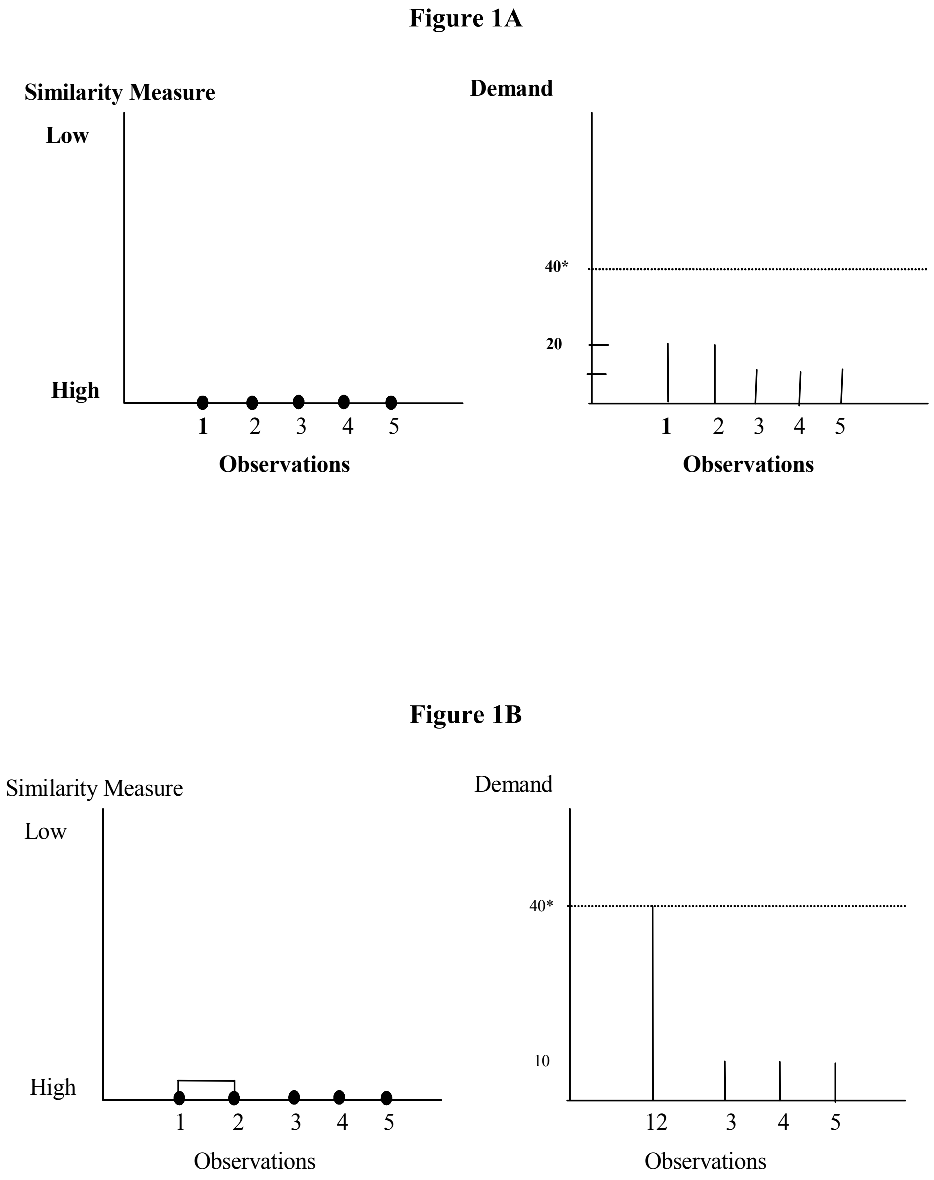


Figure 1A

Figure 1B

Figure 1C

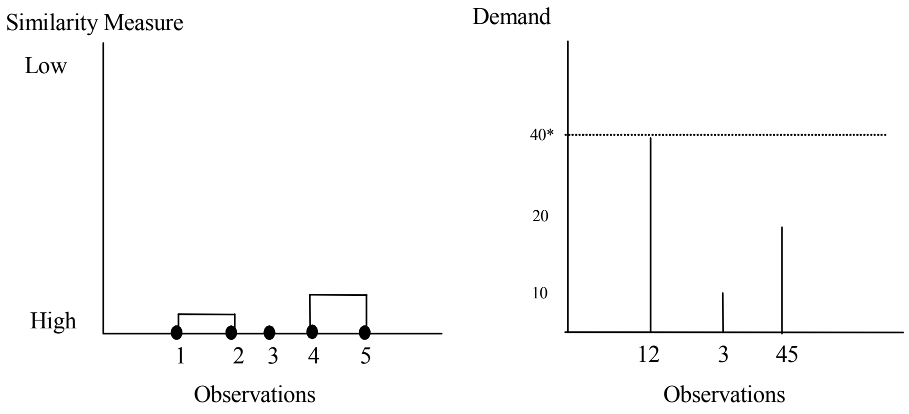


Figure 1D-1

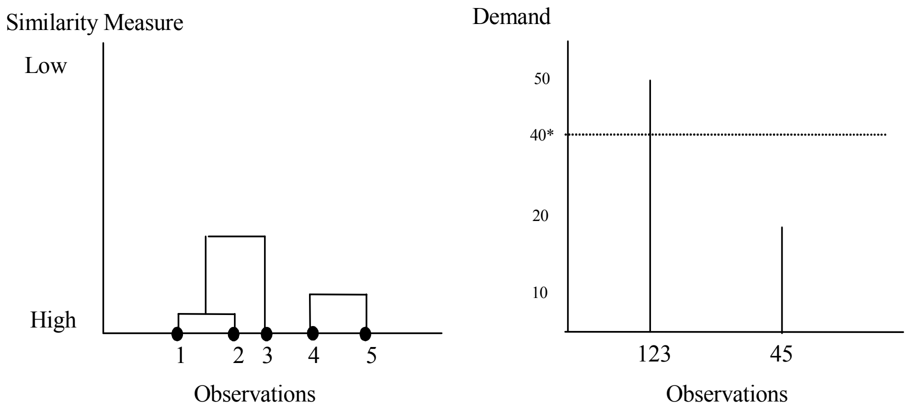
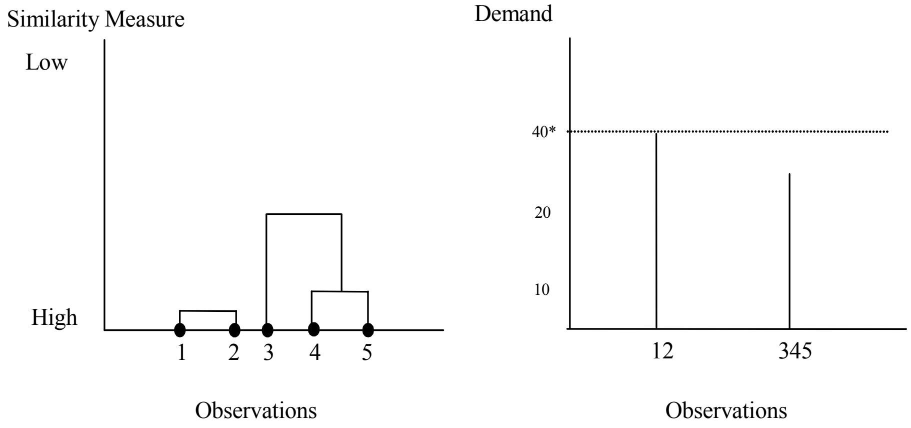


Figure 1D-2



EXAMPLE

Consider the case of five customers. Customers 1 and 2 have a demand of 20 units while customers 3, 4, and 5 have a demand of 10 units. The vehicle capacity is 40 units. Customers 1 and 2 are located nearest to each other, followed by customers 4 and 5, customers 2 and 3, customers 3 and 4, and customers 1 and 5. In the context of a vehicle routing problem, the similarity measure is equivalent to distance. Thus, customers 1 and 2 have the highest similarity, Figure 1A.

Note that when all the customers are considered individually, the similarity measure is the highest. In the first step, customers 1 and 2 are clustered because it results in the smallest decrease in similarity (see Figure 1B). After customer 4 and customer 5 are clustered (see Figure 1C); customer 3 is clustered with the cluster containing customers 1 and 2 (see Figure 1D-1). Note that clustering customer 3 with customers 1 and 2 violates the vehicle capacity constraint. The classical clustering heuristics do not consider constraints and would therefore join the clusters. In the last step, the two remaining clusters would be clustered. Grouping all customers into one cluster results in the lowest level of commonality.

When using the side constraint as a stopping criterion as in Min (1987) and Min (1996), three clusters will be used; a cluster for customers 1 and 2, a cluster for customer 3, and a cluster for customers 4 and 5. In other words, Min's approach stops when the best possible agglomeration first violates the vehicle capacity constraint.

When clustering with side constraints is used, joining customer 3 with customers 1 and 2 also results in violating the constraint. Hence, customers 1, 2 and 3 would not be clustered. The capacitated clustering approach then considers joining customer 3 with customers 4 and 5 (Figure 1D-2). Note that the resulting similarity measure will be worse than it would have been if customers 1, 2, and 3 were to be clustered. However, the similarity measure might still satisfy the termination criterion and the resulting cluster is vehicle capacity feasible. In the final step, the heuristic would attempt to cluster the two remaining clusters. Note that this last step violates the vehicle capacity constraint and is therefore not feasible. The result is two clusters: 1-2 and 3-4-5.

While capacitated hierarchical clustering has received limited attention, various authors have investigated the problem of the k-means capacitated clustering problem. In the k-means clustering problem, the number of clusters is determined *a priori*. Each observation has to be assigned to one of these k clusters and the cluster capacity cannot exceed a given threshold. Note that in hierarchical clustering, the focus of our research, the number of clusters is not determined *a priori*.

Mulvey and Crowder (1979) use Lagrangian relaxation to solve the k-means clustering problem. Mulvey and Beck (1984) extend that research to the capacitated clustering case. Their heuristic starts with a random set of k centers. Customers are assigned to these clusters based on regret value, where regret value is defined as the difference in distance between the nearest and second nearest cluster. As in k-means clustering, the process is iterative and cluster centers are recalculated after all customers are assigned. These new centers are used in the next phase. The process stops when the cluster centers no longer change. The heuristic approach solutions are within 5% of the lower bound for their randomly generated test problems.

Osman and Christofides (1994) develop a hybrid heuristic approach to solve the k-means CCP. Their heuristic approach combines Tabu search and simulated annealing. The hybrid approach outperforms both the Tabu search and simulated annealing approaches for their data sets.

Barreto et al. (2007) develop four heuristics for the capacitated routing problem. Their first heuristic is similar to the approach used by Min (1987). The capacity constraint serves as a stopping criterion for merging of subsets.

The second heuristic is a two-phase heuristic approach for the k-means clustering problem, where the number of clusters formed is equal to the number of available vehicles. In the first phase, the capacity constraints are ignored. Note that this likely results in clusters that violate the capacity constraint. In the second phase, customers from a cluster that violates the capacity constraint are moved to a cluster that can absorb them without violating the capacity constraint. The transfer is based on the relative proximity of the customers to the clusters. If this does not result in a feasible solution, the number of vehicles (i.e., clusters) is increased by one.

The third heuristic is a direct assignment heuristic. As in k-means clustering, the number of clusters is known *a priori*. A set of k customer locations is used as initial customer centers and the remaining customers assigned to these clusters. Unlike k-means clustering, the initial locations are not chosen arbitrarily. The chosen locations are located on the boundary, i.e., chosen based on farthest neighbor.

For the final heuristic, Barreto et al. (2007) extend the direct assignment heuristic. New customers are assigned not solely based on the initial customers' locations but on the location of all customers in the clusters. For their test cases, the first heuristic (one-phase) hierarchical method outperforms the other heuristics. Of the clustering heuristics used, average linkage provides the most consistent results while Ward's method is the least consistent.

HIERARCHICAL CLUSTERING WITH SIDE CONSTRAINTS

As explained in Section 3, clustering uses a distance matrix to determine which subgroups, i.e., observations and/or clusters, will be clustered. In this section, we adapt existing hierarchical clustering approaches to account for side constraints. The algorithm and a description of the approach follow. In the pseudo code we use single linkage to determine which clusters to merge. The code can easily be adapted for alternative approaches (e.g., complete linkage or average linkage) by changing the between cluster distance calculation in step 3.

At each stage, the two nearest subsets are considered for clustering. For this potential cluster, we check whether the resource requirement of the cluster exceeds the resource availability. When the constraints are satisfied, the two subsets are clustered. The distance matrix is modified and the two nearest subsets are considered for clustering. When a constraint is violated, the subsets are not clustered and the next best subsets are considered for clustering. Note that this code is applicable to single linkage, complete linkage, average linkage, and Ward's method. The only difference between these methods is the calculation of the similarity matrix that is used to determine the next subsets to be considered for clusters.

Algorithm for Capacitated Clustering Approach

1. *Start with N clusters or subsets, where each subset contains a single observation. Calculate a N x N distance or similarity matrix* ***D*** *= {d(i,j)}, where d(i,j) is the distance between observations i and j.*
2. *Consider the two nearest subsets for clustering. If the constraints are satisfied, cluster the subset and label the resulting cluster. If clustering the subsets violates the constraints, consider the next nearest subsets for clustering.*
3. *Update the distance matrix. Add the newly formed cluster and remove the merged clusters. The distance between the new cluster K and cluster L can be calculated by:*

$$d_{S_K S_L} = \min_{k \in S_K, l \in S_L} \{d(k,l)\}$$

4. *Go to step 2 and repeat N-1 times*

Thus, in our approach, the clustering procedure is not stopped when the best possible agglomeration violates a constraint. Rather, the approach investigates the next best alternative. If at a given stage, clustering of the remaining clusters does not create a capacity feasible cluster, clustering is terminated.

Potentially, this clustering approach could create some bad clusters. Consider the following example. A company has the vast majority of its customers in the Midwest, one customer in New

York and one customer in California. At the last stage, potentially a cluster could be formed consisting of the customer in New York and the customer in California. While placing these customers on the same route might actually reduce the total travel distance, it would create a very lengthy route. When subsequently the clusters are assigned to warehouses, these savings might turn out to be a mirage. The travel distance incurred by assigning customers to the nearest open warehouse might be significantly smaller than the travel distance incurred when the customers are combined on a route. Additionally, increasing the number of clusters allows for redesigning the outgoing routes from each depot and to reallocate the customers to the depots (Perl, 1983, p. 98). We therefore consider additional stopping criteria.

Because hierarchical clustering techniques do not require any assumptions about the number of clusters to be formed or about the underlying structure, the number of clusters to use is a decision variable. Stopping criteria are proposed to prevent clustering of observations that would create a feasible cluster, but are distant from one another. In this research we consider three stopping criteria: cutting value, minimum number of clusters, and change in cluster variation or elbow rule.

The first criterion is the approach used in Min (1987) and Min (1996). Observations are clustered with a hierarchical clustering approach. The dendogram or tree diagram is then “cut” such that for the number of clusters chosen all the clusters satisfy the constraint.

The second criterion is the ‘natural’ stopping point for the capacitated clustering approach. It is the minimum number of clusters that can be achieved with the clustering approach given the constraint. Perl (1983, p. 98) argues that the greater flexibility associated with using more routes likely results in a difference between the minimum number of routes and the optimal number of routes.

The last criterion is based on the notion that when two distant observations are clustered, the within cluster variation increases. Variation has been used in a number of stopping criteria developed in the prior literature. For instance, Crutcher and Joiner (1977) apply a variance criterion to hierarchical clustering approaches in a meteorology setting. In particular, they propose stopping the clustering heuristic when the variation explained by $k+1$ clusters is not significantly more than the variation explained by k clusters. When the number of clusters is plotted against the variation explained by these clusters, a significant change in variation explained results in a break or elbow. Hence, Crutcher and Joiner (1977) propose using an elbow rule to determine the number of clusters.

Caliñski and Harabasz (1974) propose using a variance ratio criterion or pseudo F-statistic. Milligan and Cooper (1985), in a comparative study of 30 stopping criteria, report that the pseudo F-statistic performs well for their data sets. The method requires the within-cluster variation and the between-cluster variation. The within-cluster variation is denoted by SSW, sum of squares within clusters. In a clustering context, SSW_K provides a measure for the dispersion of the observations within the K clusters. The between cluster variation is denoted by SSB_K, sum of squares between clusters. It is a measure for the separation of the clusters. The total sum of squares (SST) is the sum of SSW_K and SSB_K. The within-cluster variations for cluster S is calculated by:

$$SSW(S) = \sum_{i \in S} ||i,\mu(S)||^2 \quad (1)$$

where, $||i,\mu(S)||^2$ is the squared distance from customer i to cluster S's centroid, i.e., μ(S). The clustering process partitions the data set such that all customers are assigned to exactly one cluster, (2) and (3), and all the clusters contain at least one observation, (4);

$$\bigcup_{j=1}^{K} S_j = C \quad (2)$$

$$S_i \cap S_j = \varnothing \text{ for all } i, j \neq i \quad (3)$$

$$S_j \neq \varnothing \text{ for all } j \in C^C \quad (4)$$

where C is the set of all customers and C^C is the set of all clusters. The total within-cluster variation for all K clusters in C^C is then

$$SSW_K = \sum_{j=1..K} SSW(S_j) \quad (5)$$

SST is equivalent to the within cluster variation for the case where only one cluster, j, is formed.

$$SST = \sum_{i \in C} ||i,\mu(C)||^2, \quad (6)$$

where $||i,m(C)||^2$ is the squared distance of customer *i* to the centroid, i.e., μ(C).
From SSW_K and SST, the between-clusters variation, SSB_K, can be derived:

$$SSB_K = SST - SSW_K. \quad (7)$$

Note that this stopping criterion can be used for capacitated and uncapacitated hierarchical clustering. From SSB_K and SSW_K we can calculate the mean square for between clusters, MSB_K, and the mean square within clusters (MSW_K). From MSB_K and MSW_K a pseudo F-statistic can be calculated.

$$F_K = [SSB_K/(K-1)] / [SSW_K/(N-K)] = MSB_K / MSW_K \quad (8)$$

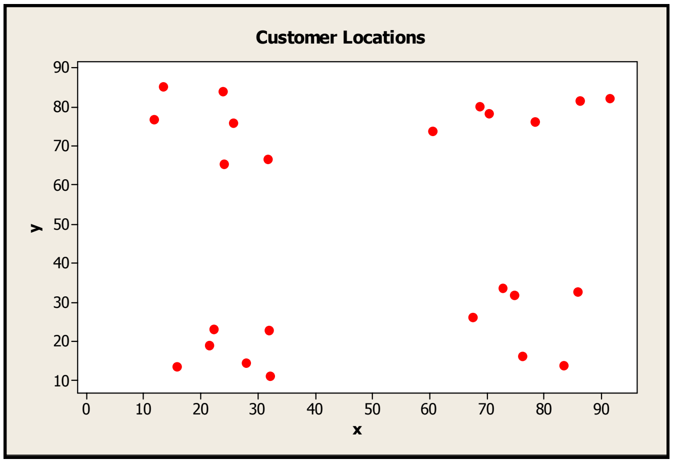
where

F_K = the pseudo F-statistic when K clusters are formed,
SSB_K = sum of squares between clusters for K clusters,
SSW_K = sum of squares within clusters for K clusters,
K = number of clusters formed, and
N = number of observations or customers.

For our stopping rule, we use the change in the pseudo F-statistic to measure the change in variation explained. Starting with the minimum number of clusters required, we calculate the relative change in the pseudo-F statistic if one more cluster were to be used. When the change is relatively small, the number of clusters is increased by one and the next relative change is calculated until a relatively large change is found. When the change is relatively large, an incumbent number of clusters has been found. If increasing the incumbent number of clusters by one results in a small change in the pseudo-F statistic, the incumbent solution is the number of clusters used for the change in variation stopping rule. If the change is relatively large, a new incumbent solution has been found and we increase the number of clusters by one. In this research, we have defined a large change as a change of more than 10%.

The example below applies the stopping criterion to pseudo-randomly generated data. Data was generated such that the underlying structure consists of four distinct clusters (See Figure 2.)

Figure 2: Example data for clustering stopping criteria



For this problem instance, all hierarchical clustering approaches (i.e., single linkage, average linkage, complete linkage, and Ward's method) provide similar results. The data reported below is for the single linkage approach. The within- and between-cluster variations for three clusters, four clusters, and five clusters are reported in the ANOVA table below (see Table 1).

When using three clusters, the within-clusters variation (SSW_3) is relatively large compared to the between-clusters variation (SSB_3). When using four clusters, the within-cluster variation (SSW_4) is relatively small compared to the between-clusters variation (SSB_4). As expected, the accompanying pseudo F-statistic for three clusters is significantly smaller than the pseudo F-statistic for four clusters. Hence, the change in variation stopping rule would suggest using four clusters.

Table 1: ANOVA Table for Stopping Criterion Data

Number of Clusters (K)	Source of Variation	SS	df	MS	Pseudo F_k	Fcrit
3	Between Clusters	23763.7	2	11881.8	17.6617	3.5
	Within Clusters	14127.6	21	672.7		
4	Between Clusters	33970.6	3	11323.5	57.7637	3.1
	Within Clusters	3920.7	20	196.0		
5	Between Clusters	34582.9	2	11881.8	49.6519	3.1
	Within Clusters	3308.4	21	174.1		

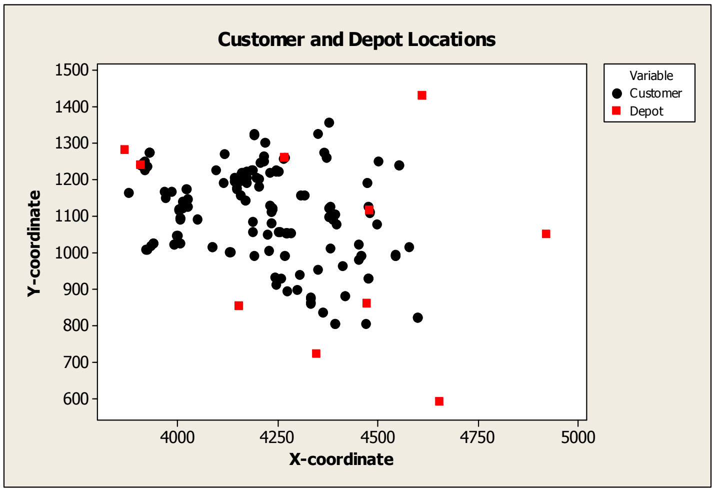
Because we compare the between-cluster variation to the within-cluster variation, we could compare this ratio, i.e., pseudo F-statistic, to a critical F value. Note that we do not use the F-statistic to test a hypothesis. Rather, we use the F-statistic as a benchmark to quantify the difference among different numbers of clusters. The result indicates that using three or more clusters would be significantly better than using only one cluster. The change in within-cluster variation rule, or elbow rule, compares the value of the F-statistics and supports the assessment that there are four underlying clusters.

CAPACITATED CLUSTERING APPLICATION

We test our capacitated clustering approach and stopping rules on the data provided by Min (1996). In his paper, Min uses data from the transportation division of a US company that produces and sells hardware. The distribution system serves 134 customers and 27 vendors. The vehicle capacity is 850 units. The data, as reported in Min (1996), is shown in Figure 3. The coordinates for the depots are those reported in Min's dissertation (1987).

In the location routing literature, clustering is used as a first step to solve the location routing problem. While from a statistical perspective there is a benefit to utilizing more clusters, i.e., more cohesive clusters, in a location routing context, there is a cost associated with increasing the number of clusters. In particular, increasing the number of clusters results in more routes and thus lower vehicle utilizations (see Table 2). We therefore assess the performance of the three alternative stopping rules to the data provided in Min (1996). Because in the paper no data was disclosed about warehouse capacities and costs, we have made the following assumptions. Warehouse capacity is 3,500 units and warehousing costs are $268 per warehouse. Note that: (1) the warehouse capacity is sufficient for the route to warehouse allocation decisions made in Min's dissertation (1987), and (2) the warehouse costs are those used by Barreto et al. (2007) in their modification of the problem provided in Min (1996). Because the fixed warehouse costs are relatively small compared to the routing costs, we perform sensitivity analysis for the fixed warehouse costs.

Figure 3: Location data from Min (1987)



To solve the location routing problem, we use the following heuristic approach. First, we cluster the customers with the capacitated clustering heuristic. For each cluster, we solve the TSP with Helsgaun's (2000) implementation of Lin-Kernighan (1973). We then treat each cluster as a single node and solve the facility location problem (FLP). This approach differs slightly from the approach used by Min (1996). In particular, Min uses the distance from the cluster centroids to the depots rather than the TSP costs.

In Table 2, the results for the heuristic approach are presented. Our example shows that using 23 clusters (the minimum number of required) rather than the 27 clusters (Min's cutting criterion) reduces the overall costs by 5.3%. For this instance, using 32 clusters (change in variation) results in higher overall costs. Note that because the clusters are vehicle capacity feasible, i.e., customer demand for each cluster is less than 850 units, each cluster is equivalent to one truck route.

Table 2: Pseudo-F and Costs by Number of Clusters

Clusters	Pseudo-F	Cost	Avg. Vehicle Utilization
23	58.1590	10634	83.48
24	63.5905	10835	80.00
25	71.7453	11049	76.80
26	74.1047	10978	73.85
27	75.9826	11229	71.12
28	73.6598	11192	68.58
29	74.4470	11309	66.21
30	73.8522	11427	64.00
31	105.328	11638	61.94
32	149.433	11753	60.00

Clusters is the total number of clusters created
Pseudo-F is a measure for ratio of between and within cluster variation
Cost is the sum of the warehousing and routing costs
Avg. Vehicle Utilization is the average percentage of vehicle capacity used

This result suggests that while the variation within clusters increases when the number of clusters decreases, the higher vehicle utilizations results in lower overall transportation costs for the example problem. Note that cost is not a convex function.

In our example, the warehousing costs are relatively low. Sensitivity analysis shows that when we vary the warehousing costs relative to the routing costs, the difference in costs between 23 clusters and 27 clusters remains in the same order of magnitude (see Table 3). A similar result is found when comparing 23 clusters and 32 clusters.

Table 3: Sensitivity Analysis Fixed Warehousing Costs

Adjustment factor	23 Clusters	27 Clusters	Gap (%)	32 Clusters	Gap (%)
0.25	9428	10023	6.3	10547	11.9
0.50	9830	10425	6.1	10949	11.4
1.00	10634	11229	5.6	11753	10.5
1.50	11382	12033	5.7	12557	10.3
2.00	12052	12837	6.5	13361	10.9
3.00	13392	14319	6.9	14969	11.8

Gap is the percentage by which the costs are higher (lower) than the costs for 23 clusters.
Adjustment factor is the factor by which the base case for the fixed warehousing cost is multiplied.

