大葉大學 九十三 學年度　研究所博士班　招生考試試題紙

系所別	組別	考試科目（中文名稱）	考試日期	節次	備註
管理研究所博士班	乙	論文評述	6月21日	第二節	共乙頁

註：考生可否攜帶計算機或其他資料作答，請在備註欄註明（如未註明，一律不准攜帶）

請就所提供的三篇論文中，選擇一篇閱讀並回答下列問題**【請務必於答案卷中，註記作答的論文編號，否則以零分計算】**：

1.請說明該篇論文的主要研究目的、所要解決的問題以及所提出的解決問題的方法【請務必自行整理說明】。(30%)

2.請說明該篇論文的優點以及缺點，並寫出如何改進該論文的缺點。如果沒有缺點，則寫出為什麼沒有缺點的理由【請務必自行整理說明】。(40%)

3.請就整篇論文內容訂定合適它的英文題目名稱，並用英文寫出約200-300字左右的英文摘要。(30%)

編號：1

Introduction

It has long been recognized that the effectiveness of decision making is influenced by many factors. Among these are the time available before the decision must be rendered, the experience of the decision maker, and the quality of the data needed for the decision. Although ideally the data used should be of high quality, in practice this often is not the case, for reasons that range from the cost of obtaining quality data to the inherent difficulty or even impossibility of doing so for certain data types. Nevertheless, experienced decision makers, especially ones who have worked in a particular milieu for a sufficient period of time, develop a feel for the nuances and eccentricities of the data used and intuitively compensate for them. As organizations increasingly move to stored repositories such as data warehouses, this intuitive feel is not preserved for many who extract data from such sources to support their particular needs.

One solution would be to capture some of the knowledge regarding the data's quality along with the actual data values. Data tagging to provide information regarding the data has long been proposed (Wang and Madnick 1990); however, it is not clear how or if decision makers would use this data quality information. Chengalur-Smith et al. (1999) define data quality information (DQI) to be metadata that addresses the data's quality. Clearly, any benefits that accrue from providing information about the quality of the data must outweigh the cost of obtaining and maintaining this metadata. Although logic dictates that DQI would be of benefit, it is also plausible that the benefit of such information would vary considerably depending upon the circumstances.

The effect of providing data quality tags on decision making in controlled environments was explored by Chengalur-Smith et al. (1997, 1998, 1999). They determined that DQI does indeed influence the decision made, but also found differences in the amount of influence based upon factors such as complexity of task, decision strategy, and format of the DQI. Unfortunately, their work provides no guidance as to when it would be appropriate to include DQI in databases. For this, it is necessary to study the behavior of actual decision makers vis-à-vis DQI. Experienced decision makers might find DQI of greater value than inexperienced ones. Also, one would suppose that time constraints (time available for the decision) and time pressures (perceptions about time constraints) would impact the use of DQI. The purpose of this paper is to explore the use of DQI by decision makers, who have various experience levels and who are required to make decisions under various time conditions, to determine who would most benefit from the inclusion of DQI and under what circumstances. Information of this sort would guide data managers as to whether to include DQI in databases.

Background and Research Model

The diverse uses of data and the increased sharing of data that has arisen as a result of the widespread introduction of data warehouses have exacerbated deficiencies with the quality of data (Ballou and Tayi 1999, Haisten 1995). Researchers have identified many facets or dimensions of data quality such as accuracy, timeliness, completeness, consistency, relevance, and so forth (Ballou and Pazer 1995, Wang and Strong 1996). Wang and Strong (1996) investigated characteristics of data quality from the data consumer perspective and found that of the various dimensions of data quality, data accuracy (used interchangeably with reliability) was most important. For this study, we chose to focus on a single dimension of data quality, namely accuracy. A review of the field of data and information quality research may be found in Wand and Wang (1996) and Wang et al. (1995).

Decision Processing

Decision making is a response to problems where the problems include choices from among a set of alternatives (Kingma 1996). Over the last half-century significant research has been conducted on decision-making processes and strategies. Simon (1957) defined a rational model of decision making in which decision makers consider all aspects of all alternatives before making a decision. However actual decision making often falls short of the rational ideal (March and Simon 1958). Some of the reasons that decision making falls short of the ideal is that knowledge is incomplete, experience of the consequence is incomplete, there is limited amount of time to explore all alternatives, and humans do not calculate perfectly (March and Simon 1958). These factors have influenced our choice of factors to consider in this study.

Payne et al. (1993) identified seven decision-making strategies and combinations thereof. These strategies may be grouped into two fundamental types—weighted additive and conjunctive. In weighted additive decision making, weights are assigned to each attribute and scores are assigned based on how closely an attribute matches the goals of the decision problem. For each alternative, each attribute is evaluated and the resulting score is multiplied by the weight. These values are then summed to produce an overall score, and the alternative with the largest score is chosen. This summation and weighting process allows for an alternative to be chosen that may have weak scores on some attributes, but high scores on other attributes. In contrast, the conjunctive strategy sets minimum acceptable levels for each attribute. If an alternative has a low score on even one attribute that alternative will be rejected.

In the bounded rationality model (March and Simon 1958), decision makers look for heuristics to reduce task demands. When making a decision, individuals will use a compromise strategy that minimizes their cognitive effort (Payne et al. 1993) and will

1-2

ignore less relevant information in complex problems (Grether et al. 1986). Thus the nonuse of DQI, if DQI is not recognized as relevant, may be a function of the level of task complexity.

Task

The degree of task complexity is implied by the number of cells in the decision space that is constructed by building a matrix of decision choices (alternatives) and decision criteria (attributes). Prior research has indicated that 20 cells represent a relatively simple task, while complex tasks may have as many as 40, 60, or 80 cells (Payne et al. 1993).

Chengalur-Smith et al. (1997, 1998, 1999) used a 20-cell apartment selection task to study the use of DQI in a simple task setting. For a complex task they used a business site selection task with 42 cells; their subjects had no problem with these levels of complexity. The subjects were college seniors working under the same time constraints. (By contrast, our study uses novices, experienced managers, and professionals under various time conditions.) Morrow et al. (1992) stated that experience is only important when a task is difficult enough to call on the domain-relevant knowledge; therefore, it was important to develop a task that was more complex than the one that was accomplished by college seniors.

A newly developed job transfer task with 63 cells fits these requirements. This task describes seven alternative jobs along with nine attributes of each job. Subjects are requested to rank the jobs according to predefined weights and scores of each attribute. (See Appendix A). This task provides the ability to examine results by domain-specific experience as well as by general experience. Some people have not changed jobs, some have changed jobs once or twice, and some have changed jobs multiple times.

Decision Outcome

In a multiattribute decision-making task an actor chooses one alternative from among many alternatives or ranks all alternatives from most preferred to least preferred. Each alternative is described by several attributes. The values of these attributes form the basis for the actor's decision, such as buying a car based on attributes such as price, safety record, and maintenance record. The