An important observation is that the number of warehouses and the locations of these warehouses is the same for the range of clusters considered above. Hence, the lower costs achieved by combining customers into 23 clusters could potentially be reduced by optimizing the routes after the warehouse decision has been made. This will however require additional computational time.

SUMMARY AND CONCLUSION

In the location routing literature, clustering has been used to proxy routing costs. Because hierarchical clustering does not require selection of the total number of clusters upfront, we have proposed two stopping rules. We developed a constrained clustering heuristic to solve the LRP that utilizes these stopping rules. We provide an application of our approach with a single constraint but the approach is easily extended to account for multiple constraints.

While various stopping rules have been proposed and tested in the prior literature under a variety of settings, our results suggest that consideration of alternative stopping rules in a location routing setting is appropriate. For our example problem, the minimum number of clusters stopping rule finds a lower cost solution than the cutting measure used in the prior literature. We note that the number of warehouses selected and their locations are the same for the range of clusters we consider.

REFERENCES

Barreto, S., Ferreira, C., Paixão, J. & Sousa Santos, B. (2007). Using Clustering Analysis in a Capacitated Location-Routing Problem. *European Journal of Operational Research. 179* (3), 968-977.

Caliñski T. & Harabasz, J. (1974). A Dendrite Method for Cluster Analysis. *Communications in Statistics* 3, 1-27.

Crutcher, H. L. & Joiner, R. L. (1977). Another Look at the Upper Winds of the Tropics. *Journal of Applied Meteorology 16,* 462 -476.

Garey, M. R. & Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP – Completeness*. San Francisco: Freeman.

Helsgaun, K. (2000). An Effective Implementation of the Lin–Kernighan Traveling Salesman Heuristic. *European Journal of Operational Research 126* (1)*,* 106-130.

Jain, A. K. & Dubes, R. C. (1988). *Algorithms for Clustering Data*. Upper Saddle River, NJ: Prentice-Hall.

Johnson, R. A. & Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis* (5th ed.). Englewood Cliffs, NJ: Prentice Hall.

Lin, S. & Kernighan, B. W. (1973). An Effective Heuristic Algorithm for the Traveling Salesman Problem. *Operations Research 21,* 498–516.

Milligan, G. W. & Cooper, M. C. (1985). An Examination of Procedures for Detecting the Number of Clusters in a Data Set. *Psychometrika 50*, 159-79.

Min, H. (1987). *The Vehicle Routing Problem with Product/Spatial Consolidation and Backhauling.* PhD Dissertation. Ohio State University

Min, H. V. (1996). Consolidation Terminal Location-Allocation and Consolidated Routing Problem. *Journal of Business Logistics 17,* 235-263.

Mojena, R. (1977). Hierarchical Grouping Methods and Stopping Rules: An Evaluation. *The Computer Journal 20* (4)*,* 359-363.

Mulvey, J. M. & Beck, M. P. (1984). Solving Capacitated Routing Problems. *European Journal of Operational Research 18,* 339-348.

Mulvey, J. M. & Crowder, H. P. (1979). Cluster Analysis: An Application of Lagrangian Relaxation. *Management Science 25* (4)*,* 329-340.

Osman, I. H. & Christofides, N. (1994). Capacitated Clustering Problems by Hybrid Simulated Annealing and Tabu Search. *International Transactions in Operational Research 1* (3)*,* 317-336.

Perl, J. (1983). *A Unified Warehouse Location-Routing Analysis*. Ph.D. Dissertation. Northwestern University.

INFORMATION SYSTEMS SECURITY AND SAFETY MEASURES: THE DICHOTOMY BETWEEN STUDENTS' FAMILIARITY AND PRACTICE

Ewuuk Lomo-David, North Carolina A&T State University
Li-Jen Shannon, Sam Houston State University

ABSTRACT

Information systems security and safety measures (ISSSM) are attributes that, if properly implemented, contribute to the safety of computer systems, networks and information. This proper implementation will prohibit or delay viruses, malware and hackers from continuing to plague the digital environment. It is our contention in this study that the problem of data and cyber insecurity could be reduced if more systems users become familiar with and use our suggested ISSSM. Information on the relationship between familiarity with and usage of safe computing practices is needed to address this problem. This study analyzes the relationship between students' familiarity with ISSSM and actual usage of these measures on a daily basis. We use survey data from a sample of 867 students for the study. Results indicate that familiarity with ISSSM translates into practical use for six of the ten attributes. The six attributes are simple passwords, sophisticated passwords, daily computer system scan, scan of email attachments, anti-virus software, and firewalls. That four attributes that did not show significant relationships between familiarity and usage underscore the need for educational institutions to supplement methods of disseminating information about safe-computing to students.

INTRODUCTION

One burning issue concerning information security and safety in contemporary digital computing is how university students' computing behaviors enhance or depreciate the safety and security of information in their domain. The overwhelming interest in the subject of digital information systems security has focused on the coder and distributor of virus and spam ware programs all over the internet. The human access component that requires careful protection of data by the end-user has recently become a subject of major discourse. Since the world has millions of students who access the internet every minute of every day, it is imperative for safety and security of information focus to shift to this large group of users to determine if there is a concomitance between what they are familiar with and what they actually practice. Also, the incessant connectivity of corporate and educational digital communication infrastructure and critical information exchange via the World Wide Web created a state of unsurpassed vulnerability

(Crowley, 2003) that is genie-like in scope. This vulnerability calls for a concerted effort to determine if end-users' familiarity with and usage of ISSSM are related.

In 1996, the National Research Council for information security alert and the 1998 Decision Directive 63 by the President on the vulnerability of critical data in cyberspace is indicative of the importance of the problem. To solve this problem requires training and education in management information systems and security specialization degrees. In the same line of thought (Zhang, 2005) agrees that to ensure security of information and avoid spyware invasion of systems require avid vigilance and education in information security issues. Also, the end user needs further education on current computer protection and privacy methodologies and all students should be computer-security literate. Security awareness (Siponen & Kajava, 1998) steadily evolved through the years in three stages: "drawing peoples' attention on security issues, getting users acceptance, and getting users to learn and internalize the necessary security activities." In terms of drawing people's attention to the challenges of information technology, the Federal Executive Council of Nigerian in 2001 approved a National Information Technology Development Agency (NITDA) (Federal Executive Council) to bring information technology closer to the people by ensuring that "the entire citizenry is empowered with information technologies through the development of a critical mass of IT proficient and globally competitive manpower." The organization of the paper from this point on is as follows: related literature, purpose of the study, methodology, data analysis, results, discussion, conclusion, and recommendation for further research.

RELATED LITERATURE

The vocabulary that covers information security is vast but for the sake of brevity, we are going to limit our related literature discourse to password protection security interests.

Password

A password is a system protection or identity releasing, must-remember non-sensible or sensible combination of characters or word that grants or denies access to proprietary systems. Passwords can be categorized into simple and sophisticated. The simple passwords are easy to remember, easy to guess and non-hacker proof. Sophisticated passwords are more difficult to hack and require a combination of letters, numbers, and special characters to make them effective. On the whole, passwords can be algorithmically hashed by a person with avid interest in doing so. In explaining the password concept, (Weinshall and Kikpatrick, 2004) described it as a self-certifying method that requires a conscious effort to recollect. They argue that passwords should be seen as less perfect and therefore advocate the use of human natural characteristics for identification. The first level of software protection for any system is to understand how to create a password and use it to log into a protected system. Passwords, though most popular among the known authentication schemes, is the weakest (Stoller, 2009) because it can be stolen, it may be forgotten and unwittingly openly exhibited. Several researchers have proposed different approaches to password creation and use. Passwords' robustness is an attribute that can prevent unauthorized access to proprietary systems (Oreku & Jianzhong, 2009). Password authentication systems must identify the user and charge a fee based on the number of times of usage without implementing a password table that lends itself to a "replay attack" (Lin & Chang, 2009). Password authentication must be required to identify users who want entry into systems (Chang, Chen, & Hwang, 2004).

Organizations boost the confidence of their clients by having a password requirement combined with preregistered questions and answers in their database. The drawback to the preregistered question and answer such as "what high school did you attend?" is that the answer to such a question cannot be too difficult to obtain by an ardent intruder. We suggest that the registrant be given the opportunity to create at least ten questions and provide the answers concomitantly. At entry beyond the password permission zone, the requester is asked to provide answers to random selection of questions. When the answer provided matches the answer in the database, the requester is given entry permission to the system.

Pervasiveness of security problems

The unfortunate continuous success of intrusions into systems and the attendant loss of capital, money, man hours, and goodwill are attributed to several influences. A study by Teer, Kruck, and Kruck (2007) found that students are not the most savvy when it comes to protecting their passwords. They often allow others to make use of and share their passwords. In social engineering circles, releasing a password to a persistent imposter is not as burdensome as a full-fledged attack on a computer system (Mitnick & Simon, 2002), but it does promote avoidable vulnerability. Research found that using social engineering physical approach to solicit usernames and passwords successfully netted 80% of respondents who released their user names and 60% who released their passwords (Orgill, Romney, Bailey, & Orgill, 2004). In a study at Sydney University, researchers used bogus email to ask students to provide their passwords and usernames for purposes of system upgrade. The result was that 47% of the participants succumbed to the prank (Greening, 1966).

Sometimes institutions understand the challenges that privacy poses but they do not employ new technology for privacy enforcement (Brodie, Karat, & Feng, 2005). The enforcement of privacy policies combined with password use for data protection can mean better data and system handling. Misgivings about negative publicity drive companies that have suffered intrusions to withhold information from the public. This is contained in CSI/FBI Computer Crime Security Survey which also indicates that security incident reporting has increased from 20% to 25% (Gordon, Martin, Loeb, Lucyshyn, & Richardson, 2006). Some companies have strong sentiments about reporting of security breaches because the knowledge, if made public, will present an imperfect persona of the organization (Roberts, 2005). Reporting of intrusions can elicit client legal actions but more importantly is the fact that failure of organizations to install breach control mechanisms such as firewalls and anti-virus software is tantamount to contributing to the problem. A study explored how security breach announcement affected market reactions and thus the value of firms but found the result inconclusive (Cuvosoglu, Mishap, & Raghunatan, 2004).

PURPOSE

The primary purpose of this study is to determine if there is a significant relationship between familiarity with ISSSM on the one hand and actual usage on the other. In other words, do students who say they are familiar with ISSSM also practice the use of these measures in their daily affairs with computers?

DEFINITIONS

Familiarity refers to having a general knowledge of the existence and may be rudimentary functions of an ISSSM. Use refers to having the intellectual, theoretical and practical capacity to apply each ISSSM when the circumstance calls for it.

METHODOLOGY

This research is part of a larger study that is exploring the understanding and appreciation of ISSSM across educational institutions in Turkey, Republic of China (ROC), and Nigeria. The data for this study was extracted from 24 questions that formed three parts of the larger study. In the first section students were asked to indicate on a three-point Likert-type scale whether they are unfamiliar, somewhat familiar or extremely familiar with a given security measure. In the second section they were asked to indicate the percentage of times, (<31%, 31-50%, and >50%), that they use each measure on a regular basis. The third section solicited some demographic information. The survey instrument was widely critiqued by other researchers in Nigeria, Europe and U.S. for redundancy, ambiguity and readability of questions. To administer this instrument, professors at a random sample of 20 out of the 90 member universities of the National Universities Commission that accredits institutions of higher learning in Nigeria were contacted to participate in the study. They administered the survey to their students. The surveys were sent as email attachments to enable participants to download and digitally make their selections and return the instruments via email. Prior to full blown administration of the questionnaire, a 100-person pilot test was conducted to ensure that the statements were easy to understand.

DATA ANALYSIS

Descriptive statistics and cross tabulations were used to analyze the data in this study. In Table 1A are the frequencies/percentages of levels of familiarity (unfamiliar, somewhat familiar and extremely familiar) and levels of usage (<31%, 31%-50%, > 50%).

Table 1A: Frequencies (Percentages) of Levels of Familiarity with and Usage of ISSSM

	Familiarity			Usage (percent of the time used)		
Security Measures	Unfamiliar	Somewhat Familiar	Extremely Familiar	<31%	31-50%	>50%
Simple passwords	69(8%)	202(23%)	596(69%)	217(25%)	99(11%)	551(64%)
Sophisticated passwords	757(87%)	85(10%)	25(3%)	780(90%)	52(6%)	35(4%)
Daily computer system scan	432(50%)	164(19%)	271(31%)	74(9%)	191(22%)	602(69%)
Scan of email attachments	651(75%)	79(9%)	137(16%)	110(13%)	291(34%)	466(54%)
Anti-virus software	484(56%)	171(20%)	212(25%)	510(59%)	230(27%)	127(15%)
Password on email attachments	682(79%)	79(9%)	106(12%)	712(82%)	152(18%)	3(.3%)
Biometric authentication	819(95%)	38(4%)	10(1%)	847(98%)	11(1%)	9(1%)
Firewalls	544(63%)	184(21%)	139(16%)	673(78%)	135(16%)	59(7%)
Intrusion detection systems	386(45%)	180(21%)	301(35%)	759(88%)	69(8%)	39(5%)
Multifaceted authentication systems	818(94%)	41(5%)	8(1%)	853(98%)	8(1%)	6(1%)

Table 1b: Frequencies (Percentages) of Levels of Familiarity with and Usage of ISSSM

	Familiarity		**Usage (percent of the time used)**	
Security Measures	**Unfamiliar**	**Familiar**	**<31%**	**31-100%**
Simple passwords	69(8%)	798(92%)	217(25%)	650(75%)
Sophisticated passwords	757(87%)	110(13%)	780(90%)	87(10%)
Daily computer system scan	432(50%)	435(50%)	74(9%)	793(91%)
Scan of email attachments	651(75%)	216(25%)	110(13%)	757(88%)
Anti-virus software	484(56%)	383(45%)	510(59%)	357(42%)
Password on email attachments	682(79%)	185(21%)	712(82%)	155(18.3%)
Biometric authentication	819(95%)	58(5%)	847(98%)	20(2%)
Firewalls	544(63%)	323(37%)	673(78%)	194(23%)
Intrusion detection systems	386(45%)	481(56%)	759(88%)	108(13%)
Multifaceted authentication systems	818(94%)	49(6%)	853(98%)	14(2%)

RESULTS

Demographics

One thousand one hundred surveys and 867(79%) usable responses were distributed and returned respectively. Demographics information are as follows: Gender: female (54%), male (46%); Classification: undergraduate (63%), graduate (38%); Major: Arts & Sciences (29%), Business (37%), Engineering (18%), others (16%); Level of experience in computing: Expert (46%), Very good (22%), Good (18%), Poor/Novice (14%).

Comparative Descriptions of ISSSM

The following are relevant descriptions of each measure accompanied by representative graphics. Figures 1A to 10B are paired graphical representations of the contents of Table 1A.

Simple Passwords

Specifically, Figure 1A indicates that while 69% of respondents are extremely familiar with or aware of simple passwords, 64% use it fifty percent of the time and greater (see Figure 1B). Usage of simple passwords by only 64% of students is not extremely impressive considering the fact that even a simple password is necessary to keep some data safe, maintain some system integrity and delay some intrusions. We expected more than 70% of respondents to actively use simple passwords more than 50% of the time.

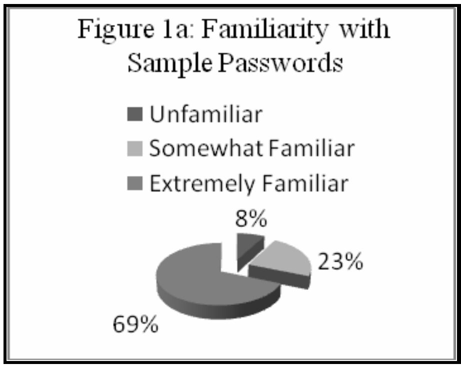


Figure 1a: Familiarity with Sample Passwords

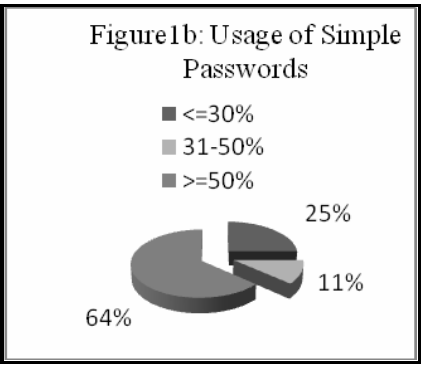


Figure1b: Usage of Simple Passwords

Sophisticated Passwords

Figures 2a and 2b illustrate that 87% of respondents are unfamiliar with sophisticated passwords. To add to that, only a dismal 4% use it more than 50 percent of the time. This should raise an alarm because the non-use or non-application of sophisticated passwords by 96% (100%-4%) of students less than 50% of the time is a perilous contribution to the problem of system compromise.

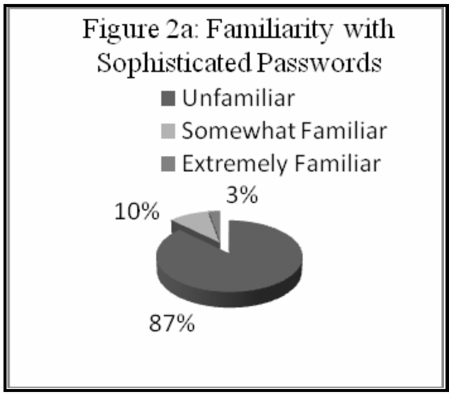


Figure 2a: Familiarity with Sophisticated Passwords

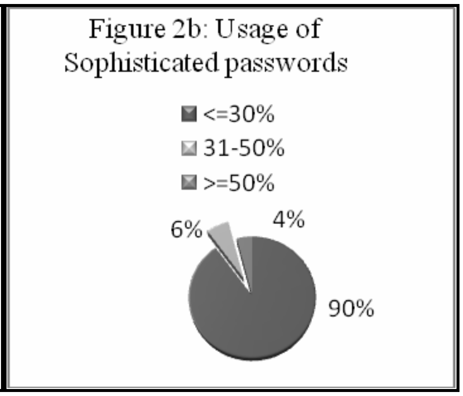


Figure 2b: Usage of Sophisticated passwords

Daily Computer System Scan

Figures 3a and 3b show that 50% of respondents are unfamiliar with daily computer systems scan but 69% use it more than 50% of the time. Because daily computer system scan is an automatic process in contemporary computing most people may probably know that it is happening during the boot process and interpret that as using it. The test of actual usage may come from response to a system glitch combined with the need to successful execute an immediate system scan.

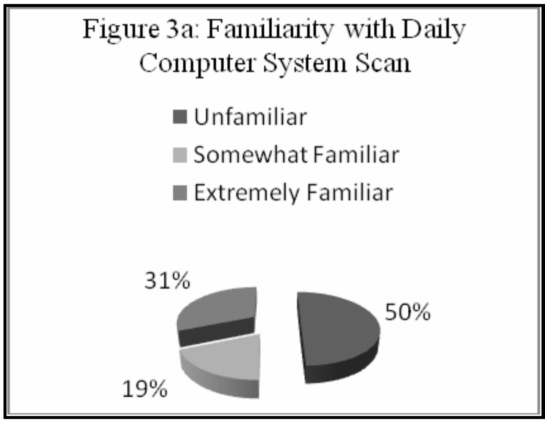


Figure 3a: Familiarity with Daily Computer System Scan

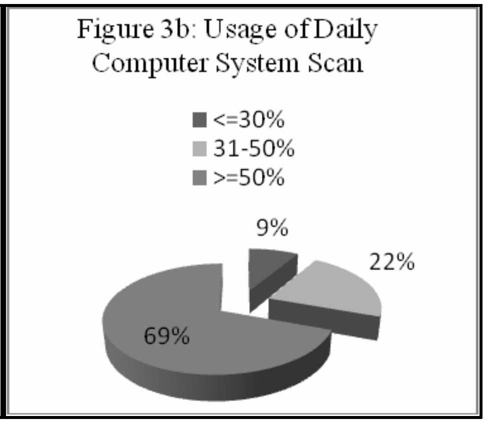


Figure 3b: Usage of Daily Computer System Scan

Scan of Email Attachments

Figures 4a and 4b indicate that 75% of respondents are unfamiliar with scan of email attachments while 54% use it more than 50% of the time. Again, since email scanning is generally an automatic process respondents may consider familiarity and usage to fall in the same realm of understanding and therefore claim usage.

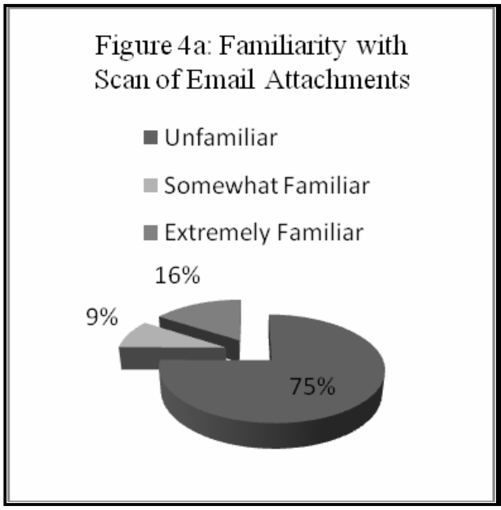


Figure 4a: Familiarity with Scan of Email Attachments

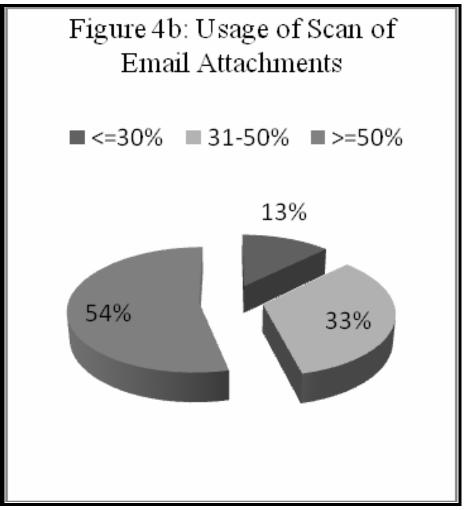


Figure 4b: Usage of Scan of Email Attachments

Anti-virus software

Figures 5a and 5b indicate that 56% of respondents are unfamiliar with anti-virus software but only 15% use it more than 50% of the time. Most computer systems today have preinstalled anti-virus software or have an online access to one and therefore usage may be automatic. The data that indicates that only 15% use it more than 50% of the time may be a reflection of those who do not have an online access to anti-virus software and therefore have to purchase and install their own copy.

Figure 5a: Familiarity with Anti-virus Software

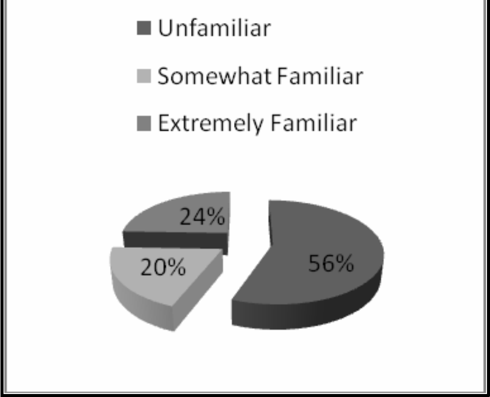
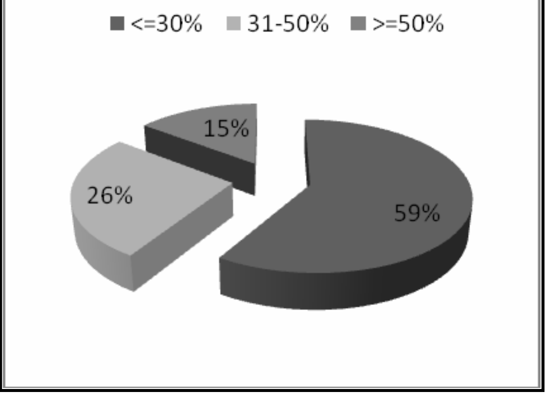


Figure 5b: Usage of Anti-virus Software



Password on Email Attachments

Figure 6a indicates that 79% of respondents are unfamiliar with creation of a password, building it into a file and attaching the file to an email message. Figure 6b indicates that less than 1% use passwords on email attachments more than 50% of the time. That 79% of respondents are unfamiliar with password attachment and a dismal less than 1% use it for more than 50% of the time should be unacceptable in contemporary computing. This is a reflection of findings in previous studies (Teer, Kruck, & Kruck, 2007) and (Aytes & Connolly, 2004) that show students disinterest in computers and data safety.

Figure 6a: Familiarity with Password on Email Attachments

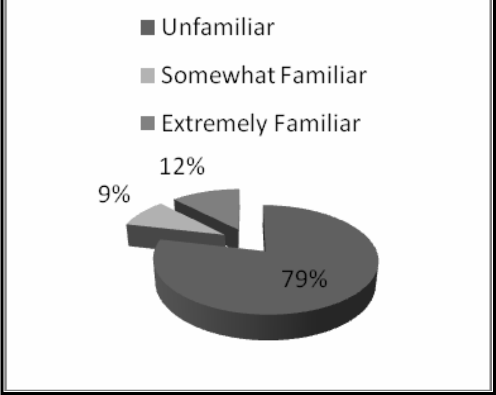
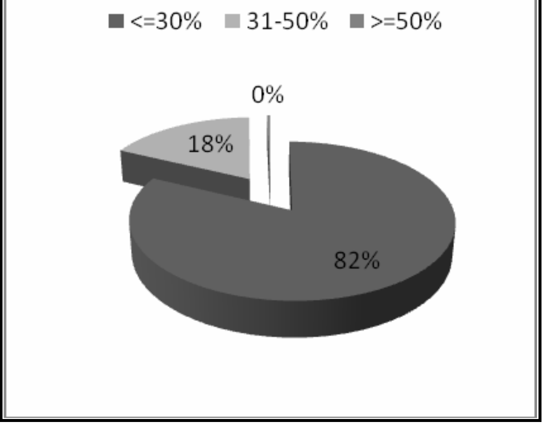


Figure 6b: Usage of Passwords on Email Attachments



Biometric Authentication

Figure 7a indicates that 95% of respondents are unfamiliar with biometric authentication while only 1% uses it more than 50% of the time (Figure 7b). Since biometric authentication uses the uniqueness of what humanity already has such as finger printing or retinal scanning and we do not have to make an effort to remember anything such as passwords, it is a technology that should be required to interface between all systems users and systems. It is hardly surprising to learn that 95% of students are unfamiliar with biometric authentication.

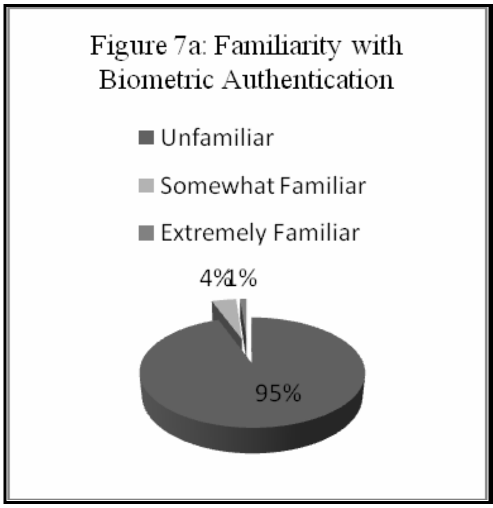


Figure 7a: Familiarity with Biometric Authentication

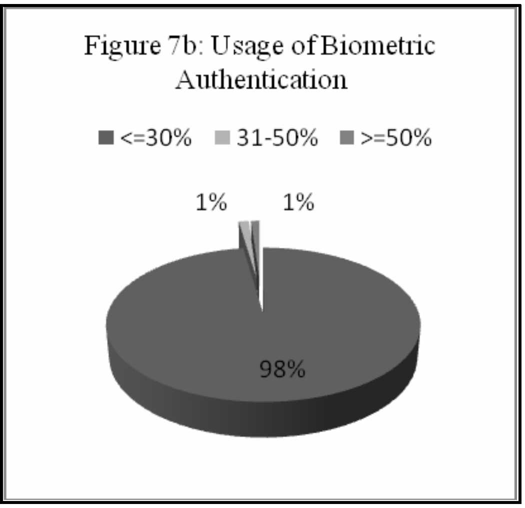


Figure 7b: Usage of Biometric Authentication

Firewalls

Figure 8a shows that 63% of respondents are unfamiliar with firewalls while only 7% use it more than 50% of the time (Figure 8b). Firewalls filter incoming traffic before they arrive at the computer station and therefore their presence may not be apparent to the non-savvy user.

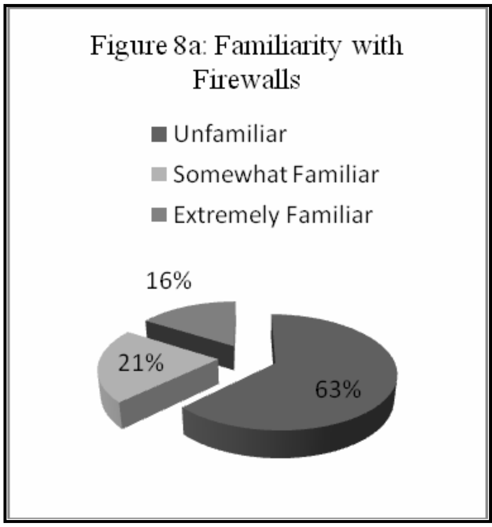


Figure 8a: Familiarity with Firewalls

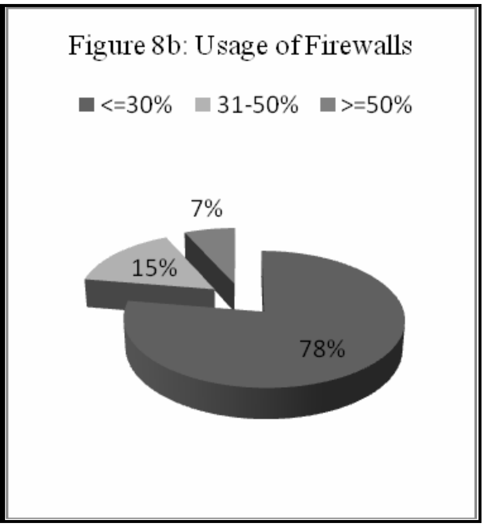


Figure 8b: Usage of Firewalls

Intrusion Detection Systems

Figure 9a shows that 44% of respondents are unfamiliar with intrusion detection systems while 4% (Figure 9b) use it more than 50% of the time.

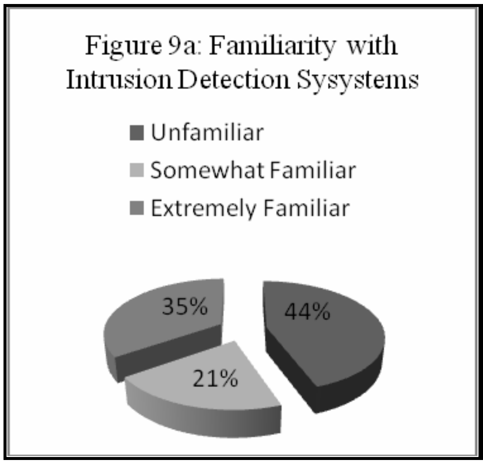


Figure 9a: Familiarity with Intrusion Detection Sysystems

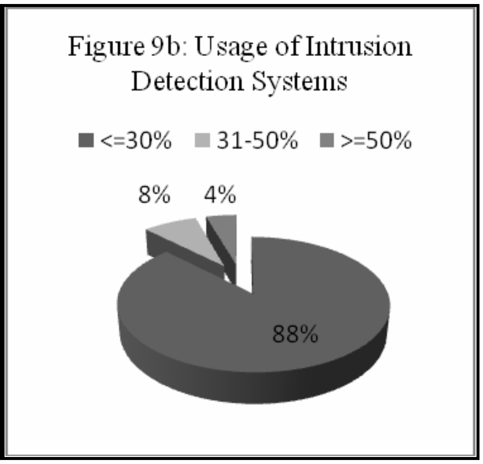


Figure 9b: Usage of Intrusion Detection Systems

Multifaceted Authentication Systems

Figure 10a shows that 94% of respondents are unfamiliar with multifaceted authentication systems while less than1% (Table 10b) uses it more than 50% of the time.

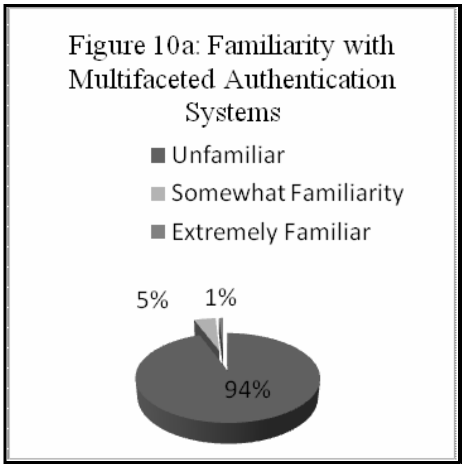


Figure 10a: Familiarity with Multifaceted Authentication Systems

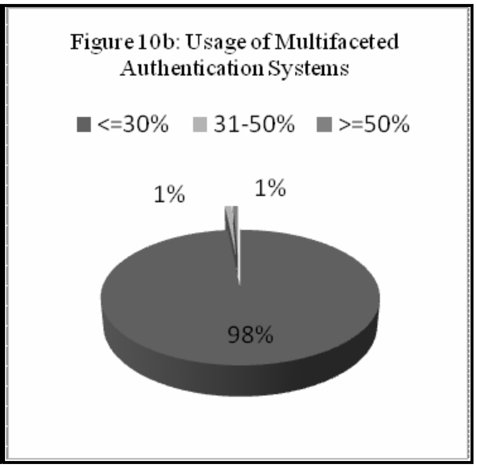


Figure 10b: Usage of Multifaceted Authentication Systems

HYPOTHESES TESTED FOR THIS STUDY

Table 2 shows the null hypotheses for this study.

Table 2: Hypothesis for the Study	
1	There is no significant relationship between familiarity with and usage of simple passwords as security measure.
2	There is no significant relationship between familiarity with and usage of sophisticated passwords as security measure.
3	There is no significant relationship between familiarity with and usage of daily computer system scan as security measure.
4	There is no significant relationship between familiarity with and usage of scan of email attachments as security measure.
5	There is no significant relationship between familiarity with and usage of anti-virus software as security measure.
6	There is no significant relationship between familiarity with and usage of passwords on email attachments as security measure
7	There is no significant relationship between familiarity with and usage of biometric authentication as security measure.
8	There is no significant relationship between familiarity with and usage of firewalls as security measure.
9	There is no significant relationship between familiarity with and usage of intrusion detection systems as security measure.
10	There is no significant relationship between familiarity with and usage of multifaceted authentication systems as security measure.

Table 3 shows the results of SPSS 15 cross tabulations and Chi-Squares of familiarity with and usage of information systems security and safety measures.

Table 3: Cross Tabulations and Chi-Squares Analyses of Familiarity with and Usage of Information Systems Security and Safety Measures

	Familiarity versus Usage	Chi-Square Value	df	Significant at .05
$H_0$1	Simple passwords	20.506	4	.000*
$H_0$2	Sophisticated passwords	12.81	4	.012*
$H_0$3	Daily computer system scan	16.215	4	.003*
$H_0$4	Scan of email attachments	105.283	4	.000*
$H_0$5	Anti-virus software	11.749	4	.019*
$H_0$6	Passwords on email attachments	5.832	4	0.212
$H_0$7	Biometric authentication	7.733	4	0.102
$H_0$9	Intrusion detection systems	8.9	4	0.064
$H_0$10	Multifaceted authentication systems	0.852	4	0.931

*Significant at $p<.05$

Hypothesis 1: Simple passwords. We did find a significant relationship between familiarity with and usage of simple passwords at the .05 level.

Hypothesis 2: Sophisticated passwords. We found a significant relationship between familiarity with and usage of sophisticated passwords at the .05 level.

Hypothesis 3: Daily computer systems scan. We found a significant relationship between familiarity with and usage of daily computer systems scan at the .05 level.

Hypothesis 4: Scan of email attachments. We found a significant relationship between familiarity with and usage of Scan of email attachments at the .05 level.

Hypothesis 5: Anti-virus software. We found a significant relationship between familiarity with and usage of anti-virus software at the .05 level.

Hypothesis 6: Passwords on email attachments. We found no significant relationship between familiarity and usage of scan of email attachments at.05 level.

Hypothesis 7: Biometric authentication. We found no significant relationship between familiarity with and usage of biometric authentication at the .05 level.

Hypothesis 8: Firewalls. We found a significant relationship between familiarity with and usage of Firewalls at the .05 level.

Hypothesis 9: Intrusion detection systems. We found no significant relationship between familiarity with and usage of intrusion detection systems at the .05 level.

Hypothesis10: Multifaceted authentication systems. We found no significant relationship between familiarity with and usage of multifaceted authentication systems at the .05 level.

DISCUSSION

The discussion below will focus on the six ISSSM that showed significant relationships between familiarity and usage at the .05 level. The other four that did not show significant relationship will not be discussed in detail.

Significant ISSSM: Regarding simple passwords, we found that a large percentage (69%) of respondents is extremely familiar with it and do actually use it in their daily access to computer systems. Sixty four percent of respondents use it more than 50% of the time. When we added the 64% of those who use it more than 50% of the time to the 11% who use it between 31% to 50% of the time, we found that 75% of respondents use simple passwords 31%-100% of the time. This indicates that being familiar with simple passwords does translate into its use. It should be noted that simple passwords do not protect malicious entry into a system as much as sophisticated passwords. As for sophisticated passwords 87 percent of respondents are unfamiliar with it and 90% use it less than 31% of the time. In this case, unfamiliarity with sophisticated passwords translates into non-use. This is understandable because familiarity has to precede usage which in turn bolsters familiarity.

Daily computer system scan occurs each time you turn on the computer. Even though fifty percent of respondents are unfamiliar with daily computer system scan, 69% use it more than 50% of the time. Since the process is automatic, even those who are unfamiliar with it have their computing devices scanned during each access.

Scanning email attachments ensures that viruses embedded in files do not infect a computer system. In most systems this is an automatic process but there are computers that require users to manually scan the system for viruses. Seventy one percent of respondents are unfamiliar with scanning systems for viruses even though 54% use the process more than 50% of the time. The high percentage of users may be due to the fact that most system scans for viruses in file attachments are automatic and require no input from users.

Anti-virus software is designed to protect computer systems from being infected by viruses. Fifty six percent of respondents are unfamiliar with anti-virus software. Only 15% use it more than 50% of the time. The unfamiliarity of a large percentage of respondents explains the very low 15% usage. Because anti-virus software scanning is an automatic process, several end-users may not be aware that their computers are constantly being scanned for viruses. It is occasionally that end-users will conduct a manual scan for viruses because the automatic process is obsolete and needs updating. Firewalls are either software or hardware that filters information coming into computer systems. Sixty three percent of respondents are unfamiliar with firewalls while 78% use it less than 30% of the time. Because firewalls are designed to automatically protect computer systems, most users may neither be familiar with it nor know that they use it, if they are not very savvy in computing.

CONCLUSION

Based on the results from this study, we conclude that students who are familiar with the functions of simple passwords are also practical users and therefore may have some simple protection of their systems or files. The predictive power of hypothesis one was p=.000, indicating a high probability that the two factors, familiarity and usage are significantly related. Regarding sophisticated passwords, a relationship does exist between familiarity and usage thus indicating that these respondents' files can be protected better than others. The significant relationship found between familiarity and usage of daily computer system scan indicates that these respondents can have a safer computing experience. Familiarity and usage are highly related in the case of scanning of email attachments thus meaning that these respondents' emails attachments will not be easily infected by viruses. In the case of anti-virus software, a relationship does exist between familiarity and usage thus confirming that respondents who are familiar with this factor also use it to ensure that system integrity is not compromised. Testing familiarity with and usage of firewalls indicated a strong relationship. This attests to the fact that respondents who are familiar with firewalls use it to shield their systems from invasion by rogue software.

Familiarity and usage did not show a significant relationship in the cases of placing passwords on email attachments, biometric authentication, intrusion detection systems and multifaceted authentication systems. This lack of significant relationship is an indication that familiarity in these cases does not translate into usage and therefore renders computer systems less safe.

To secure the network servers and protect the users' privacy, many experts suggested providing options of access control and authentication. While many companies have been providing various services to satisfy the customers' needs which include digital advertisement, marketing, music, gaming, video, network, and many others, it is vital to make the application development environment publicly available so that it becomes easier for application developers to apply security to programs designed for the users (Ahamad, 2008). Moreover, the digital protection awareness program might be invaluable in higher education institutions to prepare our prospective employees entering into the digital work life environment of the 21 century.

FURTHER RESEARCH

Further research using the same framework might be conducted and targeted at Fortune 500 companies. It will be interesting to find the status of familiarity with and usage of ISSSM in the healthcare industry and among academics as well. This study might be replicated in the US, Canada, Europe and other African countries. The current study identified six ISSSM that lend themselves to familiarity and usage. A study should be conducted to identify the attributes that make these six ISSSM amenable to the concomitance of familiarity and usage. Appendix A shows the instrument used for this study.

REFERENCES

Ahamad, M. (2008). *Emerging cyber threads report for 2009.* Georgia: Georgia Tech Information Security Center.

Aytes, K. & T. Connolly (2004). Computer security and risky computing practices: A rational choice perspective. *Journal of Organizational and End User Computing,*16(3), 22-40.

Brodie, C., C.Karat & J.Feng (2005). *Usable security and privacy: as case study of developing privacy management tools.* Pittsburgh, PA: SOUPS, (July), 6-8.

Chang, C.C., K.L. Chen & M.S. Hwang (2004). End-to-end security protocol for mobile communications with end-user identification/authentication. *Wireless Personal Communication,* 28(2), 95-106.

Crowley, E. (2003). Information System Security Curricular Development. *Proceedings of the 4th Conference on Information Technology Curriculum*, 249-255.

Cuvosoglu, H., B. Mishap & S. Raghunatan (2004). The effect of Internet Security Breach Annoucement on Market Value: Capital Market Reactions for Breached Firms and Internet Security Developers. *International Journal of Electronic Commerce*, 9, 69-104.

Federal Executive Council, Nigeria (2008). *National Informatioin Technology Development Agency*. 12(3), Retrieved December 3, 2008 fromhttp://www.nitda.gov.ng.

Gordon, L., L. Martin, M. Loeb, W. Lucyshyn, & R. Richardson, (2006). CSI/FBI Computer Crime and Security Survey.

Greening, T. (1966). Ask and Ye Shall Receive: A Study in Social Engineering. *ACM SIGSAC Review* 14(2), 8-14.

Lin, I. & C. Chang (2009). A countable and time-bound password-based user authentication scheme for the application of electronic commerce. *Information Science*, 179, 1269-1277.

Mitnick, K. & W. Simon (2002). *The Art of Deception: Controlling the Human Elements of Security.* Indianapolis, IN: Wiley Publishing, Inc.

National Research Council (1996). U.S. Policies Should Foster Broad Use Of Encryption Technologies. Retrieved July 15, 2009, http://epic.org/crypto/reports/nrc_release.html

Oreku, G.S. & L. Jianzhong (2009). End-User Authentication (EUA) Model Password Security. *Journal of Organizational and End-User Computing* 21(2), 28-,(16 pages).

Orgill, G.L., et al. (2004). The urgency for effective user privacy education to counter social engineering attacks on secure computer systems. *Proceedings of the 5th Conference on Information Technology Education,* 171-181.

Presidential Decision Directive (1998). The Clinton Administration's Policy on Critical Infrastructure Protection: Presidential Decision Directive 63, Retrieved July 16, 2009 from http://www.fas.org/irp/offdocs/paper598.htm,

Roberts, K. (2005). Security Breaches, Privacy Intrusions, and Reporting of Computer Crimes. *Journal of Information Privacy & Security,* 1(4), 22-33.

Siponen, T.M. & J. Kajava (1998). The dimensions and categories of information security awareness. *Proceedings of the IFIP TC11 14th Internaitonal Conference on Information Security*.

Stoller, J. (2009). Authentication-passwords and beyond. *CMA Management,* 44(3), 44-46.

Teer, F.P., S.E. Kruck & G.P. Kruck (2007). Empirical Study of Students' Computer Security - Practices/Perceptions. *Journal of Computer Information Systems,* 47(3), 105-110.

Weinshall, D. & S. Kirkpatrick (2004). Passwords you will never forget. *ACM*. Vienna, Austria: ACM, April, 1399-1402.

Zhang, X. (2005). What do consumers really know about spyware? *Communications of the ACM* 48(8), 44-48.

Appendix A

INFORMATION SYSTEMS SECURITY AND SAFETY MEASURES QUESTIONNAIRE					
	This survey is designed to obtain information about security practices regarding computer and information technology usage. Your responses will remain anonymous and you will not be identified in any way.				
	SECTION I: FAMILIARITY AND CONFIDENCE WITH COMPUTER SECURITY MEASURES				
	Please circle your level of familiarity/confidence with the computer security measures below.				
1	Use of simple passwords to protect your computer data and software.	Not Familiar	Somewhat Familiar	Extremely Familiar	
2	Use of sophisticated passwords to protect your computer data and software.	Not Familiar	Somewhat Familiar	Extremely Familiar	
3	Daily computer system scan	Not Familiar	Somewhat Familiar	Extremely Familiar	
4	Scan of email attachments	Not Familiar	Somewhat Familiar	Extremely Familiar	
5	Functions and usage of anti-virus software	Not Familiar	Somewhat Familiar	Extremely Familiar	
6	Placements of passwords on email attachments before sending.	Not Familiar	Somewhat Familiar	Extremely Familiar	
7	Functions of biometric authentication as a security measure	Not Familiar	Somewhat Familiar	Extremely Familiar	
8	Functions of firewalls as security measures	Not Familiar	Somewhat Familiar	Extremely Familiar	
9	Functions of intrusion detection systems as security measures	Not Familiar	Somewhat Familiar	Extremely Familiar	
10	Functions of multifaceted authentication systems	Not Familiar	Somewhat Familiar	Extremely Familiar	
	SECTION II REGULARITY OF USAGE AND PRACTICE OF SECURITY MEASURES				
	Indicate on average the percentage of times you use or practice the following measures when computing.				
11	Use of simple passwords to protect your computer data and software.	<=30%	31-50%	>50%	
12	Use of sophisticated passwords to protect your computer data and software.	<=30%	31-50%	>50%	
13	Daily computer system scan	<=30%	31-50%	>50%	
14	Scan of email attachments	<=30%	31-50%	>50%	
15	Functions and usage of anti-virus software	<=30%	31-50%	>50%	
16	Placements of passwords on email attachments before sending.	<=30%	31-50%	>50%	
17	Functions of biometric authentication	<=30%	31-50%	>50%	
18	Functions of firewalls as security measures	<=30%	31-50%	>50%	
19	Functions of intrusion detection systems	<=30%	31-50%	>50%	
20	Functions of multifaceted authentication systems	<=30%	31-50%	>50%	

INFORMATION SYSTEMS SECURITY AND SAFETY MEASURES QUESTIONNAIRE					
	SECTION III DEMOGRAPHICS				
21	Please circle your gender.	Female	Male		
22	Please circle your classification	Undergrad	Graduate		
23	Please circle your major in the university	Arts/Sciences	Business	Engineering	Other
24	Rate your knowledge/experience with computers	Expert	Very good	Good	Poor
	Thanks for participating in this survey. Please write any comments here:_______________________________				

THE MACHINIST'S SEQUENCING DILEMMA

Arben Asllani, University of Tennessee at Chattanooga
Charles White, University of Tennessee at Chattanooga
Cynthia White, University of Tennessee at Chattanooga

ABSTRACT

We examine a single machine sequencing problem of n-jobs which are applied to a single product. While each job adds value to the product, the accumulated product value is always at risk due to a given chance that the next job will fail. Once a job fails the product is considered defective and cannot be repaired. The goal is to find the sequence of jobs which minimizes the expected value of failure. This sequencing problem, which we call the "machinist's sequencing dilemma", depends on the value of each job and the likelihood that the job will fail. We offer three solutions to this problem: a decision tree approach, a 0-1 mathematical programming, and a genetic algorithm.

INTRODUCTION

A highly qualified machinist complained to one of the authors of this paper that his job was intolerably stressful. He routinely performed a difficult machining procedure at the end of a sequence of other machining processes on a particular part. As the last machining process, the machinist was responsible for finishing a very expensive part due to the value added by the other machining jobs. If the machinist made a mistake, the entire part would have to be scrapped negating the value of all other work done on the part being created. He complained that if his machining operation were the first in sequence, he would have less stress. In this situation, if he ruined the part by cutting the metal too deeply, he would only be responsible for his labor. As the last machinist, he was essentially responsible for everyone's work, since a failure on his part would result in failure for all previous work. As an example, suppose a part must be (1) cut from stock metal, and then (2) machined on a lathe to 0.02 inch for the length of the stock, and then (3) threads cut on one end of the piece, and finally (4) a notch cut for a locking key at the opposite end of the threaded part. Any machining operation has the potential to ruin the piece if improperly done. As the part is passed from one machinist to another, the piece becomes more valuable due to the previous machining operations. If the last person cutting the notch makes a mistake, the work of all previous machinists is lost along with the final machinist's actions.

Although the example cited above is rather trivial, the machinist who complained to the author described a part he worked on that had a large number of very expensive machining processes. An example of such a process would be a military tomahawk made by RMJ Tactical, LLC. This device requires the metal be (1) straightened [1% failure rate], (2) Blanchard grinding

[6% failure rate], (3) profile machining [5% failure rate], (4) edge chamfered [failure rate 2%], (5) machining the "beard" [3% failure rate], (6) machining the cutting edge [5% failure rate], (7) spike chamfering [5%failure rate], and (8) handle injection molding [23% failure rate]. Each of these eight operations are independent and results in a ruined product upon failure. The most cost efficient sequence of operations is not intuitively obvious.

In the machinist's sequencing dilemma (MSD), *n* jobs must be performed on a given product. Associated with each job *j* is a value[1] p_j (j= 1, 2… n) and a probability of failure f_j with $0<f_j<1$. The jobs will be processed sequentially until one job fails. In that case the product becomes defective and not repairable. The value of jobs accumulated up to that point will be lost. If no job fails the product is not defective and no losses occur. We assume for the problem that set-up times are part of the cost of the machining problem; that is, as the part is transferred from one machining operation to the next, the total cost incurred includes set-up time, machining, inspecting, and transferring the product to the next work station. Each machining operation is considered totally independent of all other operations. We also assume that no machining operation affects the probability of success or failure of future machining operations.

For any sequence S, the accumulated value of the *k*-th job is the value added of the first *k* jobs calculated as $C^S_{\max(k)} = p_{s(1)} + p_{s(2)} + ... + p_{s(k)}$, where $p_{s(I)}$ is the processing time of job *i* in sequence S. Job processing is statistically independent and so for any sequence *S* the probability that *k*-th job will be carried out is $Q^s_k = q_{s(1)} q_{s(2)} ... q_{s(k-1)}$ where $q_{s(i)} = 1 - f_{s(i)}$ is the probability that job *i* for (*i*=1…k) in sequence *S* will not fail. The central question of this paper is to provide how should a series of jobs, each with a probability of failure and with an independent value added to a product be sequenced in order to minimize the risk of losing the accumulated cost. The accumulated cost at risk involved in this scenario is sequence dependent and not intuitively obvious. Improper sequencing in this type scenario can lead to substantial losses; that is, optimal solutions may result in significant savings over suboptimal approaches.

The purpose of this paper is to introduce and provide a theoretical formulation for the machinist's dilemma sequencing problem. We also offer three solution techniques (decision tree, mathematical programming, and genetic algorithm) which provide optimal or near optimal solutions to the problem.

LITERATURE REVIEW

The subject of scheduling and sequencing has been one of the most studied areas among scholars and practitioners. A number of survey papers (Graham et al., 1979; Lawler et al.,1982; Lawler, 1983; Lenstra and Rinnooy Kan, 1985; Lawler et al, 1993) have been written on this subject as well as the books by Pinedo(1995) and Conway et al. (2003). One of the major results of the scheduling research is the successful addressing of many interesting and important problems through use of optimization and heuristic solution approaches (Bellalouna and Jaillet, 2007). Stochastic scheduling is among these interesting and challenging problems and has always been the focus of the research. However, stochastic cases of scheduling have been mainly limited to random processing times (Soroush, 1996; Xia et al., 2008) random release time (Hsieh et al. 2006), and random due dates (Cai and Zhou, 2005). A comprehensive discussion of stochastic scheduling problems is provided in Chapter 8 of Pinedo (1995).

The MSD is an entirely different stochastic scheduling problem. We assume that the solution for this scheduling problem is used for a given long period of time and that for this time horizon, the set of jobs to be processed on a daily basis varies. Further, we assume that the machinist cannot re-optimize and will usually use a pre-designed sequencing heuristic hoping that the expected loss of accumulated value will be minimized. As such, this sequencing problem belongs to a family of scheduling problems, whose common characteristic is the explicit inclusion of probabilistic "elements" in the problem definitions. These problems are known as a-priori scheduling problems. This approach was used for the stochastic case of traveling salesman problem when probabilistic elements change (Jaillet 1988; Bianchi 2002).

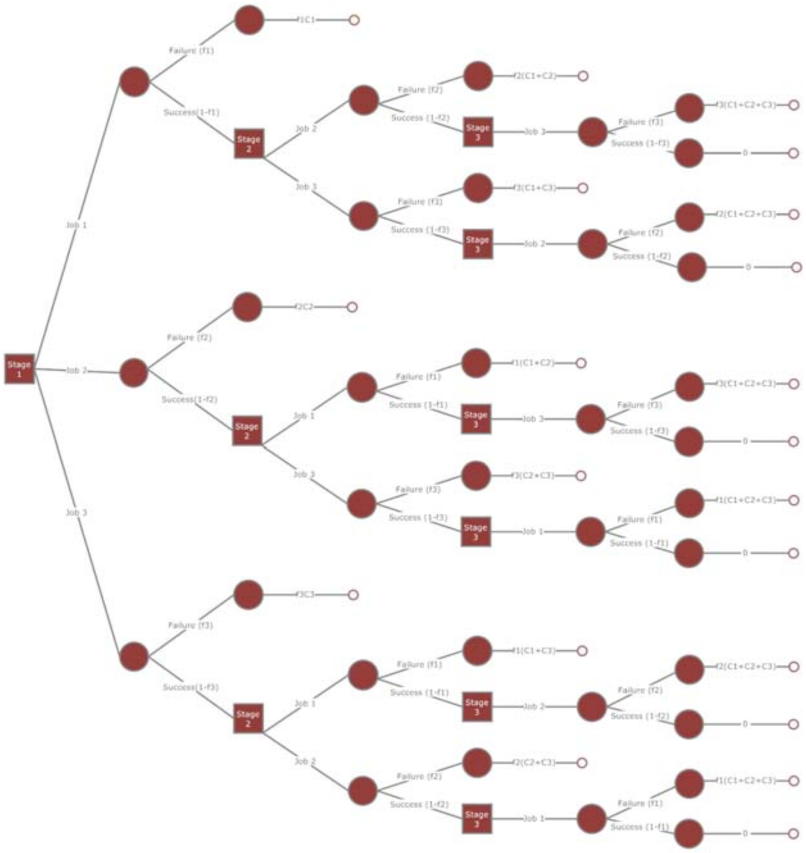
Considering the random failure rate as probabilistic element and using priory approach, we consider MSD as a single-machine sequence-dependent scheduling problem. Also, we assume the added value from each job is proportional to the job's processing time. As such, MSD can be reformulated as sequence-dependent scheduling model, where the objective function is to minimize the expected value of lost value-added under the constraint of each job having a probability of failure. Sequence dependent scheduling problems for a single machine (deterministic or probabilistic) with the objective to minimize the loss, usually involve setup times, known as the traveling salesman problem (TSP) and described in the famous paper by Gilmore and Gomory (1964). More work in the sequence dependent processing times is provided by Bianco et al. (1988), Tang (1990), and Wittrock (1990). Beside setup-times, job sequencing in a single machine becomes sequence dependent when due dates and tardiness are considered (Lawler, 1977; Lin and Ying, 2007; and Szwarc, 2007). Most recently, the research in single machine sequencing has focused on stochastic environments (Black et al. 2005; Xia et al. 2008) and dual or multiple criteria sequence dependent setup time scheduling (Lee and Asllani, 2004).

The MSD problem is quite different than the typical sequence dependent setup time. The thrust of the MSD is that the value at risk for any given job depends on the failure rate for that job and processing time of the job. Contrary to TSP where the setup time for a job depends only in the direct previous job, in the MSD problem the setup cost depends on the complete list of jobs already sequenced in the sequence. Dosa and He (2006) consider the scheduling problem with machine cost and rejection penalties and provide an online algorithm which seeks to minimize the sum of the makespan, the cost for purchasing machines, and the total penalty of all rejected jobs. Also, Cheng and Sun (2008) use a dynamic programming approach to minimize the makespan of scheduled jobs plus the total rejection penalty and the total completion time of scheduled jobs plus the total rejection penalty. Because there has been little previous academic work into this problem and because the cost of improperly sequencing these jobs can be high, there exists a need to search for a solution for the MSD. This paper will provide several initial steps toward this goal.

DECISION TREE ANALYSES FOR THE MSD

The problem addressed in this paper assumes that each job is 1) independent, 2) of constant value and 3) has a known and invariant risk of failure. Independence means that any machining job can be sequenced as desired and does not depend upon completion of previous work. This assumption allows a complete combinatorial solution without restriction. As discussed in the conclusion section, this assumption is liberal and frequently violated in real situations. It allows for the most difficult case with regard to a mathematical solution, however. The assumption of constant value describes a case each operation is valued at a cost that is known. This assumption is completely met where work is paid for on a contract basis. The final assumption of invariant risk of failure assumes the process has a well defined history in which failure rates can be determined.

Figure 1: Decision Tree for the MSD with Three Jobs



A decision tree is one of the most systematic tools of decision-making theory and practice. Since the MSD is a complex multistage decision problem, the decision trees can be helpful to understand this sequencing problem and provide an optimal solution. The MSD is a good candidate for the decision tree because the machinist needs to take into account the choices of jobs made at earlier stages of production as well as the possible outcomes of decisions at later stages of the sequence.

Figure 1 shows a decision making tree for the MSD with three jobs to be sequenced. Squares represent decisions the machinist can make. The lines that come out of the first square indicate that the machinist in the first stage must make a decision to process job 1, 2, or 3. Circles show various circumstances that have uncertain outcomes.

Once job 1 is selected, the lines that come out of the circle in the upper left section of the tree denote possible outcomes: f1 chance that job will fail and the value at risk is C1 or (1-f1) chance that the job will not fail, and the machinist has to make a decision (square) to process job 2 or 3. Each path that can be followed along the decision tree, from left to right, leads to a specific outcome.

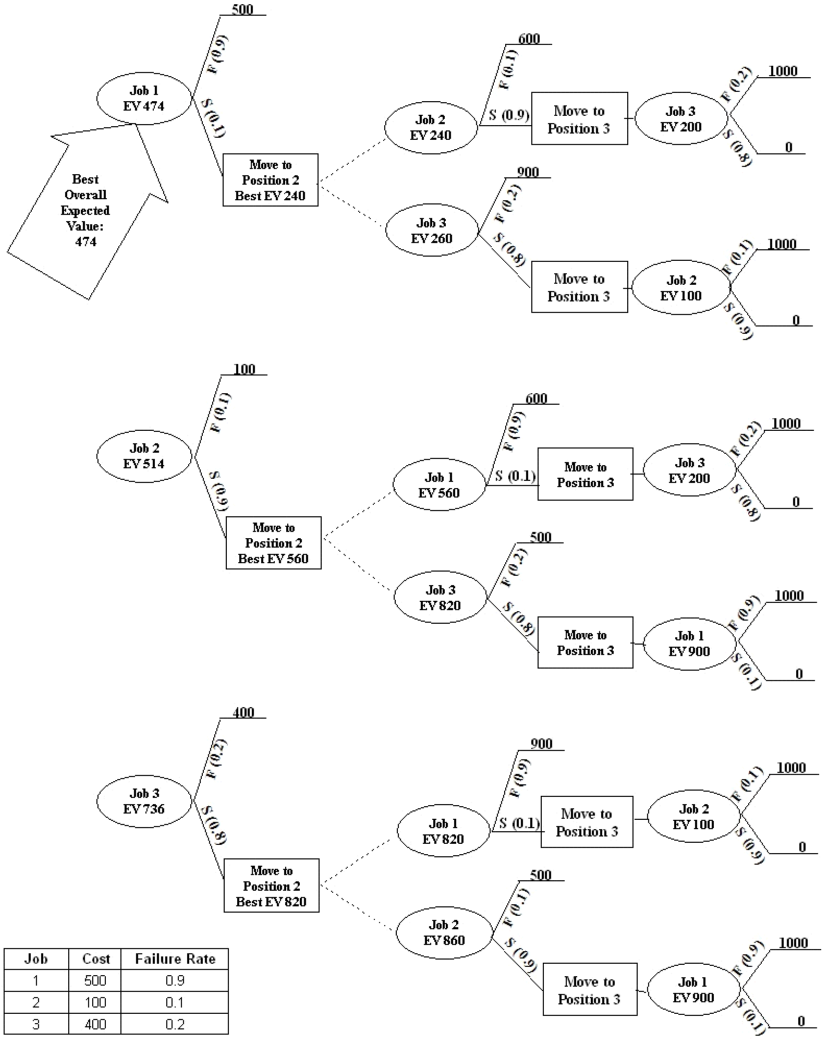
Once the decision tree is completed, the machinist can generate specific recommendations on what would be the best choice. For each square, the expected value of the choice can be calculated starting at the latest stages first. Calculations for a numerical example with three jobs are shown in Figure 2.

As shown, the machinist will start at the third stage where the value at risk for all branches is either 1000 (500+100+400) if the last job fails or zero if the last job and all the preceding jobs do not fail. The probability that job 3 will fail is .2, which lead to an expected value of 200 (.2*1000). Staying in the same branch, when job 2 is selected, there is a .1 chance of failure and .9 chance of success. The values at risk are respectively 600 (500+100) and 200, as explained earlier. This leads to an expected value of the node equal to 240 (.1*600 + .9*200). When compared with the similarly calculated value for the other branch of 260, the machinist would select the path with the minimum expected value of 240.

Again, remaining in the same branch, when job 1 is selected, there is a .9 chance of failure and .1 chance of success. The values at risk are respectively 500 and 240, as previously explained. This leads to an expected value of the node equal to 474 (.9*500 + .1*240). This is the smallest value when compared with the similarly calculated value for the other two branches of 514 and 736. As such, the optimal solution for the machinist will be to first select job 1, then job 2, and finally job 3.

Decision tree analyses can be used as an effective tool to find an optimal solution to the MSD a-priori scheduling problem. However, the tree can become very complex when the number of jobs under consideration increases. Even the use of previously created templates is not a practical option, because the number of jobs in various scenarios may vary and templates are practical only when the number of jobs to be sequenced is already known prior to the solution process.

Figure 2: Solving the MSD with Decision Tree



Job	Cost	Failure Rate
1	500	0.9
2	100	0.1
3	400	0.2

MATHEMATICAL PROGRAMMING FOR THE MSD

In this section, we first provide a 0-1 mixed integer non-linear programming formulation of the MSD. Notations for the non-linear model are:

j $= 1, \ldots , n$ Job index used as a unique identifier for each job;
k $= 1, \ldots, n$ Jjob index used to identify the position of a job in a given sequence;

p_j = Processing time for job j;
f_j = Probability of failure for job j;
x_{jk} = 1 if job j is assigned to the k-th position in the sequence;
0 otherwise;
$C(k)$ = Completion time for the job in the k-th position in the sequence;
$P(k)$ = Processing time for the job in the k-th position in the sequence;
$f(k)$ = Probability of failure for the job in the k-th position in the sequence; and
$E(k)$ = Expected value of failure for accumulated makespan in the k-the position in the sequence.

Using the above notations, a 0-1 mixed integer non-linear formulation is presented:

$$\textit{Minimize} \quad Z = E(1) \tag{1}$$

subject to

$$\sum_{j=1}^{n} x_{jk} = 1 \qquad \text{k = 1, …, n} \tag{2}$$

$$\sum_{k=1}^{n} x_{jk} = 1 \qquad j = 1, ..., n \tag{3}$$

$$P(k) = \sum_{j=1}^{n} x_{jk} p_j \qquad k = 1, ..., n \tag{4}$$

$$C(k) = \sum_{s=1}^{k} P(s) \qquad k = 1, ..., n \tag{5}$$

$$f(k) = \sum_{j=1}^{n} x_{jk} f_j \qquad k = 1, ..., n \tag{6}$$

$$E(n) = f(n)C(n) \tag{7}$$

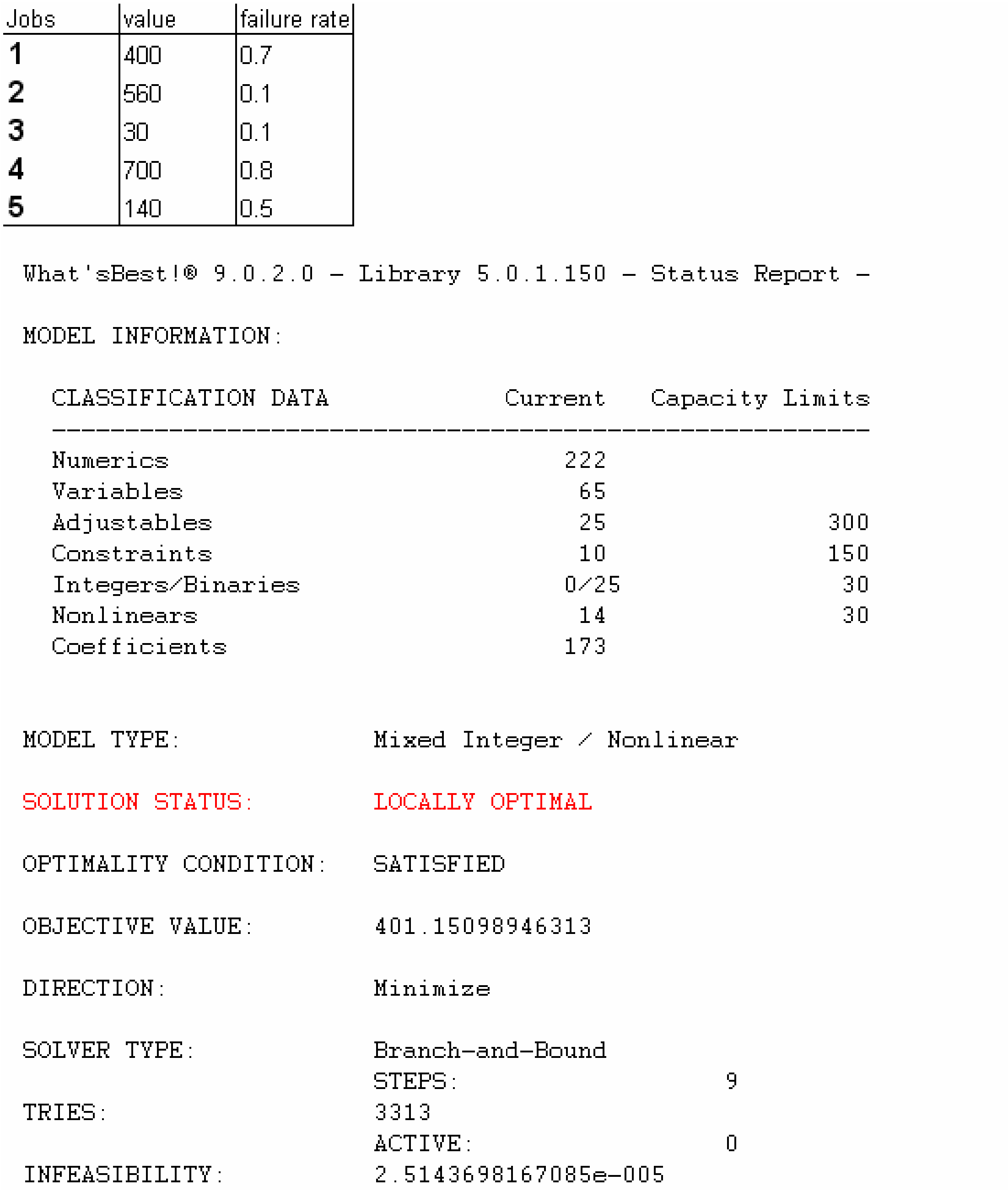
$$E(k-1) = f(k-1)C(k-1) + [1-f(k-1)]E(k) \qquad k=2,...,n \tag{8}$$

$$j = 1, \ldots, n \quad k = 1, \ldots, n \qquad (9)$$

and

$$C(k), P(k), f(k), E(k) \qquad k = 1, \ldots, n \qquad (10)$$

Figure 3: Using Excel to Solve the MSD with five jobs



Jobs	value	failure rate
1	400	0.7
2	560	0.1
3	30	0.1
4	700	0.8
5	140	0.5

```
What'sBest!® 9.0.2.0 - Library 5.0.1.150 - Status Report -

MODEL INFORMATION:

  CLASSIFICATION DATA            Current   Capacity Limits
  --------------------------------------------------------
  Numerics                          222
  Variables                          65
  Adjustables                        25                300
  Constraints                        10                150
  Integers/Binaries                0/25                 30
  Nonlinears                         14                 30
  Coefficients                      173


MODEL TYPE:               Mixed Integer / Nonlinear

SOLUTION STATUS:          LOCALLY OPTIMAL

OPTIMALITY CONDITION:     SATISFIED

OBJECTIVE VALUE:          401.15098946313

DIRECTION:                Minimize

SOLVER TYPE:              Branch-and-Bound
                          STEPS:                  9
TRIES:                    3313
                          ACTIVE:                 0
INFEASIBILITY:            2.5143698167085e-005
```

. . . .

Equation (1) is the objective function. The goal here is to minimize the expected value of failure in the first position. Note that this is a recursive function which includes all the decisions made in all positions. Equations (2) and (3) assure that only one job is assigned to each sequence position and a job is assigned to only one sequence location, respectively. Equation (4) determines the processing time, and equation (5) determines the accumulated value at risk of the *k*-th position in the sequence. Equation (6) identifies the probability of failure of the job in the *k*-th sequence position. Equations (7) and (8) recursively identify the expected value of failure for all positions 1 through *k*. Finally, equation (9) and (10) represent the integrality and non-negativity constraints.

As shown in Figure 3, we formulated and solved an MSD problem with 5-jobs. We used What'sBest!® 9.0.2.0, an Excel based package designed to solve mathematical programming models. If *n* is the number of jobs to be sequenced, our model has presented in the previous section has n^2+4n variables. When the number of jobs increases, the number size of the model with respect to decision variables will increase significantly. For example, a five-job scheduling problem requires 45 decision variables and a 10-job scheduling problem requires 140 decision variables. When more than 20 jobs are considered, the number of decision variables and constraints can become well above 500. Further, equations (7) and (8) make this a non-linear programming model and the complexity of the problem increases. Such large models make the implementation of the proposed mathematical model very time consuming and impractical.

However, since the solution provided by such models is optimal, operations schedulers should consider the use of such models when the number of jobs is relatively small. Scheduling complexity in such cases can also be avoided by preparing a user-friendly interface for the purpose of data entering and solution interpretations

GENETIC ALGORITHM

The general conceptual design of the genetic algorithm that we propose is based on the guidelines provided by Houpt and Houpt (1998). Specific details, such as, the design of the crossover and mutation operators are based on the work of Lee and Asllani (2004). Genetic algorithms, when applied to scheduling, view sequences or schedules as individuals, which are members of a population. Each individual is characterized by its fitness value. The fitness of an individual is measured by the associated value of the objective function. The proposed algorithm is shown in Figure 4. The proposed genetic algorithm consists of several steps as seen below.

Step 1: Create Population

* *Define Job Structure*. Each job consists of several members, such as job name, processing time, and probability of failure.

* Define Sequence Structure. Each sequence consists of an N-dimensional array of job structures

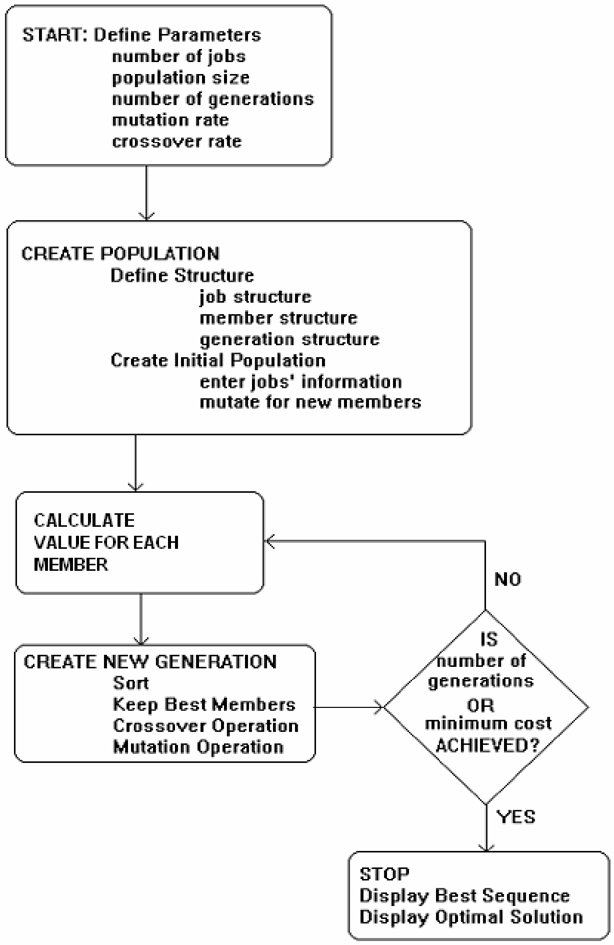
* *Define Population Structure*. Each generation consist of a vector of sequences. For the first generation, applying a mutation operator to the initial sequence creates the initial population. (See details for the mutation operator in Step 3).

Step 2: Evaluate Cost

The fitness value (or cost) of each sequence is used as an optimization objective function for the algorithm. This value is equal to *E(1)*, which is calculated by induction using:

$$E(k-1) = f(k-1)C(k-1) + [1-f(k-1)]E(k) \text{ for } k=2,\ldots,n \quad \text{where} \quad E(n) = f(n)C(n)$$

Figure 4: Algorithm for Genetic Programming



Step 3: Create New Generation

After sorting the population members of the previous generation in an ascending order based on the fitness value, the process of creating a new generation consist of three main steps:

* *Keep best members*. The process of assigning the first few best members from the old generation to the new generation will ensure a gradual improvement of the solution. The algorithm also saves the best sequence as a candidate optimal solution.
* *Crossover operator*. The following crossover operator code is suggested. This operator generates new sequences and ensures the feasibility of the algorithm.

-- Step a: *Select sequences PARENT 1 and PARENT 2 as two sequences from the old generation.*

-- Step b: *Generate k as a random number between 0 and N, where N is number of jobs in the sequence.*

-- Step c: *Select the first k members of PARENT 1 and save them in the new OFFSPRING.*

-- Step d: *Complete the rest (N-k) members of the OFFSPRING by following the following rules: If the rest of the members from PARENT 1 appear in the MOTHER sequence, then add the appearing member to the OFFSPRING following the same order they appear in the PARENT 2 sequence.*

* *Mutation operator*. The crossover operator is focused on creating alternative solutions around the best solutions achieved so far. In order to avoid the risk of remaining in the local optima, a mutation operator is suggested. For sequencing problems, mutation can be achieved by swapping two random jobs in a given sequence. The algorithm for this process consists of the following steps:

-- Step a: *Randomly generate two integers, k and s, between 0 and N, where N is number of jobs in the sequence.*

-- Step b: *Swap jobs that are in the k-th and s-th position in a given SEQUENCE*

The process of creating new generations will continue until a given number of generations is achieved or the cost of a given solution achieves an acceptable level. Lee and Asllani (2004) describe a list of the main parameters that impact the efficiency of their algorithm. Since the algorithm we propose in this paper is very similar and only simpler (does not consider any setups between jobs), we conclude that GA is an efficient tool to solve the MSD problem. For example, contrary to the mathematical formulation, we can conclude that the number of jobs does not affect the performance of the algorithm. Also, the algorithm shows significantly better performance when the population size selected by the programmer increases over the same number of jobs. The same can be concluded for crossover rate, the algorithm performs better when the number of members created through the crossover algorithm in each generation is increased.

CONCLUSIONS

This paper provides three alternative solution methodologies for an NP hard scheduling problem: Machinist's Sequencing Dilemma. Operation schedulers can use either a decision tree, 0-1 mixed integer-programming model, or a genetic search algorithm to solve such a problem. Decision tree is simple and easily to visualize, mathematical programming always ensures an optimal solution, and genetic-based algorithm does not always guarantee an optimal solution. On the other hand, the number of jobs that need to be sequenced does not affect the genetic algorithm performance. In contrast, decision tree and mathematical programming model becomes very complex and even unmanageable when the number of jobs is increased. We provide the following practical recommendations that can be used to select one of the above three solution methods. Decision tree can only be used if the number of jobs to be sequenced is small, such as four or less. Mathematical programming can be used when achieving an optimal solution is important and when there are more than four jobs to be sequenced. For a more practical solution and when the number of jobs is large (more than 10), we suggest using the proposed genetic algorithm.

The problem addressed in this paper assumes that each job is 1) independent, 2) of constant cost and 3) of constant risk of failure. These assumptions allow for the computation of expected loss at any point in the solution by multiplying all previous values added by the probability of loss for the k^{th} job. These assumptions may be violated in certain situations, however.

The first assumption of independence is easily violated by the physical properties of the piece of work. Our problem assumes any process can be placed at any point in sequence. Often one process must precede another. For example, a hydraulic piston rod must be lathed before fastening threads can be cut. Therefore regardless of the value added and risk of failure of the processes, the lathing must precede the thread cutting. Our problem also assumes that value added calculations use a measure of total time to complete a process. In fact, a lengthy individual process may be botched in the first few seconds of work rather than at inspection when all work is completed. Finally, risk of failure may be a function of the number of previous processes completed. Damage to an already machined section of a piece of work may be more likely if adjacent places on the work have already been machined and are susceptible to being ruined.

The final area for research is to develop an algorithm that allows a scheduler to determine either the first process or the final process to be scheduled independently of all non-optimal solutions. Our method presented in this paper requires all possible solutions be considered before an optimal solution can be determined. For processes with a small number of operations, this approach is feasible and not time consuming. For determining the combinatorials of this type of problem requires the calculation of a factorial; that is, the possible number of sequences of K processes is K!. The natural break point of effort for the decision tree solution presented in this paper is 5 or perhaps 6 operations (5! = 120 combinations; 6! = 720 combinations). For pieces of work requiring 20 machining processes, the possible number of sequences (20!) equals 2.43 X 10^{16} and is not feasible using our methods. If, on the other hand, an algorithm can be found to find the first process, that process can be removed from the alternative list of choices and the algorithm can be repeated on the remaining choice continuing until a final solution is determined.

REFERENCES

Bellalouna, M., & Jaillet, P. (2007). A Priori Parallel Machines Scheduling. Retrieved November 2007, from http://web.mit.edu/jaillet/www/general/sche220506-mb-pj.pdf

Bianchi, L., Gambardella L.M., & Dorigo M. (2002). An ant colony optimization approach to the probabilistic traveling salesman problem. *Lecture Notes in Computer Science*, 2439, 883—892.

Bianco, L., Ricciardelli, S., Rinaldi, G., & Sassano, A. (1988). Scheduling Tasks with Sequence Dependent Processing Times. *Naval Research Logistics*, 35, 177-184.

Black, G. W., McKay, K. N., & Morton, T. E. (2006). Aversion Scheduling in the Presence of Risky Jobs. *European Journal of Operational Research*, 175, 338-361.

Cai, X., & Zhou, X. (2005). Single-Machine Scheduling with Exponential Processing Times and General Stochastic Cost Functions. *Journal of Global Optimization*, 31, N. 2, 317-332.

Cheng, Y., & Sun, S. (2008) Scheduling Deteriorating Jobs with Rejection on Dominant Machines. *Journal Journal of Shanghai University*, 12 (6), 471-474.

Conway, R. W., Maxwell, W. L., & Miller, L. W. (2003). *Theory of Scheduling*. Courier Dover Publications.

Dosa, G., & He, Y. (2006) Scheduling with Machine Cost and Rejection. *Journal of Combinatorial Optimization*, 12 (4), 337-350.

Gilmore, P.C. & Gomory, R. E. (1964). Sequencing a One-State Variable Machine: A Solvable Case of the Travelling Salesman Problem, *Operations Research*, 12, 655-679.

Graham R. L., Lawler E. L., Lenstra J. K., & Rinnooy Kan A. H. G. (1979). Optimization and approximation in deterministic sequencing and scheduling: A survey. *Annals of Discrete Mathematics*, 5:287—326.

Hsieh, J., Chang, P. & Chen, S. (2006). Genetic Local Search Algorithms for Single Machine Scheduling Problems with Release Time, In Book Series: *Knowledge Enterprise: Intelligent Strategies in Product Design, Manufacturing, and Management*, 207/2006, 875-880.

Houpt, R. L, & Houpt, S. E. (1998). *Practical Genetic Algorithms*. John Wiley and Sons, New York, NY.

Jaillet, P. (1988). A priori solution of a traveling salesman problem in which a random subset of the customers are visited. *Operations Research*, 36, 929—936.

Lawler E. (1983). Recent results in the theory of machine scheduling. In A. Bachem, M. Grotchel, & B. Korte, (Ed.), *Mathematical Programming: The State of the Art*. Springer, Berlin.

Lawler E., Lenstra J., & Rinnooy Kan A. (1982). Recent developments in deterministic sequencing and scheduling. In M. Dempster, J. Lenstra, & A. Rinnooy Kan, (Ed.), *Deterministic and Stochastic Scheduling*. Reidel, Dordrecht.

Lawler E.L., Lenstra J.K., Rinnooy Kan A.H.G., & Shmoys D.B. (1993). Sequencing and scheduling: Algorithms and complexity. In S.C. Graves, A.H.G. Rinnooy Kan, & P.H. Zipkin, (Ed.), *Handbooks in Operations Research and Management Science*, Vol 4, (pp 445—522). North Holland.

Lawler, E. L. (1977). A Pseudopolynomial Algorithm for Sequencing Jobs to Minimize Total Tardiness. *Annals of Discrete Mathematics*, 1, 331-342.

Lee, S. M. & Asllani, A. (2004). Job scheduling with dual criteria and sequence-dependent setups: mathematical versus genetic programming. *Omega*, 32, N. 2, 145-153.

Lenstra J. & Rinnooy, Kan A. (1985). Sequencing and scheduling. In J. Lenstra and A. Rinnooy Kan, (Ed.), *Combinatorial Optimization: Annotated Bibliographies*. Wiley, Chichester.

Lin, S. W., & Ying, K. C. (2007). Solving single-machine total weighted tardiness problems with sequence-dependent setup times by meta-heuristics. *The International Journal of Advanced Manufacturing Technology*, 34, No. 11-12, 1183-1190.

Pinedo M. (1995). *Scheduling: Theory, Algorithms and Systems*. Prentice Hall.

Soroush, H. M. (1996). Optimal Sequences in Stochastic Single Machine Shops. *Computers and Operational Research,* 23, No. 7, 705-721.

Szwarc, W. (2007). Some Remarks on the Decomposition Properties of the Single Machine Total Tardiness Problem. *European Journal of Operational Research*, 177, 623-625.

Tang, C. S. (1990). Scheduling Batches on Parallel Machines with Major and Minor Setups. *European Journal of Operational Research*, 46, 28-37.

Wittrock, R. J. (1990). Scheduling Parallel Machines with Major and Minor Setup Times. *International Journal of Flexible Manufacturing Systems*, 2, 329-341.

Xia, Y., Chen, B., & Yue, J. (2008). Job sequencing and due date assignments in a single machine shop with uncertain processing times. *European Journal of Operational Research*, 184, 63-75.

This is a combined edition containing both Volume 12, Number 1, and Volume 12, Number 2

Articles for Volume 12, Number 2

INSTANT MESSENGER COMMUNICATION IN A MULTINATIONAL CORPORATION

Stephen Hunt, Sam Houston State University

ABSTRACT

This research investigates contemporary business communication practices regarding instant messaging. The research will explore the strengths, weaknesses and best management practices of instant messaging in a small multinational company in the Oil & Gas industry.

The research methodology includes employee interviews, employee surveys, and direct observation. The interviews provided guidance for the development of the employee surveys.

The findings are expected to show that instant messaging is a highly utilized communication method in business and has become essential to efficient business operations. The conclusions from this research will equip managers to minimize the risks and maximize the benefits of instant messaging communication in their organizations.

INTRODUCTION

In the 1990's software developers released a communication tool called instant messaging. This new communication method allowed users to communicate through a common instant messaging program. America Online was the first company to successfully attract a strong instant messaging customer base. The majority of AOL's customers were young and technologically savvy. These young consumers quickly made AOL IM a success.[1]

The first generation of IM consumers who flocked to these early IM programs has entered the workforce. They bring with them a level of technological comfort that is changing the face of modern business communication. Managers and business communication professionals should educate themselves about the impact instant messaging technology will have on organizations.

PURPOSE

The purpose of this research is to identify strengths, weaknesses, and the way employees use instant messaging to support daily operations. The research also seeks to determine which subsections of employees are using instant messaging communication most frequently.

RESEARCH METHODS

A survey and several interviews were conducted to collect data regarding instant messaging use in a small international oil trading company, Mabanaft, Inc., based in Houston, Texas. Senior managers at Mabanaft authorized a survey of employees in North America, Europe, and Asia. The survey participants included all Mabanaft employees from these regions. Twenty-seven surveys were completed and returned, for a 60 per cent response rate, which provided very strong data for analysis.

The survey consisted of ten multiple choice questions. The questions focused on employee use of instant messaging systems. Also, several demographic questions were used to provide categories for further analysis. Below are the questions from the survey:

Employee Use of Instant Messaging Systems

Q1. Rank the following communication methods based on how frequently you use each during an average workday, 1 being the most, 4 being the least:

Telephone
Email
Instant Messaging
Personal Contact

Q2. What percentage of your total business communication is through instant messaging?

0 percent to 20 percent
21 percent to 40 percent
41 percent to 60 percent
61 percent to 80 percent
81 percent to 100 percent

Q3. Please rate the following statements by selecting strongly agree, agree, neutral, disagree or strongly disagree

Instant messaging is convenient to use while working from home.
Instant messaging is easy to access while traveling.
I feel that instant messaging is secure.
My preferred method of communicating at work is instant messaging.

Q4. Please rate the following statements by selecting always, frequently, sometimes, rarely or never.

Instant messaging is useful to communicate with colleagues in other companies.
I use other communication methods to supplement instant messenger conversations.
I communicate with colleagues in other countries with instant messaging.

Q5. Please rate the following statements by selecting strongly agree, agree, neutral, disagree or strongly disagree. I believe instant messaging is:"

Impersonal
Unreliable
Unprofessional
Efficient

Employee Use of Instant Messaging Systems

Q6. Please rate the following as strongly agree, agree, neutral, disagree or strongly disagree.

Managers implementing a new instant messaging system in their organization should:
Monitor all incoming and outgoing messages.
Save all instant messenger communication.
Use a messenger program designed for business use.
Use a free messenger program, such as Yahoo or MSN.
Limit Instant Messaging to internal communication

Q7. I work for:

Mabanaft Inc.
Mabanaft BV
Mabanaft PTE
Other (please specify)

Q8. My home office is located in:

North America
Europe
Africa
Asia
South America

Q9. I currently work in the following area:

Finance
Accounting
Operations
Trading
Marketing
Management
Other

Q10. My current age is between:

20 and 30
31 and 40
41 and 50
51 and 60
61 and 70

RESULTS

The Mabanaft survey provided several interesting pieces of information. The most significant findings are listed below.

♦ When asked how frequently they use instant messaging to communicate with colleagues in other countries, 23.1 per cent of Mabanaft employees selected “Always” and 42.3 per cent chose “Frequently”.

- 65.4 per cent of Mabanaft employees responded that they communicate with other organizations through instant messaging at least frequently.
- The survey revealed demographic information for employees who use instant messaging for more than 40per cent of their daily business communication. Forty-five percent of North American employees use instant messaging at this rate, compared to 27.3 percent of Europeans and Asian colleagues.
- The survey revealed that 91 per cent of Mabanaft employees who use instant messaging for more than 40 per cent of their daily business communication are between the ages of twenty and forty.

DISCUSSION

The most significant findings from the survey relate to international business communication, not only the use of instant messaging among employees but also between employees and external stakeholders. These results provide a preview of organizational communication changes for which managers must prepare.

International organizations incur significant costs to communicate effectively between international locations. Instant messaging can help managers significantly reduce these costs and improve communicative efficiency in their organizations. The regular use of instant messaging communication allows inexpensive and efficient communication among the international Mabanaft offices.

Another interesting result from the survey was the significant use of instant messaging communication with external business partners. The nature of a trading business requires fast communication between brokers, buyers, and sellers. This study found that traders use instant messaging to contact brokers, buyers, and sellers. Schedulers receive pipeline movement information from pipeline companies and other schedulers. The operations department uses instant messaging to stay updated on vessel movements. Instant messaging is the fastest and most efficient method of communicating between all the organizations involved in moving Mabanaft's products.

Finally, the survey indicates younger members of the workforce are more likely to utilize instant messaging for business communication. The data shows a tendency for employees less than 40 years of age to use instant messaging for 40 per cent of their daily business communication. Managers must take notice of this trend and should develop systems to manage instant messaging communication.

In 2004, the US Air Force decided to add secure instant messaging portals to their communications network. The Air Force made this decision because military personnel were already using AOL and Yahoo IM to conduct daily military affairs.[2] Rather than ban an effective communication method, the Air Force took steps to effectively manage the grassroots movement toward IM communication. This trend will likely intensify as more members of the generation that made instant messaging a booming business enter the job market.

The following list of strategies emerged from this study. They are a starting point for managers interested in preparing their organization for business communication in the future.

DO

1. Archive all incoming and outgoing messages.
2. Have a clear instant messaging policy, even if the company does not officially use instant messaging.
3. Educate employees about the potential legal liability they expose the company to through instant messaging.

DON'T

1. Use instant messaging for highly sensitive communication.
2. Ban instant messaging without serious consideration. Employees have and will continue to use it without authorization.
3. Wait to develop systems for instant messaging until they become a problem.

CONCLUSION

Clearly there is great potential for growth in instant messaging. Many companies are already utilizing instant messaging in daily business operations. Our survey of Mabanaft employees showed that 48.1 per cent either agree or strongly agree that their preferred method of communication is instant messaging. IBM employees send an average of 6 million internal instant messages on a daily basis. In 2006, they were sending 4 million internal instant messages daily. [3] Managers cannot ignore the growth of instant messaging communication.

As the generation of consumers that made Yahoo, MSN and AOL household names enters the workforce, instant messaging usage will increase. Our research at Mabanaft showed that instant messaging accounts for over 40 per cent of the daily business communication for 55.5 per cent of Mabanaft employees between 20 and 30 years of age. By comparison, 35.3 per cent of Mabanaft employees between 31 and 70 years of age use instant messaging for 40 per cent of daily communication. This trend can be seen in other organizations around the world. Junior members of the labor force are bringing instant messaging programs into organizations around the world.

Billions of dollars in commodities, stocks, bonds and commercial goods are traded every day using instant messaging technology.[4] The strong international communication, easy message archiving, high level of efficiency and easy implementation will lead more organizations to use instant messaging. The productivity and cost reduction benefits of instant messaging far outweigh the potential negatives to businesses.

ENDNOTES

[1] Jeff Tyson and Alison Cooper, "How Instant Messaging Works," *How Stuff Works,* http://communication.howstuffworks.com/instant-messaging1.htm

[2] William Jackson, "Air Force abuzz over instant messaging app," *Government Computer News,,* http://gcn.com/articles/2004/06/06/air-force-abuzz-over-instant-messaging-app.aspx

[3] Robert McGarvery, "Instant Messaging comes to the office," *Executive Travel,* http://www.executivetravelmagazine.com/page/Instant+Messaging+comes+to+the+office?t=anon

[4] Kim Kyoungwha and Nesa Subrahmaniyan, "An Unlikely Trading Hub in Asia," *International Herald Tribune,* http://www.iht.com/articles/2005/12/01/bloomberg/sxtrades.php

REFERENCES

Anthony Reuben, "What is it like being an oil trader?" *BBC News,* http://news.bbc.co.uk/2/hi/business/7250554.stm

Nancy Flynn, *Instant Messaging Rules*, (New York: AMACOM, 2004), 145-155.

UNDERSTANDING THE LACK OF MINORITY REPRESENTATION IN GRADUATE PROGRAMS IN COMPUTER SCIENCE AND INFORMATION TECHNOLOGY: A FOCUS GROUP STUDY OF STUDENT PERCEPTIONS

Sharad K. Maheshwari, Hampton University
Anne L. Pierce, Hampton University
Enrique G. Zapatero, Norfolk State University

ABSTRACT

This paper is an effort to delineate factors impacting lack of representation of minority students at the graduate level education in information technology fields: computer science and computer information systems. The research was conducted in three Virginia institutions: Hampton University (HU), Norfolk State University (NSU), and Virginia State University (VSU). The paper examined basic factors impeding interest of undergraduate computer science and information technology students in graduate education. Based on our findings a few strategies are suggested which could possibly lead to higher interest, hence, better recruitment and retention of minority students in graduate programs in these fields.

The research shows that students' lack of interest in graduate education was due to four basic factors. These factors were 1: lack of information about graduate school and admission process to the graduate programs, 2: perceived value of graduate education, 3: financial considerations and 4: perceived educational preparedness. It was also found that undergraduate school supervisors (teachers, advisors, and administrators), family and friends have a direct impact on the students' intention and interest in graduate education. Furthermore, students' interest level decreases as they move from underclassman status to upperclassman status. Based on these findings, a few basic recruitment strategies are proposed.

Keywords: Graduate Minority Students, Graduate Computer Science Education, Recruitment of Graduate Minority Students, Focus Group Study, Computer Science Students in HBCUs.

INTRODUCTION AND LITERATURE REVIEW

There is a growing concern over the under representation of women and minorities in the natural sciences and engineering fields, including computer science. There is a large body of research material which documents this fact. The focus of this research was to look beyond undergraduate education and to investigate the lack of representation of minorities in graduate and post graduate education in the field of computer science/information technology. Some of the relevant research is included here in the following section.

Grandy (1994) conducted a study among college seniors who registered to take the Graduate Record Examination (GRE) test and who were majoring in natural sciences, mathematics, computer sciences, and engineering (NSME.) A stratified sample of 1,651 such college students was collected. The goals of the survey were to identify some of the factors that may lead NSME majors to change fields for graduate school, analyze differences among ethnic groups remaining in NSME, and analyze differences between male and female NSME majors who plan to remain in NSME. The research mainly focused on gender and ethnic differences in NSME majors planning graduate study in their fields. Results showed that the decision to leave NSME was uncorrelated with gender, race, or GRE scores. Detailed analysis of gender and ethnic differences among NSME majors planning to continue in their fields showed small to moderate differences on many dimensions. There were gender and ethnic differences in salary expectations, importance on making a contribution to society, and preferences for various job activities.

The under representation of women and minorities in information technology (IT) professions is also well documented (National Science Foundation [NSF], 2000). In fact, recent statistics show that the IT workforce is comprised of less than 30 percent female and less than five percent minority professionals (Council of Higher Education Accreditation [CHEA], 2000). The Computing Research Association survey on graduate students shows that, between 1993- 2003, African American enrollment in Ph.D. programs in computer science/computer engineering remained 1% or 2% of total Ph.D. enrollment in these majors (Vegso, 2005). Several recent research studies have been done to determine the reasons why such an employment gap exists despite the relatively high demand and attractive salaries for IT workers (Houston-Brown, 2002; Baylor, 2003), and many more studies have documented the underlying reasons for a similar gap that exists in science, math, and engineering professions in general (Landis, 1985; Cohoon, 1999; Thom, Pickering, & Thompson, 2002; and Armstrong & Thompson, 2003). In a recent publication, Cohoon, & Aspray (2006) reviewed the existing literature for the causes of the gender gap in the information technology field and possible strategies to rectify this problem. These studies point to the well documented “digital divide,” which limits minorities’ access to computing technology; inadequate K-12 preparation, especially in math and science; and a critical lack of counseling and mentoring as key reasons for lack of recruitment and retention of minority students in IT majors.

Gates, Teller, Bernat & Cabrera (1999) studied the affinity research group model which provides students with opportunities to learn, use, and integrate the knowledge and skills that are required for research with the knowledge and skills that are required for cooperative work. Membership in affinity groups is dynamic, i.e., old members graduate and new members join in; and students come to the groups with different levels of knowledge and skills. Because of this, an annual orientation is needed for new members to facilitate their understanding of the philosophy and goals of the affinity model, understanding of the research goals of the projects to which they are assigned, learning of the basis of the cooperative paradigm, and awareness of group expectations. More importantly, the orientation develops new members' basic understanding of the research process and provides information about available resources. The orientation is also important for established members. It provides them with an opportunity to renew their commitment to the group, improve their research and cooperative group skills, and processes within the group with the goal of improving the group's effectiveness. The orientation also allows faculty mentors to become aware of members' misgivings and expectations of the affinity group experience. It also provides a chance to the faculty member to reevaluate the goal of the model and its success.

Eide & Waehrer (1998) examined the expectations of attendance in a graduate program and its payoffs affect in the selecting the undergraduate major. Results explain why some students choose to major in fields associated with poor job prospects for undergraduate degree holders. The option to attend graduate school is not a significant factor in choosing to major in computer science/engineering. Women are significantly less likely to select majors associated with higher future wages. The research effort is generally concentrated on undergraduate college education, hence, focus is recruitment and retention of K-12 students to science and technology related majors in college.

In addition, a number of research studies have identified “best practices” in programs that seek to address these gaps. Model programs at institutions as diverse as the California State University, Northridge (Landis, 1985), Case Western Reserve University (Boulding, 1985), Texas A&M (Graham, Caso, Rierson & Lee, 2002), Arizona State University (Fletcher, Newell, Anderson-Rowland, & Newton, 2001), the University of Maryland (Armstrong & Thompson, 2003), and the Oklahoma Alliance for Minority Participation in Science, Mathematics, Engineering, Technology and Education (Mitchell, 1997), as well as an ongoing studies at several institutions in different programs that are successful at recruiting and retaining women, have incorporated a number of strategies: summer bridge programs; academic enrichment activities; tutorial services; ongoing peer, faculty, and professional counseling and/or mentoring; and cooperative and internship experiences both on and off campus. However, while all of these studies offer valuable insights into the underlying reasons and possible solutions to the lack of minorities in the IT workforce, few of these programs specifically target African-American students, and most are broadly focused on science, math, engineering, and technology majors. There are also efforts made to educate teachers and counselors about causes of lack of representation of minorities and women in technology related fields (Nicholson, Hancock, & Dahlberg, 2007.) This research focuses on altering teachers’ and counselors’ perspectives that could bring change in the attitude of minority and female students towards technology related fields.

RESEARCH MOTIVATION

All of the above studies are pointing to one fact that there is a shortfall of minority students in the field of science and technology including computer science. This shortfall increases substantially as one starts to look beyond the undergraduate level to the masters and doctoral levels. It is, therefore, imperative to understand the underlying factors that impede the progress towards graduate education among minority students in the field of computer science/information technology. Furthermore, it is important to identify some strategies that encourage minority students to pursue graduate education in these fields.

This study is an attempt to understand the factors which hinder students from pursuing graduate and post graduate education in computer science related fields. Also an attempt is made to provide some possible paths to design strategies to reduce this shortfall. The study was conducted in three Historically Black Colleges and Universities (HBCUs) in Virginia that offer an undergraduate computer science program. The institutions involved were Hampton University (HU), Norfolk State University (NSU) and Virginia State University (VSU.)

DATA COLLECTION

Two different methods were used to collect the data. This study utilized focus groups in the first phase of data collection and a written survey in the second phase of data collection. In the first phase, two focus groups were conducted in each institution involved with the project in the fall and spring semesters of 2004-2005 academic year. It was not possible for logistical reasons to invite a random group of students for 30-40 minutes of discussion on graduate education. Focus groups consisted of junior or senior level classes in which 25-35 minutes of class time was devoted to an open ended discussion of graduate education: its need, its value, their perceptions, etc. The role of the faculty member was limited to ask a probing question, whenever, there was a general pause in the discussion. No audio or video recording was made. The faculty member took notes about the discussion. Total of six informal focus groups were conducted in three institutions. The participation in the focus group was voluntary; students were told the purpose of the research and were given the option to opt out of the discussion if they so desired. The results of these focus groups were summarized to identify the underlying themes.

A written survey was conducted in the second phase of data collection. The survey instrument (Appendix 1) was developed based on available research and the experience gathered from the focus groups. The main objective of the survey instrument was to determine the influence of family members, friends, teachers, mentors and student’s background on their interest and intention to attend graduate school. The survey was administered using clustered sampling technique among randomly selected junior and sophomore classes in all three participating institutions in the Spring 2005 semester. As with the focus group, students were given the option not to participate if they so desired. Results from the focus group and survey are discussed in the next section.

RESULTS AND ANALYSIS

Patterns within the focus group results were identified after tabulation of all major points from the three participating institutions. The results of the focus group indicated four major themes for lack of representation and interest among students in the graduate education in computer science/information systems. These major themes were: a) graduate school admission process and awareness, b) job and financial issues, c) perceived value attached to graduate education, and d) perceived level of preparedness.

Graduate School and Admission Process and Awareness

In all three institutions students repeatedly said that they were either not very aware of the process of graduate school admission or didn't consider the graduate school admission processes to be easy and straightforward. Students also mentioned a lack of understanding of availability of funding, and other options available to finance their graduate education. This was one area where more personal contacts and better mentoring/nurturing can help students to consider graduate school as an option while planning for career choices. A summary of the major points in this area is included below in Table 1.

Table 1: Summary of Focus Group Themes Related to the Admission Process

No.	Comments on the Admission Process Theme
1.	Applying to grad school is too long a process – application, GRE, letters, applying for funding, etc. GRE preparatory courses were attended only by 3 students. Some mentioned the hassle of application—long process, preparation for exams, GRE, etc. or Grad school application process is long. Involves separate preparation. Students say that they don't have that much time.
2.	Preparation is important, mostly on the strategies for taking the graduate admission tests. Graduate admission tests (both GMAT and GRE) are perceived as a major hurdle "…you know these tests are skewed against minorities…", "…we do not test well on these types of exams…", "…I heard that if you retake it, they still use the lowest score…", "…a friend of mine told me it was impossible to prepare for it [GMAT] since it [GMAT] does not test knowledge…", "…I was advised to take it 'cold' [without any preparation]…"
3.	Students in general are not aware of the number of opportunities for graduate school support. Lack of information about graduate school or students not being aware of graduate school support: RAs, TAs, LAs, etc.
4.	Perception of graduate school being very hard; it requires a lot of reading and research work. Undergraduate education is challenging enough, not ready for more academic work. Struggled at undergraduate and hence do not think attending graduate school is possible.
5.	Lack of research experience.
6.	Their undergraduate school does not offer graduate program.
7.	Affirmative Action set asides for graduate school admissions for minorities are gone "…see what has happened at the University of Michigan…", "…the current administration [federal government] has an agenda to destroy Affirmative Action, especially in higher education…"

Table 1: Summary of Focus Group Themes Related to the Admission Process	
No.	Comments on the Admission Process Theme
8.	Most graduate programs are in majority institutions. Some minority students feel they might be shunned or kept apart from the majority. Other schools have a lot more resources, which translated to some fear of other schools being harder or superior. Perceptions as of other schools are teaching more at the undergraduate level as these other schools have more resources to prepare their students better. Department's ranking was also an issue in the mind of a few students. Perception that they are not prepared adequately. Certain cutting edge courses are not taught or are not part of the curriculum in these schools (for example information security.)
9.	Difficult to connect with other schools in graduate program.

Job Market/Financial issues

Students also indicated that the cost of college (undergraduate education) was high and they wished to start earning money. Furthermore, they were very aware of the fact that the average salary in their discipline was higher than many other disciplines as well as the job market was very good. That is, they would find a job easily and they would be able to command a decent salary. This is an indication of a major obstacle for promoting graduate education to our students. Graduate schools have to compete directly with the job market in CS/IT area. In other words, computer science graduate programs have to provide more incentives including better financial support to attract students. A summary of major points in this theme area is provided in Table 2.

Table 2: Summary of Focus Group Themes Related to the Job Market	
No.	Comments on the Job Market Theme
1.	Going to graduate school in CS/IT is not as necessary to secure a good job compared with other majors in education or liberal arts. Salaries are good enough after undergraduate degree; why bother with more education. Postpone graduate education from the immediate future after undergraduate degree. Currently, main goal is to find a job and start making money and become financially independent. Main motivator is earning enough money to become independent – we need to ask about influence of parents here. Most of the time what you learn at undergraduate school is enough to get a job.
2.	"…the opportunity loss of delaying employment in lieu of higher education…" Loans are too much to pay after undergraduate or cannot afford more loans for graduate education. The concern about repaying student loans for undergraduate education while accumulating more debt for graduate school. .
3.	Some students are first-time family graduates from a college. Why bother to extend my education? Why wait to finish my education when there are jobs waiting for me?
4.	Learning can take place at the workplace; why go to grad school. You will learn more on the job. Experience is more important than academic work.
5.	Corporate world needs mix of business education and technical skills; MBA is more important for career than masters in CS/IT. Graduate degree in business is more valued in their mind as that is considered a ticket to moving up on the corporate ladder. "Will I get a job after master's degree or will I get a better job after masters? Master's degree may be an overkill of education."

Table 2: Summary of Focus Group Themes Related to the Job Market	
No.	Comments on the Job Market Theme
6.	Financing grad school is an issue. Want to go to grad school only if employer pays for it and can be done on a part time basis. “I don’t want to pay for the education; will consider grad school only if it is paid for.”

Perceived Value Attached to the Graduate Education

This was the most recurring theme. Students on several occasions mentioned that graduate school in computer science was not going to add much market value to their careers. Two main factors which were heavy on the students’ minds were salary potential and skill building potential of graduate education. The differential in salary with one or two years of experience compared with one or two years of advanced college education was not in favor of the latter. Second, students were clearly indicating that education other than a graduate or post graduate degree in computer science was more financial rewarding (e.g., MBA or professional/technical certifications.) Table 3 has a summary of major points in this area.

Table 3: Summary of Focus Group Themes Related to the Value of Graduate Education	
No.	Comments on the Value of Graduate Education Theme
1.	Students had a view that college education (at least higher level education) is not as important as obtaining certificates in the technical skills. Certificate programs are more valuable than grad school.
2.	Learned enough “why bother with more education?” “How will it help? Programming is an art so experience is more important.” One can teach a lot of programming and other technical things to oneself, once you have a baseline understanding. Why do you need grad school?
3.	“What is the motivation of going to grad school?” Want to work and make money. Is it worth the difference in salary vs. assistantships over the long term?
4.	Long term perspective is missing. Education is equated to what you can make, what you can afford to buy and buy it now.
5.	Not passionate about any further education. Field is too broad and is difficult to stay focused.

Perceived Level of Preparedness

Students also showed some lack of confidence. They indicated a fear of graduate education and research. They also indicated that they don’t see many role models at the graduate level. Lack of knowledge about graduate programs also was a factor here. It seems that students who were interested wanted more information along with some role models who could alleviate their fears about difficulty and rigor of graduate school education. A summary of major points in this area is included in Table 4.

Table 4: Summary of Focus Group Themes Related To The Level of Preparedness

No.	Comments on the Perceived Level of Preparedness Theme
1.	Department's ranking was also an issue in the mind of a few students. Students are not aware of the number of schools accredited by the same organization as theirs.
2.	There was a general lack of information about grad schools. This may be due to the fact that they are not considering graduate school or recruitment efforts for graduate schools are not very strong. "It is my opinion that we do not have an abundance of role models we could emulate." "We need to create a "graduate school" mentality among our students."
3.	Perception as if other schools with more resources are teaching better and preparing their students at higher level. Other schools have a lot more resources; some fear other schools are very hard/ superior.
4.	Students are not aware of the general education focus placed on the undergraduate programs versus the special knowledge developed at the graduate level.
5.	Lack of emphasis on research at the undergraduate level. Students do very little guided research. Most schools do not offer a research methods undergraduate course. The closest they get to experimental design is in a second course of statistics which is compressed with management science topics into a single course offering (DSC 376 -- Statistics and Quantitative Methods). Lack of research preparedness. This had to do with perceived (or perhaps very real!) quantitative skills weakness. This affects the students' performance in calculus, statistics, management science, operations management, and other quantitative courses which serve as a foundation to do research.
6.	GPA isn't high enough/Afraid of GREs. Difficult to connect with other schools in graduate program.
7.	Undergraduate is challenging enough/Graduate school is too challenging

At the next step survey analysis was performed. A total of 153 surveys were collected. 18 surveys were excluded from the analysis as nothing was filled on those surveys other than the college name and other demographic information. The statistical analysis was done using SPSS 12.0.1. Summary of demographic information is provided in Tables 5 through 10. Most of the survey participants (68%) were junior level students. There was no significant difference in the sample composition according to classification of the students from three institutions. Over 90% of the sample respondents were between 20-24 years old. The majority of students were from urban areas (58%) and there was no significant difference based on urban-rural mix among three schools under study. Gender distribution (approximately 50% male and 50% female) was very homogenous. 38% of students reported their GPAs between 2.51-3.50; approximately 28% reported GPAs below 2.50 and 29% reported above 3.51 GPA. GPA distribution among the participating schools was very similar as well, however, VSU's students reported slightly lower overall GPAs.

This indicated that samples from the three different schools under considerations were very similar according to age, gender, urban background and GPA. Hence, it can be assumed that the overall sample was homogenous for statistical purposes.

Table 5: Frequency Count by Schools

School Name	Type	Frequency
HU	Private	45
NSU	Public	46
VSU	Public	44
TOTAL		135

Table 6: Distribution of Classifications by Schools

School	Freshman	Sophomore	Junior	Senior
HU	3	13	28	1
NSU	1	13	32	0
VSU	1	7	32	4
TOTAL	5	33	92	5

Table 7: Distribution of Age by Schools

School	Under 20	20-22	23-24	25-26	over 26
HU	2	36	5	0	1
NSU	0	30	10	1	4
VSU	0	33	7	0	3
TOTAL	2	99	22	1	8

Table 8: Distribution of Urban-Rural Background by Schools

School	Urban	Rural
HU	29	14
NSU	22	18
VSU	28	14
TOTAL	79	46

Table 9: Distribution of Gender by Schools		
School	Male	Female
HU	27	16
NSU	18	26
VSU	23	20
TOTAL	68	62

Table 10: Distribution of GPA by Schools					
School	2.00-2.50	2.51-3.00	3.01-3.50	3.51-4.00	Over 4.00
HU	10	6	8	17	2
NSU	9	12	13	4	5
VSU	19	7	5	8	3
TOTAL	38	25	26	29	10

Earlier focus group gave us some indications that students showed a lack of confidence in their preparation for graduate school. To further investigate this, we added four questions in the survey about their interest and understanding of mathematics and pure sciences and their relationship with computer science and related fields. Approximately 55% of students responding saw no relevance of mathematics courses with computer science courses (Table 11.) 90% of the responding students saw no relevance between pure sciences and computer science/information technology courses (Table 12.) This was a very high proportion of students who saw foundation courses (mathematics and physical sciences) for computer sciences being irrelevant to the curriculum. Although it is also clear from this fact that mathematics and physical sciences were not contributing to their perceived lack of preparedness. It needs further investigation to better understand the causal relationship as to why student feel they are not well prepared for graduate level education with various curriculum components.

Table 11: Distribution of Relevance of Mathematics by Schools			
School	Yes	No	Don't Know
HU	21	22	1
NSU	16	28	0
VSU	18	24	0
TOTAL	55	74	1

Table 12: Distribution of Relevance of Pure Sciences by Schools			
School	Yes	No	Don't Know
HU	6	31	7
NSU	9	34	2
VSU	2	37	3
TOTAL	17	102	12

The survey asked several questions about family, advisor, friends and teachers to determine their influence on student's inclination to attend graduate school. The instrument had a set of questions that were designed where students could indicate that they at least would consider graduate school as an option (see question numbers 40-44 in Appendix I). For further analysis a composite variable 'interest in graduate level education' was created. This composite variable had three levels for students' interest in the graduate education in computer science/information technology: 'yes', 'no', and 'no preference.' The positive interest was indicated via five separate questions on the questionnaire (Appendix I: question numbers 40-44), definitive lack of interest for graduate school was based on a larger question set of 16 questions in the instrument (Appendix I: question numbers 23-38) and the rest of students were considered to have no preference. A summary of interest in the computer science/information technology graduate school is provided in Table 13. Approximately 33% of students showed some level of interest in graduate education.

Table 13: Interest in Computer Science Graduate Education		
Interest Grad School	Frequency	Percentage
No Preference	19	14.1%
No	71	52.6%
Yes	45	33.3%

The cross tabulation of the interest in the graduate education data with students' classification showed a different picture (Table 10.) There was a significant drop in the "interest in graduate education variable" from sophomore to junior level. The drop was even more significant as they reached the senior level; however, sample size for the seniors was very small.

One of our objectives was to establish whether interest in computer science/information technology graduate education was influenced by family members, friends, advisors, teachers and administrators. There were nine different questions on the instrument asking students to describe their parents', siblings', professors', advisors', administrators', and friends' attitudes towards graduate education. These questions gave respondents wide latitude: seven levels ranging from "never discussed" to "insisting upon going to graduate school." After tabulation of responses for the attitude towards computer science/information technology graduate education question for each group, a composite variable 'group member attitude toward graduate education' was created for each subgroup. The composite variable reduced the seven levels for attitude into two levels reflecting a positive attitude towards computer science/information technology graduate education and absence of a positive attitude towards computer science/information technology graduate education. Positive attitude was defined as the following response from the student about a group members' attitude: "think it would be good for me," "expect me to go" or "are insisting I go." If respondent selected any other choice for attitude towards computer science/information technology graduate education, then it was not considered positive. These choices included the answers "never discussed," "think I should not go," "leave it up to me" or "want me to go into a different field in graduate school."

Table 14: Distribution of Interest in Computer Science/Information Technology Graduate Education by Classification

Interest in Graduate School	Freshman	Sophomore	Junior	Senior
No Preference	60%	9.1%	14.1%	0.0%
Yes	20.0%	42.4%	31.5%	20.0%
No	20.0%	48.5%	54.3%	80.0%

To establish relationship between attitudes of different group members with students' interest in computer science/information technology graduate education, Pearson's bivariate correlation coefficients were calculated. Table 15 summarizes these correlation coefficients and shows the statistical significance (p-value) for a two-tailed test. Every peer group (siblings and friends) and every superior group (parents, and school officials) have statistically significant relationships with the students' interest in the computer science/information technology graduate education, except the college-graduate siblings. The result in the college graduate sibling subgroup may be influenced by the small sample size of that group (18% of surveyed students indicated having a college graduate sibling.)

It is important to note that a much higher degree of correlation exists between the interest in graduate education and positive attitude of professors, advisors and college administrators compared with any other group. It suggests that students who observe a positive reinforcement about graduate education from professors, advisors and administrators have a higher likelihood of considering graduate education. Among the family and friends, father's attitude has the highest influence on the student's interest in graduate education.

Table 15: Bivariate Correlation Coefficients Between Interest in Computer Science/Information Technology Education and Attitude of Various Groups

Groups	Pearson's Correlation Coefficient.	Significance (2-tail)
Mother's Attitude	0.189	0.028
Father's Attitude	0.280	0.001
College_Graduate Sibling's Attitude	*0.108*	*0.213*
Non_College_Graduate Sibling's Attitude	0.173	0.044
Advisors' Attitude	0.328	0.000
Professors' Attitude	0.338	0.000
Administrator’s Attitude	0.316	0.000
Close Friends' Attitude	0.200	0.020
Other Friends' Attitude	0.221	0.010

A similar correlation coefficient analysis was conducted to establish the relationship between four major theme areas identified during the focus group stage (Table 16). These factors were ‘knowledge of graduate programs,’ ‘education preparedness,’ ‘need to work/financial considerations’ and ‘perceived value of graduate education.’ All of the correlation coefficients were highly significant. These factors have negatively influenced the interest in graduate education. That is, if financial considerations are more important, then that student is unlikely/less likely to consider graduate education, at least, in near terms. Financial factors were at the top of the list in influencing the students’ considerations regarding graduate school. However, the second most influential factor was knowledge of graduate schools’ process and programs. The education preparedness was ranked third.

The survey analysis confirmed the focus group findings. However, during the focus group discussions students repeatedly question the value of graduate education in computer science/information technology compared with work experience and technical certifications. This factor had the lowest value of correlation coefficient with interest in graduate education. This factor needs further investigation.

The survey collected data on several demographic and other variables as well. The correlation coefficients between interest in the graduate program and these demographic variables were also calculated. The interest in graduate education was not significantly correlated to any of these variables. These factors were type of institution (public-private), age, gender, background and type of high school attended showed no significant relationship with interest in the graduate programs. More importantly, grade point average (GPA) and relevance of mathematics and science to computer science/information technology graduate curriculum had insignificant relationship with the dependent variable. The only other factor which showed significance was the education level of father. A student whose father had no college diploma showed strong conviction towards finding a job rather than considering graduate education.

Table 16: Bivariate Correlation Coefficients between Interest in Computer Science/Information Technology Education and Other Independent Variables

Factors	Pearson's Correlation Coefficient.	Significance (2-tail)
Knowledge Graduate Schools and Programs	-0.500	0.000
Education Preparedness	-0.386	0.000
Want To Work/Need Money	-0.542	0.000
Value Technical Education/Experience	-0.303	0.000

CONCLUSIONS AND FUTURE RECOMMENDATIONS

Based on the results of the focus group and survey, several factors were identified that influence students to consider or not consider graduate education in computer science/information technology. These factors include both positive and negative factors. The positive factors include attitude of college superiors (teachers, advisors, and administrators), peer groups and family, especially father. The negative factors include a strong job market and highly compensated jobs in computer science/information technology, lack of information about graduate schools including the process of application, lack of perceived preparedness and poor market value (perceived or real) of graduate education. It was also found that factors like grade point average, age, gender, urban background, interest in mathematics or science do not have strong relationship with graduate education plans. Furthermore, the interest in graduate education drops significantly from sophomore level to junior level. Data showed significant drop at senior level as well but sample size was too small to confirm it.

It is evident from the data analysis that graduate education in computer science /information technology currently faces stiff competition from a strong job market in the IT sector. However, there are strong indications for possible steps which can be taken to increase students' interest in graduate education. We can make several recommendations. These recommendations include that schools should provide more information on graduate schools and admission process, organize frequent information sessions for family and if possible, simplify the graduate admission process. The graduate schools should facilitate aggressive mentoring programs through teachers, advisors, and administrators. These mentoring programs should start early like sophomore year. The schools should increase interaction with "peer-group" role models to alleviate the fear about preparedness and to enhance confidence level. It is important that students learn about graduates from their own schools or areas succeeding in graduate school to put to rest.

ACKNOWLEDGMENT

This research was funded by the National Science Foundation Grant # CNS 0420473, 2004-2005.

REFERENCES

Armstrong, E., & Thompson, K. (2003). Strategies for Increasing Minorities in the Sciences. *Journal of Women and Minorities in Science and Engineering.* 9 (2).

Baylor, S. J. (2003). *Pursuing Critical Mass: The Coalition to Diversify Computing*. Retrieved January 5, 2004. Web Site: http://www.npaci.edu/envision/v14.3/cdc.html.

Boulding, M.E. (1985). Recruitment and Retention. In Landis, R.B. (Ed.), *Handbook on Improving the Retention and Graduation of Minorities in Engineering*. New York: The National Action Council for Minorities in Engineering, Inc.

Council of Higher Education Accreditation: CHEA (2000). *Research and Information*. Retrieved March 12, 2005. Web Site: http://www.chea.org.

Cohoon, J. M. (1999). Departmental Differences Can Point the Way to Improving Female Retention in Computer Science. *ACM: SIGCSE Bulletin*. 31 (1).

Cohoon, J. M., & Aspray, W. (2006). *Women and Information Technology: Research on Under-Representation*. Cambridge, MA: MIT Press.

Eide, E., & Waehrer, G. (1998). The Role of the Option Value of College Attendance in College Major Choice. *Economics of Education Review.* 17 (1).

Fletcher, S. L.; Newell, D.C.; Anderson-Rowland, M.R. & Newton, L.D.(2001). The Women In Applied Science And Engineering Summer Bridge Program: Easing the Transition For First-Time Female Engineering Students. In *Proceeding of 31th ASEE/IEEE Frontiers in Education Conference*, Reno, NV, October.

Gates, A. Q.; Teller, P. J.; Bernat, A. & Cabrera S. (1999). A Cooperative Model for Orienting Students to Research Groups. *In Proceedings of the 29 ASEE/IEEE Frontiers in Education Conference*, San Juan, PR, November.

Graham, J.M., Caso, R., Rierson, J., & Lee, J. (2002). The Impact of The Texas LSAMP on Under-Represented Minority Students at Texas A&M University's College of Engineering: A Multi-Dimensional Longitudinal Study. In *Proceedings of the 32nd ASEE/IEEE Frontiers in Education Conference*, Boston, MA, November.

Grandy, J. (1994). *Gender and Ethnic Differences among Science and Engineering Majors: Experiences, Achievements, and Expectations*. (GRE Board Research Report No. 92-03R) Princeton, NJ. (ERIC Document Reproduction Service No. ED388502).

Houston-Brown, C. K. (2002) Perceived Barriers To African Americans And Hispanics Seeking Information Technology (Doctoral Dissertation University of La Verne, La Verne, CA). *Dissertation Abstracts International,* 63, 06A.

Landis, R.B. (Ed.) (1985). *Handbook on Improving the Retention and Graduation of Minorities in Engineering.* New York: The National Action Council for Minorities in Engineering, Inc.

National Science Foundation (2000). *NSF Workshops Report on Underrepresentation of Women and Minorities in Information Technology*. (National Science Foundation Report Number NSF PR 00-40 - June 7, 2000.)

Nicholson, K., Hancock, D., & Dahlberg, T. (2007). Preparing Teachers and Counselors to Help Under-Represented Populations Embrace the Information Technology Field. *Journal of Technology and Teacher Education.* 15 (1).

Mitchell, Jr., D. (1997). Oklahoma Alliance for Minority Participation in Science, Mathematics, Engineering, Technology, and Education. (1997 Report for "Program Effectiveness" Reviews, NSF). Retrieved January 11, 2004. Web Site: http://ls-okamp.biochem.okstate.edu/images/stories/97PER.pdf.

Thom, M.; Pickering M., & Thompson, R.E. (2002). Understanding the Barriers to Recruiting Women in Engineering and Technology Programs. In *Proceedings of the 32nd ASEE/IEEE Frontiers in Education Conference*. Boston, MA, November.

Vegso, J. (2005). CRA Taulbee Trends: Ph.D. Programs and Ethnicity. Computing Research Association Survey. Retrieved April 15, 2006. Web Site: http://www.cra.org/info/taulbee/ethnicity.html.

APPENDIX I.

Survey Questionnaire--Interest in Graduate Study in Computer Related Fields

This questionnaire is part of an attempt to determine reasons why undergraduates decide to attend or not attend graduate school in computer-related fields of study. Your responses will help focus faculty efforts in the three schools engaged in this project to encourage graduate study. Please complete the following to best of your abilities.

1. What is your major?___________________

2. What is your classification? Sophomore________, Junior______, Senior______, Other_____

For Questions 3 and 4
Choose from the following descriptors for the highest educational level of your parents:
(1) non-high school graduate
(2) high school
(3) college but no degree
(4) technical school
(5) specialized military training
(6) associate's degree
(7) bachelor's degree
(8) post-graduate non degree
(9) master's, degree
(10) certificate of advanced graduate study
(11) doctorate
(12) post-doctorate
(13) other.

Enter the appropriate number in the space provided.

3. Father____ Describe if other________________________________

4. Mother____ Describe if other________________________________

5. How many older siblings do you have?_____

6. How many preceded you in college?_________

For Questions 7 and 8

Which of the following best describes **your parents'** attitudes toward graduate school in a computer-related field for you:

(1) never discussed (2) think I should not go (3) leave it up to me

(4) think it would be good for me (5) expect me to go (6) are **insisting** I go

(7) want me to go into a different field in graduate school (8) other.

Enter the appropriate number in the space provided.

7. Father____ Describe if other______________________________

8. Mother____ Describe if other______________________________

For Questions 9 and 10

(1) never discussed (2) think I should not go (3) leave it up to me

(4) think it would be good for me (5) expect me to go (6) are **insisting** I go

(7) want me to go into a different field in graduate school (8) other.

Enter the appropriate number in the space provided.

9. College graduate sibling(s)____ Describe if other_______________________________

10. Non college graduate sibling(s)___ Describe if other_______________________________

For questions 11, 12, and 13,

In general, which of the following best describes **your professors'** attitudes toward graduate or professional school for you:

1) never discussed (2) think I should not go (3) leave it up to me

(4) think it would be good for me (5) expect me to go (6) are **insisting** I go

(7) want me to go into a different field in graduate school (8) other.

Enter the appropriate number in the space provided.

11. Advisor____ Describe if other_______________________________

12. Professors who know you well ____ Describe if other_______________________________

13. Administrators who know you____ Describe if Other_______________________________

For Questions 14 and 15,
In general, which of the following best describes **your peers'** attitudes toward graduate or professional school for you:
1) never discussed (2) think I should not go (3) leave it up to me
(4) think it would be good for me (5) expect me to go (6) are **insisting** I go
(7) want me to go into a different field in graduate school (8) other.

Enter the appropriate number in the space provided.

14. Closest friend____ Describe if other______________________________
15. Other students____ Describe if other______________________________

For Question 16,
As of today, list as many of the following that describe **your own** position toward attending graduate school in a computer-related field:
(1) I have not given it much thought
(2) I want to go to work
(3) I am not sure what I want to do after graduation
(4) I would like to go but I am afraid I will not qualify
(5) I would like to go but can't afford it
(6) I will definitely go
(7) I want to go into a different field in graduate school
(8) other.

Choose as many as applicable.
16. ____,___,____,____,____,____,___. Describe if other______________________________

Do not complete items 17-39 if you are definitely going to graduate or professional school in a computer related or other field. If you are planning on graduate study in a computer related field, please go to question number 40.

If you are currently not planning on graduate study in a ***computer science or related*** field, select a descriptor(s) from the list below that best reflects your reasoning. Please check as many as applicable.

___17. I do not have knowledge regarding graduate programs.
___18. I do not like complexity of grad school application process.
___19. I do not think I'll meet graduate school entrance requirements.
___20. I do not want to deal with the graduate school workload and difficulty.
___21. My undergraduate program did not prepare me for grad school.
___22. Graduate school is beyond my ability.
___23. I can get a job with a bachelor's degree where the pay is good, and I don't think grad school will pay off.
___24. I need to begin earning a living.
___25. I need a change of pace/lifestyle from college life.
___26. I am tired of going to school
___27. I believe my undergrad program will provide me with the skills needed to get a good job and that grad school will not add that much.
___28. I can't afford grad school.
___29. There are not enough scholarship opportunities from colleges and federal government.
___30. My family wants me to go to work.

___31. I am no longer interested in working with or studying computer-related topics.
___32. I should not have majored in a computer-related field.
___33. I value experience more than graduate education.
___34. I think management education is more important after my undergraduate degree.
___35. I don't see many good graduate computer science programs in the Historically Black Colleges Universities.
___36. I don't see many minorities' students in the graduate CS programs.
___37. I believe that technical certifications are more valuable than graduate school.
___38. I don't see many minorities' role models in computer sciences.
___39. Other, please describe___________________________________

Please respond to questions from 40-44, **if you are currently planning graduate study in a computer related field**. Select all the applicable from the list below.

___40. Have you done guided research as an undergraduate?
___41. Have you written extensive research papers or technical reports other then course related papers?
___42. Have you had an internship while in college?
___43. Have you attended special programs for graduate school preparation?
___44. Have you had a cooperative education experience while in college?

Tell us something about yourself:

45. Gender M F

46. Age Under 20 20-22 23-24 25-26 Over 26

47. State of domicile_________

48. Your high school district is best described as: Urban_____ Rural_____

49. You have attended
(1) regular public high school (2) private/faith-based high school
(3) science and technology high school/program (4) magnet high school/program
(5) charter high schools (6) other.

Please select one of the above that describes your high school education the best _________
Describe if other______________________________

50. College GPA (1) 2.0-2.50 (2) 2.51-3.00
(3) 3.01-3.50 (4) 3.51-4.00 (5) Over 4.00

51. Number of courses you have already completed in an IT, CS, MIS etc. environment______.

52. Do you think all mathematics classes required in the program have direct relationship with your current major?
Yes______ No_____ Don't know _______

53. Do you think all pure science classes required in the program have direct relationship with your current major?
Yes______ No_____ Don't know ______

54. Degree of your interest in the highest level of mathematics course you have taken, can be best described as:

Very high____ High____ Acceptable____ Low____ Very low____

55. Degree of your interest in the highest level of pure science course you have taken, can be best described as:
Very high____ High____ Acceptable____ Low____ Very low____

56. Are there any comments regarding the topics and issues referred to in this questionnaire that you would like to discuss?

SEMI-AUTOMATED IDENTIFICATION OF FACETED CATEGORIES FROM LARGE CORPORA

Jose Perez-Carballo, California State University

ABSTRACT

Several studies, suggest that interfaces that present results organized into categories or faceted hierarchies meaningful to users may help them make sense of their information problem as well as the information system itself.

This paper describes a system that generates facets in a semi-automatic way from corpora gathered from the web.

Such tool would make the work of knowledge engineers faster, easier, and cheaper. It could also be used to make search and browsing engines more effective and more user friendly.

A "facet" is an aspect of a topic (Anderson and Perez-Carballo, 2005). For example the following sets would be reasonable facets in a cooking domain: ingredients (e.g. apples, cayenne pepper, chocolate), utensils (e.g. egg slicer, funnel, grater, potato masher), processes (e.g. basting, poaching, pressure cooking), etc.

The central hypothesis of this paper is that multi-word terms that appear in a similar grammatical context are likely to belong to the same facet. For example, "cayenne pepper" and "chocolate" are likely to appear in similar contexts, which are likely to be different from the contexts in which "potato masher" and "egg slicer" appear.

The tests described here suggest the method presented is useful and effective.

INTRODUCTION

This paper describes FFID (Fast Facet Identifier), a system that can be used to compute facets from a corpus of documents. FFID uses a fast simplified clustering algorithm that allows the identification of hundreds of facet clusters from a corpus of hundreds of thousands of sentences in a very short time (seconds). The automatic identification of facets may be a very powerful tool to design better information retrieval systems. The goal of information retrieval is to support people in searching for the information they need. Given an information problem, finding relevant (let alone high quality) documents is difficult. The sheer amount of information available on line makes this a difficult problem. The size of the web is debatable (Markoff, 2005) but it must be by now at least 12,000 million pages. If each one of these web pages were printed on a standard A4 sheet of paper (21-cm wide), and put side to side on a straight line, it would take about 60 earth circumferences to lay them all down. This is a lot of information. People learn about their information problem and about the information resource they are using through interaction with the resource. Human-computer interaction is the crucial phenomenon of the information retrieval process. Fast algorithms, hardware for storage and processing, data and knowledge structures are important but useless if we do not understand how humans interact with machines when looking for information. All the techniques we use must first take into account what we are doing this for: the user. Users encounter several problems when they approach an information resource:

(1) Users seldom understand their information problem. Belkin and Croft define information need as a problematic situation where a person cannot attain their goals due to lack of resources or knowledge (Belkin and Croft, 1992).

(2) Users cannot articulate their problem, they need help constructing and refining queries. In his classic and still relevant paper from 1968 Taylor (Taylor, 1968) argues that users might have a vague information need, but it may not be clear enough for them to articulate it. Belkin's "anomalous state of knowledge" (ASK) hypothesis (Belkin, 1980) proposes that when users encounter a problematic situation, the resulting cognitive uncertainty makes it difficult for them to adequately expressing their information need.

(3) Users do not know whether what they are looking for may or may not be in the collection they are searching (Hearst, 1999).

(4) Users may not be aware that there are other interpretations for their questions (Venkatsubramanyan and Perez-Carballo, 2007).

(5) Users may not be familiar with the information resource's user interface (UI) (Xie and Cool, 2009).

(6) Users may not be able to recognize that an item is useful or relevant even after it is presented to them (Xie and Cool, 2009).

(7) The relevant documents may be lost in a large number of results returned by the information resource (Xie and Cool, 2009).

The problems listed above make it desirable to have an interface capable of supporting browsing or exploration. Or like one of my colleagues used to tell his students: "browsing is what you want to do when you don't really know what you want!" (James Doig Anderson, personal communication).

Several studies (Hearst, 2006; Venkatsubramanyan & Perez-Carballo, 2007), suggest that interfaces that present results organized into categories or faceted hierarchies meaningful to users may help them make sense of their information problem as well as the information system itself.

There are several open problems with the design of UIs that present organized results including how to generate useful groupings and how to design interfaces that can use them effectively. For a survey of such interfaces see Venkatsubramanyan & Perez-Carballo, 2007.

Two common ways of generating groupings are: document clustering and facet categorization. Marti Hearst and her group (Hearst, 2006) have described the differences between these methods (see section 2 for a discussion of the differences). The essential differences are that document clustering creates sets of similar documents and attempts to label each set in some useful way, while facet categorization uses the different aspects of a topic in order to classify it (next paragraph defines "facet"). In this paper an automatic way to generate facet categorization is presented.

A "facet" is an aspect of a topic (Anderson and Perez-Carballo, 2005). For example the following sets would be reasonable sets of facets in a cooking domain: ingredients (e.g. apples, cayenne pepper, chocolate), utensils (e.g. egg slicer, funnel, grater, potato masher), processes (e.g. basting, poaching, pressure cooking), dishes (e.g. a jiaco, bengal potatoes, bhuna khichuri, black-eyed peas, kale), herbs (e.g. basil, chicory, dill), etc.

The central hypothesis of this paper is that multi-word terms (MWTs) that appear in a similar grammatical context are likely to belong to the same facet. For example, "basil" and "dill" are likely to appear in similar contexts which are likely to be different from the contexts in which "funnel" and "grater" appear. The following examples show sentences where MWTs (shown in bold face) appear in similar contexts:

1. dip in **beer batter** and shake gently
2. dip in **seasoned flour** and shake gently to remove excess
3. garnish with **parsley**, if desired
4. garnish with **pecan halves**, if desired
5. Free scores by **Anton Bruckner** in the Werner Icking Music Archive
6. Free scores by **Luca Marenzio** in the Choral Public Domain Library - ChoralWiki –

The previous sentences provide evidence that "beer batter" and "seasoned flour" belong to a facet, "parsley" and "pecan halves" to another, and "Anton Bruckner" and "Luca Marenzio" to yet another. FFID identifies MWTs and clusters them according to the degree of similarity of the contexts in which they appear. In some cases, FFID may be able to identify a useful label for the facet itself (e.g. Composer) and well as clustering the corresponding values (Vivaldi, Brucker, etc.) under that label.

The methods followed by the system described here are not very different from the methods that librarians and indexers are taught to use when they are charged to find facets from texts (Anderson and Perez-Carballo, 2005). Facets have a very strong correspondence with grammatical categories. Librarians are taught to create sentences that describe the topic of an item they want to classify. Each grammatical category they find is made into a facet.

The following example is taken from the book by Anderson and Perez-Carballo, 2005. The topic is: "African-American missions in Western Africa". From the topic, the librarian creates a phrase, such as: "African American Baptists, including the National Baptist Convention of the U.S.A. and Lott Carey in the Southern States evangelized West Africans in the 19th century."

The verb "evangelization" corresponds to the "operation" facet. The verb's object "West Africans" corresponds to the "entity" facet. The subject "African Americans, National Baptist Convention of the U.S.A. and Lott Carey" corresponds to the "agent" facet. The propositional phrases such as "in the 19th century" and "in the Southern States" generally correspond to "time" or "place" facets.

This paper is divided in the following sections: *Introduction, Faceted Metadata and Information Exploration, System Description, Corpus Building, Term Identification Phase, Facet Identification Phase, Context Normalization, Cluster merging, Human Intervention, Comparison with Other Semi-Automatic Methods, Conclusions and Future Work.*

FACETED METADATA AND INFORMATION EXPLORATION

In the techniques that use document clustering, groups of documents are created by defining some similarity measure among them. The documents themselves are clustered, and labels are chosen for each cluster by the algorithm. Examples of this technique are: Clusty (http://clusty.com/), and Vivisimo (Rivadeneira and Bederson, 2003) (http://vivisimo.com)

The following example shows the list (generated Jan 26 2009) of clusters generated by clusty.com for the query using the terms: "tosca, opera, Puccini."

Giacomo Puccini
Opera House
Reviews
Amazon
Metropolitan Opera
Event
DVD
Pavarotti, Singing
Floria Tosca

There are several problems with the clusters in the previous example which may be problematic to users. These problems have been described by in Hearst, 2006. The cluster categorization shown is created by clustering documents. One of the problems with this kind of clustering is that the labels assigned by the program are misleading to users. For example, different aspects of the topic appear at the same level (e.g. the composer, performers, opera characters, formats such as DVD). While items that might belong to the same aspect appear at different levels. For example, "Metropolitan Opera" appears at the first level while "State Opera Prague" (not shown) appears under "Giacomo Puccini" when the user expands that item. A knowledge engineer would have created sets of conductors, singers, orchestras, venues, formats, etc. in order to facilitate browsing of the information by users.

Usability studies that compare different kinds of groupings show that systems that use facet categorizations are more useful to people (Hearst, 2006). Users seem to find facets easier to understand and use. This is the same kind of categorization that librarians prescribe to help users browse through unfamiliar topics (see for example Anderson and Perez-Carballo, 2005).

Facet categorization utilizes the aspects or "faces" of a topic in order to classify it. Facet classification often uses pre-existing metadata that has been created by human experts. Facets are similar to the traditional questions: who, what, where, when, why. They are fundamental characteristics that can be used to analyze and describe any topic.

Facet classification is not a new idea. Traditional library classifications have always been based on facets. Shiyali Ramamrita Ranganathan (1892-1972) described a facet system in the 1930s (Svenonius, 1992; Anderson and Perez-Carballo, 2005). Ranganathan suggested 5 basic facets:

(1) entities or "things," people, natural or artificial objects, animals and plants, institutions, etc.
(2) attributes - "part-of " relationships, symptoms, properties, etc.
(3) actions, processes, events
(4) places
(5) time

A faceted hierarchical classification uses a set of category hierarchies (instead of only one). Each hierarchy corresponds to a different facet (dimension or property). Any topic can be described specifying each of the relevant facets. Several researchers (English et al., 2002; Yee et al., 2003; Stoica et al., 2007) have tested interfaces that use facets in order to support information exploration and identification.

If the facet classification is not built automatically from the corpus it is difficult and very expensive to build the categories by human intellectual labor. Using a pre-built classification (like the WordNet approach described in Stoica and Hearst, 2004) it is possible to miss topics that are important in the corpus, or emerging topics (such as current events and people). In the experiments described by Stoica and Hearst (op cit) the algorithm builds categories for ingredients, dishes, cooking utensils and people. But there are facets (present in the corpus) that it does not cover, such as type of cuisine, because they do not exist in WordNet. Similar problems would occur for the corpus of a company describing its products, or a corpus containing information about current events.

SYSTEM DESCRIPTION

The following is a high level description of FFID. Precise details of each step are given in the following sections. The steps are:

Corpus building.
Term identification phase. Compute multi-word terms (MWTs) from the corpus.
Facet identification phase. Use a clustering algorithm to group the previously found MWTs into sets according to the similarity of their grammatical contexts.

CORPUS BUILDING

We used the GNU Wget spider (http://www.gnu.org/software/wget) to gather documents from the Web to be used in test corpora. Wget allows us to download a number of web pages from a chosen starting point and to traverse the web to a predetermined depth. Using this method we collected documents for the different corpora we used in our tests. Some of the corpora we used are collections of articles from Wikipedia (http://en.wikipedia.org/wiki/Main Page). No corpus or topic-specific training was used.

The classical music corpus used in our tests was built pointing the spider to the European classical music Wikipedia article and using a recursion level of 3. It has 1,162 files, occupying around 97MB, and contains about four million words.

TERM IDENTIFICATION PHASE

In our first attempts we tried to identify terms using purely statistical methods (Venkatsubramanyan and Perez-Carballo, 2004). This approach had the advantage of being domain and language independent. But too many of the terms identified were of poor quality. In our current approach we use part of speech (POS) information in order to identify terms. The steps of the term identification phase are as follows:

(1) *using the Brill tagger (Brill, 1995), tag the corpus with part of speech information.*
(2) *using predetermined regular patterns of POS tags, identify sequences of terms that are good candidates for multi-word terms. For example: a sequence of singular nouns is a candidate for MWTs.*
(3) *each MWTS is saved with the sentence where it was found. The sentence will be used later to cluster the MWTs according to the context in which they appear in the corpus.*

The Part of Speech Tagger (POST) uses the Penn Treebank II tag coding where a string of the form "/TAG" is appended to each word. Table 1 shows an explanation for the TAGs that appear in the tagged example below. For example "Austrian/JJ" means that the word "Austrian" is an adjective.

Example of a paragraph to be tagged:

Besides the nine completed numbered symphonies , principal works are the song cycles Lieder eines fahrenden Gesellen - rendered as Songs of a Wayfarer , literally Songs of a Travelling Journeyman - Kindertotenlieder - on the Death of Children - the synthesis of symphony and song cycle that is Das Lied von der Erde - Song of the Earth .

Example of the paragraph after tagging:

Besides/IN the/DT nine/CD completed/VBN numbered/VBN symphonies/NNS ,/, his/PRP $principal/JJ works/NNS are/VBP the/DT song/NN cycles/NNS Lieder/NNP eines/FW fahrenden/FW Gesellen/FW -/: usually/RB rendered/VBN as/IN Songs/NNPS of/IN a/DT Wayfarer/NNP ,/, but/CC literally/RB Songs/NNPS of/IN a/DT Travelling/VBG Journeyman/NNP -/: and/CC Kindertotenlieder/FW -/: Songs/NNPS on/IN the/DT Death/NN of/IN Children/NNS -/: and/CC the/DT synthesis/NN of/IN symphony/NN and/CC song/NN cycle/NN that/WDT is/VBZ Das/NNP Lied/NNP von/NNP der/NNP Erde/NNP -/: The/DT Song/NNP of/IN the/DT Earth/NNP ./.

Table 1: Penn Treebank II tagset (partial list)

POS Tag	Description	Example
CC	coordinating conjunction	and, or, but
CD	cardinal number	1, third
DT	determiner	a, the
FW	foreign word	d'hoevre
IN	preposition/subordinating conjunction	in, of, like, from
JJ	adjective	green
NN	noun, singular	table
NNP	proper noun, singular	John
NNPS	proper noun, plural	Vikings
NNS	noun plural	tables
PRP	personal pronoun	I, he, it, himself, his
RB	adverb	however, usually, naturally, here, good
VBG	verb, gerund/present participle	taking
VBN	verb, past participle	taken
VBP	verb, sing. present, non-3d	take
VBZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which

The following are two regular expressions used in the system in order to detect MWTs. The syntax used to specify the regular patterns uses the regular pattern syntax of the Java language:

```
(1)([^ ]+/JJ ){0,1}([^ ]+/((FW )|(NN )|(NNS )|(NNP )|(NNPS ))+
(2)([^ ]+/(NNP |NNPS ))((of/IN ){0,1}[^ ]+((FW )|(NN )|(NNS )|(NNP )|(NNPS )))*
```

Some notes about the syntax used in Java to express regular expressions: [ˆ]+ matches a sequence of any number of characters excluding space; {0, 1} means previous expression must appear either once or not at all. (FW)|(NN) matches either "FW " or "NN ". The expression identified above with (1) will match a word tagged with any of FW, NN, NNS, NNP, or NNPS. The pattern must appear at least once and can be repeated any number of times and must be preceded by at most one JJ.

In the experiments presented here only the two patterns above were used. Using the patterns listed above the following are examples of MWTs that were found in the music corpus:

disco/NN
church/NN music/NN
echoes/NNS
Carthage/FW
vocal/JJ music/NN
Gregorian/JJ chant/NN
medieval/JJ churches/NNS
Middle/NNP Ages/NNPS
Bobby/NNP McFerrin/NNP
United/NNP States/NNPS
Boyz/NNP II/NNP Men/NNS
Manhattan/NNP Transfer/NN
El/NNP Salon/NNP Mexico/NNP
Adam/NNP de/FW la/FW Halle/NNP
Cincinnati/NNP Symphony/NNP Orchestra/NNP
Le/NNP Jeu/NNP de/FW Robin/NNP et/FW Marion/NNP
Eastern/NNP Orthodox/NNP Christian/NNP Church/NNP music/NN

The patterns we chose to use for the experiments described in this paper are very simple and may not match some terms that should be matched. For example, the following name of a piece by Aaron Copland would not be matched by the previous expressions:

"Fanfare/NNS for/IN the/DT Common/NNP Man/NN"

The purpose for the system described here was to achieve higher precision at the cost of recall. In order to use the system with a corpus in a different language it would be necessary to change the POST module (part of speech tagger) as well as the patterns of POS tags used to identify MWTs. In principle, the POST should be trained for different domains for optimum performance but in the experiments we ran on different domains without specific training the results were sufficiently good even though the POST made tagging mistakes.

FACET IDENTIFICATION PHASE

The great difficulty of processing natural language is the great variability and flexibility of human language. Among other issues: there are several counterexamples for almost every rule, and world knowledge has to be taken into account as well as syntax. Our approach for this system is simpler. We are not trying to design a method that will parse and interpret correctly every possible sentence. For our method to cluster two MWTs together we only need to find some sentences in the corpus in which the MWTs appear in the same syntactic context. Thus the two sentences "Travel guide to France from Wikitravel" and "Travel guide to England from Wikitravel" provide evidence that "France" and "England" appear in similar contexts so they may belong to the same cluster. All other MWTs that appear in the same context are added to the same facet. Our hypothesis is that if two terms belong to the same facet then in a big enough corpus it should be possible to find several contexts that are similar enough for our method to cluster the two terms together. The more clusters we find that share the same two terms, the more likely it is that those terms belong to the same facet.

One of the advantages of the method described here is that it is possible to identify facets (e.g. "countries") and facet values (e.g. "France", "England") not known in advance. For example, places are recognized as places because they appear in the same kind of grammatical contexts. The sentence "they built a house in Tlayacapan" allows the system to decide that "Tlayacapan" is likely to belong to the same facet as any term X that appears in a sentence "they built a house in X". The system decides that "Tlayacapan" is a value of a facet that corresponds to the context "build a house in".

During the facet identification phase, corpus sentences containing the terms found during the term identification phase are clustered. The hypothesis is that each cluster should be a facet relevant to the corpus. In a first, proof of concept, version of FFID an implementation of a hierarchical agglomerative clustering algorithm was used. The quality of the resulting facets was encouraging. The algorithm used a similarity function to compute a distance between every pair of contexts and build a distance matrix. Even for a small collection of thousands of sentences, building the distance matrix required a significant amount of time. The performance for tens of thousands of sentences was simply not practical. If a measure of similarity is to be computed between each pair of grammatical contexts the complexity of the algorithm will be O(n2). This means that if clustering a few thousand sentences takes on the order of minutes, clustering tens of thousands may take days. This is the reason why clustering algorithms are seldom practical. The algorithm used by FFID is a fast, simplified clustering algorithm that can cluster hundreds of thousands of terms in seconds. Before the simplified clustering algorithm is described, consider the following examples of sentences found in a potentially very large list of sentences. The grammatical context of the terms shown in bold face are the tagged words that appear next to them in the same sentence either to the left or the right.

(1) *Free/JJ scores/NNS by/IN* ***Andrea/NNP Gabrieli/NNP*** *in/in the/dt choral/nnp public/nnp domain/nnp library/nnp*

(2) *Free/JJ scores/NNS by/IN* ***Giovanni/NNP Pierluigi/NNP da/NNP Palestrina/NNP*** *in/in the/dt choral/nnp public/nnp domain/nnp library/nnp*

(3) *Scores/NNS by/IN* ***Jacob/NNP Clemens/NNP non/NNP Papa/NNP*** *could/MD be/VB found/VBN in/in the/dt choral/nnp library/nnp*

A Clustering algorithm would require the comparison of each sentence with each other sentence in order to generate a distance matrix that would store the degree of similarity between every pair of sentences in the corpus. In order to avoid the O(n2) comparisons, in FFID the sentences are processed first in order to transform them into strings that, when sorted, will result in similar contexts being next to each other. The next paragraphs illustrated this method with an example.

For each sentence, LT R, where L is the sequence of words on the left of the term T and R is the sequence on the right of the term, the sentence is transformed into reverse(L); R; T , where reverse(L) represents the words in sequence L listed in inverse order. For example, the following sentence (where the term T is shown in bold face):

Free/JJ scores/NNS by/IN ***Giovanni/NNP Pierluigi/NNP da/NNP Palestrina/NNP*** *in/in the/dt choral/nnp public/nnp domain/nnp library/nnp*

is transformed into:

by/IN scores/NNS Free/JJ; in/in the/dt choral/nnp public/nnp domain/nnp library/nnp; ***Giovanni/NNP Pierluigi/NNP da/NNP Palestrina/NNP***

The reason the order of the words in the L sequence is reversed is that we use the hypothesis that the closer words are to the term T the more significant they are. So the word "by/IN" in the previous example is, in that sense, more significant than the word "Free/JJ".

Now when the transformed sentences are sorted, the terms that have the same context (same L and R) will appear together in the sorted list. The previously described simplified process is nlog(n) but would find the same clusters as a O(n2) process that uses a distance matrix built using a similarity function that returns the maximum value if L and R are exactly the same and the minimum value otherwise. A similarity function may seem more flexible than the kind of comparisons possible using the simplified algorithm but it is possible to simulate more complex similarity functions using our simplified process. For example, using different transformations we could cluster together all sentences with the same L, or sentences such that have the same prefix of size n of L and the same prefix of size m of R. In a section below we list the transformations used for the tests described in this paper. In the experiments described here we used n = m = 2.

CONTEXT NORMALIZATION

The problem now is that without a similarity function we seem to have lost the ability to do subtle comparisons between each pair of contexts. Since similarity now has been reduced to proximity in a sorted list of contexts, two MWTs must appear in almost exactly the same context in order to appear close to each other in the sorted list so they can be included in the same cluster. The consequences would be very small clusters and MWTs that should be clustered together would not be because they do not appear in exactly the same context.

To solve the previous problems a normalization module is added before sorting. This module reduces several different sentences to the same normal form. Consider the two following sentences:

Antonio Vivaldi is a very popular composer
Arcangelo Corelli was an Italian composer born in Ravenna

The previous two sentences are too different for them to cluster together using the method previously described. In other words, after sorting, the contexts "is a very popular composer" and "was an Italian composer born in Ravenna" would not appear close to each other in a sorted list of contexts. After normalization the same two sentences become:

Antonio Vivaldi ISA composer
Arcangelo Corelli ISA composer

Now the contexts of the terms have become the same. Without context normalization and a sophisticated similarity function to compare contexts, too few of them would be similar enough to allow for more than a few MWTs to cluster together. Context normalization is a process that maps several different contexts into a normal form. This is achieved by the use of regular expressions that are used to match contexts to normal forms. If a sentence matches the regular expression corresponding to a normal form, it is replaced in the list by that normal form. Example:

Deng/NNP Xiaoping/NNP and/cc other/jj leaders/nns focused/vbn on/in market-oriented/jj economic/jj development/nn

matches the expression (shown here using Java syntax):

^(or|and)/cc other/jj (([^]+/jj){0,1} ([^]+/(nn|nns|nnp|nnps))+)

so it is replaced by the normal form:

PATTERN_ISA leaders/nns;Deng/NNP Xiaoping/NNP

This kind of pattern is similar to the patterns proposed by Hearst (Hearst, 1998) in order to discover automatically WordNet relations.

Another advantage of reducing a number of sentences to a normal form is that it is possible to assign to them meaningful labels. In the previous example the label “isa leader” would be assigned to the cluster that groups all the terms with contexts “PATTERN_ISA leaders/nns.”

Table 2 shows the patterns used in the system described in this paper in order to normalize contexts. An id (such as P1, P2, etc.) is used to identify each pattern (pattern P3 was eliminated of the table because it was unsuccessful).

Table 2: Patterns used in the experiments

<table>
<tr><th>id</th><th>Normalized form</th><th>Sentence example</th><th>Pattern</th></tr>
<tr><td>P1</td><td>PATTERN_OF_THE cincinnati/nnp symphony/nnp orchestra/nnp ;conductor/NN</td><td>conductor/NN of/in the/dt cincinnati/nnp symphony/nnp orchestra/nnp</td><td>^(of/in)(the/dt)(([^]+(nn|nnp|nns|nnps))+)</td></tr>
<tr><td>P2</td><td>PATTERN_SUCHAS -artists/NNS;;Britney/NNP Spears/NNP</td><td>willing/JJ to/TO play/VB artists/NNS such/JJ as/IN Britney/NNP Spears/NNP</td><td>([^]+/(NNS|NN)) such/JJ ^as/IN</td></tr>
<tr><td>P4</td><td>PATTERN_VERB00 -absorbed/VBD Germany/NNP West/NNP;; East/NNP Germany/NNP</td><td>West/NNP Germany/NNP absorbed/VBD East/NNP Germany/NNP by/IN 1990/CD ./.</td><td>(([^]+/(NN|NNP|NNS|NNPS))+) ([^]+/VBD)</td></tr>
<tr><td>P5</td><td>PATTERN_ISA -composer/nn ;Kernis/NNP</td><td>Kernis/NNP is/VBZ an/DT active/JJ composer/NN in/IN high/JJ demand/NN</td><td>((is/vbz)|(are/vbp)) (a|an)/dt ([^]+/jj)? ((type|kind|class|sub-group|example|part)/nn of/in)*.*?(([^]+/(nnp|nn)[ ]){1,})</td></tr>
</table>

Table 2: Patterns used in the experiments

id	Normalized form	Sentence example	Pattern
P6	PATTERN_POSSESIVE - Copland/NNP;;Third/NNP Symphony/NNP	the/DT fourth/JJ movement/NN of/IN Copland/NNP ’s/POS Third/NNP Symphony/NNP	^(’s/POS)(([^]+(NN|NNP|NNS|NNPS))+).*
P7	PATTERN_INCLUDING - composers/NNS;;John/NNP Cage/NNP	with/IN other/JJ American/JJ composers/NNS including/VBG John/NNP Cage/NNP	^including/VBG (((,/,)|(-/:))){0,1}(([^]+/(NN |NNS |NNP |NNPS))+)([^]+/[^]+)*
P8	This label identifies clusters obtained without any normalization.		

CLUSTER MERGING

After normalization it is possible to identify clusters that can be merged with each other. For example all clusters following the pattern ISA <noun>, for some noun, are merged with clusters SUCHAS [^]+/jj <noun> for the same noun. For example, all the following clusters are merged into a “ISA composer” cluster:

(1) *ISA composer*
(2) *SUCHAS baroque/jj composers/nns*
(3) *SUCHAS composer/nn*
(4) *SUCHAS composers/nns*

THE FULL ALGORITHM

Step 1: Use Brill tagger to tag each sentence in the corpus with part of speech (POS) information.

Step2: Use the regular expressions listed above to identify MWTs.

Step3: For each MWTS, T, found, create I strings with the following format: “li (reverse(L));ri (R);T”, where T is the term found, L are all the words to the left of the term and R the words to the right. li and ri are transformations used to simulate a similarity function. For example, if both li and ri are the identity transformation (i.e. they return their argument unchanged) then they simulate a similarity function that returns the maximum value if the context of the two terms being compared (i.e. both the words on the right and the words on the left of the terms) is exactly the same. Another example: if both transformations are the same and are equal to a function that returns the first two words of its argument, then they are equivalent to a similarity function that returns the maximum value if the two words closest to the terms (on both sides) are the same (these are the transformations used in the tests described here).

Step4: Normalization. For each line generated by the previous process, if the line matches any of the regular expressions described previously in the normalization section, replace the line by the normal form that correspond to that regular pattern. For example the next sentence:

composers/NNS such/JJ as/IN Handel had/VBD before/IN him/PRP

becomes:

PATTERN_SUCHAS composers/NNS;;Handel

Step5: Sort all sentences. For each group of strings, “A;B;T” that share the same prefix “A;B;” create a cluster containing the terms “T”. For example: the following strings would result in clustering together the terms “Geography”, “Military”, and “Political divisions”.

(1) article/NN Main/JJ;of/in the/dt united/nnp states/nnps ./.;Geography
(2) article/NN Main/JJ;of/in the/dt united/nnp states/nnps ./.;Military
(3) article/NN Main/JJ;of/in the/dt united/nnp states/nnps ./.;Political divisions

Clusters created from lines such as:

PATTERN_SUCHAS composers/NNS;;Handel

are given labels. For example, the cluster create by all lines that begin with the contexts

PATTERN_SUCHAS composers/NNS

are given the label “composers”. Most clusters may not have such a useful label.

Step6: Cluster merging. After MWTs have been assigned to different clusters it may be possible to decide automatically to merge some of the clusters, given their automatically assigned label. In order to start the merging process, first sort all clusters by label (examples of cluster labels are “composers”, “countries”, etc). Then compare the label of adjacent clusters and apply merging rules. The only merging rule used in the experiments reported in this paper is the following: compare the labels of adjacent clusters, if either pattern is PATTERN ISA or PATTERN SUCHAS, and the last word is the same or its plural, then merge the terms of the clusters into a single cluster. Example: all the following clusters are merged into a single cluster labeled "PATTERN ISA song/nn":

(1) PATTERN ISA songs/nns,
(2) PATTERN ISA song/nn,
(3) PATTERN SUCHAS songs/nns,
(4) PATTERN SUCHAS song/nn

TESTING

This section describes the testing that was done in order to measure the quality and usefulness of the output of the system. It was decided not to measure recall, i.e. how complete is each set of facet values found. For example: what percentage of the composers mentioned in the corpus were identified. We did not measure either how complete is the set of facets. For example: of the facets mentioned in the corpus, such as composers, countries, instruments, etc., what percentage was found.

Another possible parameter that could be measured is the impact that different kinds of facets (including the ones found by this system) may have on users while they are solving information problems with the help of an interactive system. But facets have been described and used for generations by librarians (Svenonius, 1992; Anderson & Perez-Carballo, 2005), and other researchers have tested the usefulness of facets in modern search systems (Yee et al., 2003; English et al., 2002; Hearst, 2006). So we decided not to measure that parameter. The parameter we measured for this paper is what percentages of the facets discovered by the system would be judged useful for expert human indexers or knowledge engineers.

TEST DESIGN

In the section about normalization we discussed the patterns that are used to conflate many different contexts into one. Some of these patterns may generate useful clusters more often than others. Table 2 lists all the patterns used by the system. We designed a test to compare the proportion of useful clusters generated by each pattern, as well as clusters generated by contexts for which no normalization was provided (label P8 of table 2).

After all clusters are generated from the full corpus, a label is given to each cluster depending on its normalization pattern. The labels are P1 to P7 and are shown on Table 2 with their corresponding patters. P8 corresponds to clusters generated without using a normalization pattern. A file is created that contains at least 20 clusters for each label. The set of 20 clusters per label is selected at random from all the clusters generated by the system. This file was shown to a group of

3 experts that included 2 knowledge engineers and an indexer. They were asked to look at the sets and judge them either “good” or not. They were asked to interpret “good” as meaning “this cluster would have helped them shorten the time and effort required to create facet taxonomies by hand”. The experts worked together to judge the clusters and reach a judgement by consensus. This process is similar to the team work of knowledge engineers in the real world. At the end a cluster was judged good if they all agreed that it was good.

TEST ANALYSIS

Table 3 shows the data collected during the tests. The system generated 1553 clusters from the full corpus used in the experiment (46.6 MB). Of all the clusters generated 20 were chosen at random for each of 7 patterns. The percentage of good clusters was very different for each of the patterns. Clusters of type P2, for example, were judged to be useful significantly more often (85%) than clusters of type P1 (10%). In other words: some normalization patterns consistently resulted in more useful clusters than others.

More than half (56%) of the terms that appear in some cluster in the sample ended up in a cluster judged “good”. This suggests that the system would be helpful to knowledge engineers and indexers. Some kinds of clusters were generated much more often than others: 6043 P1 clusters vs 150 P4 clusters. But only 2 P1 clusters were found useful in the sample vs 8 P4 clusters.

Table 3: Test results

pattern id	all	P1	P2	P4	P5	P6	P7	P8
clusters in full output	1553	670	42	31	118	586	27	79
clusters in sample	140	20	20	20	20	20	20	20
good clusters in sample	65	2	17	8	8	3	13	14
% of good clusters in sample	46.4%	10.0%	85.0%	40.0%	40.0%	15.0%	65.0%	70.0%
terms that appear in some cluster in full output	14093	6043	321	150	1049	5083	172	1275
terms that appear in some cluster in sample	856	115	126	79	147	117	102	170
terms that appear in a good cluster in sample	483	12	117	29	91	18	77	139
% of good terms vs total nr of terms in sample	56.4%	10.4%	92.9%	36.7%	61.9%	15.4%	75.5%	81.8%
average size of clusters in full output	9.1	9.0	7.6	4.8	8.9	8.7	6.4	16.1
average size of clusters in sample	6.1	5.8	6.3	4.0	7.4	5.9	5.1	8.5
average size of good clusters in sample	7.4	6.0	6.9	3.6	11.4	6.0	5.9	9.9

HUMAN INTERVENTION

According to the analysis of our tests, not all the facets generated by the system would be useful. But the results indicate that the number of clusters generated automatically would speed up and facilitate the work of human experts.

The first step in creating any classification, including a faceted classification, is to determine what the facets should be or, in other words, what are the categories of topics and features of primary interest to the targeted user community (Anderson and Perez-Carballo, 2005). Consequently, an automatically generated faceted classification should be filtered and refined in order to include all and only the facets that will make it the most useful and usable for its users. Even if the output of this system has to be refined by an expert, it would still save a lot of time, effort, and money to start from the facets identified by the system. It is also likely that the expert would not have to have a lot of domain expertise in order to be able to refine the facets that are identified automatically.

The FFID system discovers mostly entities, agents, and places. Further regular patterns should be added to the term recognition phase in order to detect time and operations. Time could be detected by patterns that detect dates and operations by patterns that detect verb groups.

COMPARISON WITH ANOTHER SEMI-AUTOMATIC METHOD

Marti Heart's group at the University of California at Berkeley has been working with user interfaces designed to help users browse information that has been assigned hierarchical faceted metadata (Hearst, 2006). They have also worked on an algorithm, called Castanet, that generates hierarchical faceted metadata from textual descriptions of items (Stoica et al., 2007; Stoica and Hearst, 2004). Castanet uses WordNet (a large lexical database) to find facets.

The conclusions described in this section are the result of running FFID on corpora obtained from Marti Hearst's group and comparing the sets of facets identified by each system. The methods used by FFID and Castanet, and the corresponding results, are different but complementary. There are kinds of facets and corpora for which each approach works better. Castanet's terms are limited to one and two-word consecutive noun phrases. FFID identifies terms according to patterns of POS. The length of the resulting terms is not limited in FFID. Some examples of terms found by FFID are:

(1) *Café au lait*
(2) *Orange Lemon-Lime Soda Float*
(3) *No-Bake Chocolate Chip Cookie Pie*

Castanet is not able to recognize spelling variations. FFID identifies all variations used in the text and classifies them in the same category as long as they are used in the same contexts. The present version of FFID does not recognize that the variations may represent actually the same concept. For example “Licorice” and “Liquorice” are identified as terms and categorized as a “sweetener”. To our system both these terms are as different as “Cane syrup” and “Molasses”.

In FFID some verb forms are recognized as terms and categorized appropriately. For instance, terms such as “Basting”, “Poaching” and “Pressure cooking” are recognized and categorized as “methods”. If the same term appears in the corpus with different senses FFID is able to add it to several different categories. For example, a category is identified that includes “pepper”, “Szechuan Pepper”, “Cumin”, and “Garlic salt”. While there is another category that includes “pepper”, “Bell pepper”, “green pepper”, etc. Castanet, on the other hand, chooses one of the senses.

FFID is not limited by the coverage of a preset controlled vocabulary, such as WordNet. FFID will be able to recognize terms in the corpus as long as they conform to any of the preset patterns, and it will be able to categorize them as long as there is enough usage evidence in the corpus. This allows it to create categories of baroque composers, as well as correctly categorize an item that appears in the corpus as a sequence of Chinese characters as a “delicacy” as long as there is context evidence in the corpus.

The strengths of FFID compared to Castanet are also its weaknesses. Because FFID clusters terms according to the context in which they appear, if the corpus does not contain enough sentences where the terms are used FFID cannot cluster the terms. On the other hand Castanet works quite well as long as the terms can be found in WordNet. So a test of FFID that used a corpus consisting of names of scientific journals yields almost no meaningful facets because the corpus does not contain enough sentences where the terms are used. Castanet, on the other hand, is able to handle such a corpus successfully as long as the terms appear in WordNet.

CONCLUSIONS

This paper describes an algorithm that takes as input a large (hundreds of megabytes) corpus of text documents and generates automatically in a short time (seconds) a set of facets that can be used by knowledge engineers to build faceted hierarchies that facilitate classification, user browsing, and searching. The central hypothesis is that multi-word terms that appear in the corpus in similar grammatical contexts are likely to belong to the same facet. For example, the names of music composers are likely to appear in similar contexts which are different to the grammatical contexts in which the names of musical instruments appear. This technique allows the system to cluster together multiword terms without having any domain information. Since the system uses an English part of speech tagger and knowledge of English grammar, some of its modules would have to be changed in order for it to work for corpora in different languages.

A comparison of the system described with a similar state of the art system (Hearst’s Castanet) reveals that both systems are different and complementary. Each one finds facets that the other one could not find.

The tests reported in this paper suggest that the facets discovered by the system are useful for knowledge engineers and document indexers.

FUTURE WORK

The output of the present system seems to be considered significantly useful by indexers and knowledge engineers. There are many ways in which it could be made better. New and better patterns could be tested using the framework created for the present system. More specific improvements could be obtained in the following areas:

(1) generate more contexts for each sentence. The experiments described here used a transformation that uses just one context for each term: two terms to the left and two terms to the right of the term. Generating more contexts would be equivalent to having a more complex context similarity function.

(2) refine the normalization patterns described in this paper and try new ones.

(3) eliminate the patterns that the tests show not to produce as many high quality clusters as others.

(4) a new possible tool based on the same techniques described here could allow knowledge engineers to enter a multi word term and have the system look for other multi-word terms that occur in similar contexts. For example look for terms “similar” to “hot chocolate”.

(5) given a set of multi word terms that have been judged to be values of the same facet the system could gather all the contexts in which they appear and from these contexts synthesize a pattern that would have discovered those terms as well as other terms that might fall in the same facet.

(6) during the cluster merging phase it might be possible to identify some relationships between clusters. For example: when merging clusters that differ only in an adjective, as is the case between “french/jj composers/nns” and “italian/jj composers/nns”, the system could build a larger “Composers” clusters (as it does now) while recording the information that a “French composer” and an “Italian composer” is a composer (which is not done in the present version).

ACKNOWLEDGMENTS

The author would like to thank Marti Hearst for sending the data needed to compare her approach with ours, as well as her willingness to answer our questions. James Anderson read earlier versions and provided, as usual, valuable advice.

REFERENCES

Anderson, J. D. & Perez-Carballo, J. (2005). *Information retrieval design*. St. Petersburg., Fl: University Publishing Solutions.

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, 5, 133–143.

Belkin, N. J. and Croft, B. (1992). Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM*, 35(12), 29–38.

Brill, E. (1995). Unsupervised learning of disambiguation rules for part of speech tagging. *Proceedings of the 3rd Workshop on Very Large Corpora*, 1-13.

English, J., Hearst, M. A., Sinha, R., Swearingen, K., & Yee, K.-P. (2002). Hierarchical faceted metadata in site search interfaces. *Conference on Human Factors in Computing Systems,* 628-639.

Hearst, M.A. (1999). User Interfaces and Visualization. In R. Baeza-Yates & B. Ribeiro-Neto (eds.) *Modern Information Retrieval.* (pp 257-323) New York: ACM Press.

Hearst, M.A. (1998). Automated discovery of WordNet relations. In Christiane Fellbaum (ed.), An Electronic Lexical Database and Some of Its Applications. (pp 131-151), Cambridge, MA: MIT Press.

Hearst, M.A. (2006) Clustering versus faceted categories for information exploration. *Communications of the ACM,* 49(4), 59-61

Markoff, J. (2005, August 15). In silicon valley, a debate over the size of the web. *The New York Times*, Technology section.

Rivadeneira, W. & Bederson, B. (2003). A study of search result clustering interfaces: Comparing textual and zoom able user interfaces. *HCIL technical report*, (Technical Report HCIL-2003-36).

Stoica, E. & Hearst, M. A. (2004). Nearly-automated metadata hierarchy creation. *The Companion Proceedings of HLT-NAACL'04.* 117-120.

Stoica, E., Hearst, M. A., & Richardson, M. (2007). Automating creation of hierarchical faceted metadata structures. *Proceedings of NAACL HLT*. 244–251.

Svenonius, E. (1992). Ranganathan and classification science. *Libra*, 17(42):176–183.

Taylor, R. (1968). Question-negotiation and information seeking. *College and Research Libraries*, 29(3):178–194.

Venkatsubramanyan, S. and Perez-Carballo, J. (2004). Multiword expression filtering for building knowledge maps. *2nd ACL Workshop on Multiword Expressions: Integrating Processing*, 40–47.

Venkatsubramanyan, S. and Perez-Carballo, J. (2007). Techniques for organizing and presenting search results: A survey. *Journal of Information Science and Technology (JASIST)*, 4(2). 45-

Xie, I. & Cool, C. Understanding help seeking within the context of searching digital libraries. *Journal of the American Society for Information Science and Technology (JASIST)*, 60(3) (2009) 477-494.

Yee, K.P., Swearingen, K., Li, K., and Hearst, M. A. (2003). Faceted metadata for image search and browsing. *Proceedings of the SIGCHI conference on Human factors in Computing.* 401-408.

FUNCTIONAL REQUIREMENTS FOR SECURE CODE: THE REFERENCE MONITOR AND USE CASE

Ken Trimmer, Idaho State University
Kevin R. Parker, Idaho State University
Corey Schou, Idaho State University

ABSTRACT

Information assurance, data security, and corresponding issues are traditionally presented in Systems Analysis and Design textbooks as non-functional requirements. Systems analysts can enforce secure design and code as one of the essential goals of systems analysis and design by using the Reference Monitor concept as a means of requirements and design specification. The application of the Reference Monitor during the early stages of systems requirements specification via the Use Case emphasizes that information assurance is a critical functional requirement.

INTRODUCTION

Failure to incorporate security into systems requirements is a concern dating back at least a quarter of a century (Schell, Downey & Popek, 1973, Pipkin, 2000). Compounding this oversight is the lack of attention paid to security in textbooks and the exclusion of security as a functional requirement (Haworth, 2002, Trimmer, Parker & Schou, 2007).

The lack of ubiquitous system security requirements yields the 'penetrate and patch' strategy for secure code maintenance. This strategy, in addition to being costly to enforce and a source of vulnerabilities, may compromise an organization's system resources and corresponding operations when considered from an Information Assurance (IA) perspective (Schou, Trimmer & Parker, 2005).

The pervasive use of data by those both internal and external to an organization has led to Information Systems (IS) becoming a component of the organization's communications infrastructure, much as the fax and the telephone were before the broad adoption of personal computers at all levels of organizations. Once the telephone became an integral component of organizations, certain functions became dependent upon it, such as the ability to quickly place or receive orders from someone not physically located at the organization. Fax machines extended this, as they enabled orders to vendors and from customers to contain considerable detail about multiple items that may have been more difficult to clearly communicate via verbal telephone communications.

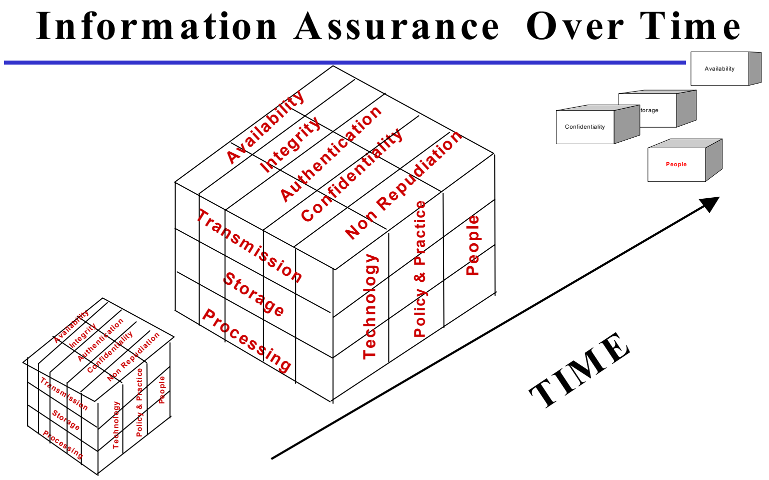
Electronic Data Interchange and e-commerce via the World Wide Web have escalated functional dependence upon IS. Furthermore, the emergence of ‘knowledge workers’ in organizations would be non-existent without IS. It is clear that the modern organization cannot exist in its evolved form without an IS. Further, the IS is unable to provide the necessary support for the dependent organizational functions unless the underlying principles of IA are considered and in place.

INFORMATION ASSURANCE

Information Assurance is an extension of computer security and information security processes. It encompasses the entire lifecycle of data and information from project inception to the retirement of the system and its contents. Because of the underlying design complexity of secure systems, security and information assurance are typically late binding design functions, if considered at all in the design phase (Schou, et al., 2005).

IA is both multidisciplinary and multidimensional. This was identified by McCumber in the representation of his model for computer security (McCumber, 1991). Spurred by the growth of the World Wide Web and e-commerce in the late 1990s, Maconachy, Schou, Ragsdale, and Welch (2001) developed the MSR model by extending McCumber’s robust information assurance model to include time as a fourth dimension, adding to Information States, Security Services, and Security Countermeasures.

Figure 1, Information Assurance as represented by Maconachy et al., 2001



Also in 2001, Maconachy et al. extended the basic information service dimensions of availability, integrity, and confidentiality with the additional dimensions of authentication and non-repudiation. The additions by Maconachy and his associates are displayed in Figure 1 (Maconachy et al., 2001). In 2008 the Joint Task Force on Computing Curricula adopted the MSR model as part of the information technology model curricula for information assurance and security (ACM, 2008, p. 73-74).

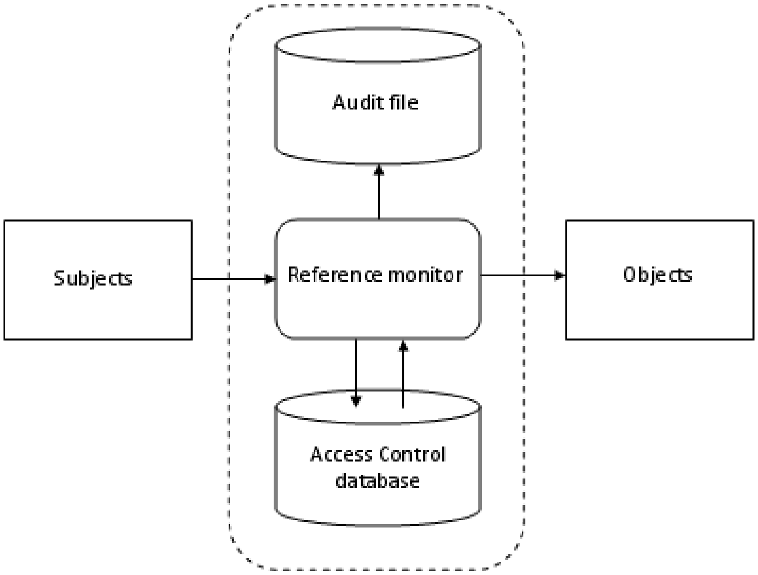
REFERENCE MONITOR

Irvine (1999, p. 3) in referencing Anderson (1972) defines the Reference Monitor (RM) as having the following requirements:

- ♦ *The access mediation mechanism is always invoked.*
- ♦ *The access mediation mechanism is tamperproof.*
- ♦ *It "must be small enough to be subject to analysis and tests, the completeness of which can be assured".*

Irvine continues her discussion of the RM, addressing the need to consider it in systems requirements as it is a broad tool that enables the systems analyst to identify abstract requirements. Trimmer et al. (2007) provide a similar argument, making a distinction between a Requirements RM and a Design RM, to be addressed as broad systems requirements and incorporated into the initial design. The incorporation of both RMs is to be performed regardless of the specific development methodology employed.

Figure 2, Concept of Reference Monitor, after Cho et al. (2008)



Furthermore, satisfying RM requirements is a basic security component of both mandatory and verified protection for software that satisfies US Department of Defense requirements for secure system controls. The RM is a component of the Trusted Computer System Evaluation Criteria (Department of Defense, 1985).

Correspondence between the service dimensions of authentication and non-repudiation in the model shown in Figure 1 and the RM are represented by Cho, Moon, and Baik (2008). This concept is displayed in Figure 2. The dotted line has been added to show the integral nature of the RM – it is implicit that it is self contained. In this representation, the subject authenticates through the RM, which uses its integral Access Control Database. Provided the subject has rights given their authentication, s/he is granted access to the objects. This access is recorded in the integral Audit File to support non-repudiation.

USE CASE

The Use Case, a component of the Unified Modeling Language (UML) (Satzinger, Jackson & Burd, 2005), represents a user as existing outside a system, making requests to the system. Traditionally, this corresponded to individuals within an organization who required specific system support to carry out their organizational functions. The advent of the Internet in business commerce further complicated the process, as supply chain enablement permits both suppliers and customers to remotely interact with the system. The Use Case serves as a requirements gathering tool not only for the UML methodology, but also for more traditional analysis and design modeling tools such as Data Flow Diagrams and Entity Relationship Diagrams (Whitten, Bentley & Dittman, 2004).

The representation of the RM by Cho et al. (2008) in Figure 2 is illustrative of applying the RM concepts as an underlying condition for the Use Case. Cho et al. (2008) use the RM as the focal point for an end user's home gateway model. They provide a user scenario as an example of user access and maintenance of a temperature control service. This scenario also presents the events in another type of UML Diagram, the Sequence Diagram.

The following discussion addresses the use of the RM, using the representation in Figure 1. This discussion focuses on the general accessing of information by an employee (A) and an online consumer (B). In both scenarios the user gains access to the information through a web portal.

A. *An employee of an agricultural firm is out of town, and needs to process an expense transaction and check on the status of a prior request. The employee gains access to the Internet via a secure WiFi at their hotel, and proceeds to the corporate website. After selecting 'Secure Login', the employee enters their user identification and password. This is passed to the RM, which provides access to the employee's web page, which provides access to only those corporate information resources to which the employee has rights. The employee selects Expense Transactions, and the RM is again engaged,* *providing the employee read/write access to New Transactions and read - only access to Transaction Status and Transaction Reports for Expense items.*

B. *A consumer Googles a product made by the same agricultural firm and finds that they can place a direct order of $1000 or more without going through a distributor. The consumer selects the product and places it in a Shopping Cart. At this point, the consumer has only read access to an online catalog. When the consumer is done selecting products and quantities and makes the 'Purchase' request, they are led to a site that asks if they are a registered user or if they would like to proceed as a guest. If they are a registered user, they will be asked for a user name or email address and password. The RM will then be invoked and the consumer will be authorized to proceed to the transaction and gain access to a set of choices similar to those seen by the Employee in scenario A, with corresponding read/write and read-only privileges. The Guest will be taken to a screen that will allow them to write one and only one transaction.*

In both cases, the RM validates the user authentication and records a corresponding transaction unseen by the user. Access to the system is necessary for either user to perform their corresponding functions. Furthermore, as discussed by Cho, et al. (2008), the RM also checks for user services, thereby calling into play an additional component of Figure 1, Availability. Another characteristic addressed in Figure 1 is Integrity, as the user has write access to only new transactions. The final of the original three dimensions in the McCumber Cube, Confidentiality, is also addressed by the RM in that only those employees, groups of employees, and customers performing functions are granted rights to certain data elements and applications.

The corresponding Use Cases must contain an "Includes" of the RM by each specific process requested by the user. By considering the RM during requirements modeling, the underlying data elements and processes will enable the user to complete the specific tasks associated with their function either internal or external to an organization. Although it can be argued that in times of system outage a user could resort to manual systems to perform their function, such actions could lead to a compromise of system integrity and corresponding user functions and should be discouraged. The role of the RM and IA is even more necessary for the completion of the knowledge management functions in the modern organization.

CONCLUSION

Systems designers must begin to incorporate secure code concepts throughout the analysis and design process. By requiring that the concept of the Reference Monitor be considered as a functional requirement in Use Case Diagrams, the designer will incorporate authentication and non-repudiation throughout the systems development life cycle, regardless of the methodology chosen. By including the RM, the designer will be forced to consider the related information characteristics of availability, integrity, and confidentiality under the umbrella of Information Assurance – critical functional requirements for the modern organization.

REFERENCES

ACM. (2008). ACM Computing Curricula Draft, Information Technology Volume. Retrieved December 21, 2008, from http://campus.acm.org/public/comments/it-curriculum-draft-may-2008.pdf

Anderson, J. P. (1972). Computer Security Technology Planning Study. *Technical Report ESD-TR-73-51*, Air Force Electronic Systems Division, Hanscom AFB, Bedford, MA.

Cho, E., C. Moon & D. Baik (2008). Home Gateway Operating Model using Reference Monitor for Enhanced User Comfort and Privacy. *IEEE Transactions on Consumer Electronics, 54*(2), 494-500.

Department of Defense (1985). Trusted Computer System Evaluation Criteria. DoD 5200.28-STD.

Haworth, D. (2002). Security Scenarios in Analysis and Design, The SANS Institute.

Irvine, C. E. (1999). The Reference Monitor Concept as a Unifying Principle in Computer Security Education. *Proceedings of the IFIP TC11 WG 11.8 First World Conference on Information Security Education*, 27-37.

Maconachy, W. V., C.D. Schou, D. Ragsdale & D. Welch (2001). A Model for Information Assurance: An Integrated Approach. *Proceedings of the 2001 IEEE Workshop on Information Assurance and Security*, 306-310.

McCumber, J. (1991). Information Systems Security: A Comprehensive Model. *Proceedings 14th National Computer Security Conference*. 328-337.

Pipkin, D. (2000). *Information Security: Protecting the Global Enterpris*e. Upper Saddle River, N.J.: Prentice Hall PTR.

Satzinger, J.W., R.B. Jackson & S.D. Burd (2005). *Object-Oriented Analysis & Design with the Unified Process.* Boston, MA: Thomson/Course Technology.

Schell, R.R., P.J. Downey & G.J. Popek (1973). Preliminary Notes on the Design of Secure Military Computer Systems, MCI-73-1, The MITRE Corporation, Bedford, MA 01730. Retrieved December 21, 2008, from http://seclab.cs.ucdavis.edu/projects/history/CD/index.html#sche73

Schou, C., K. Trimmer & K.R. Parker (2005). Forcing Early Binding of Security Using a Design Reference Monitor Concept in Systems Analysis and Design Courses. *Proceedings of the International Conference on Informatics Education and Research*, 321-331.

Trimmer, K., K.R. Parker & C. Schou (2007). Forcing Early Implementation of Information Assurance Precepts throughout the Design Phase. *Journal of Informatics Education Research, 9*(1), 95-120.

Whitten, J.L., L.D. Bentley & K.C. Dittman (2004). *Systems Analysis and Design Methods (Sixth Edition)*. Boston, MA: McGraw-Hill/Irwin.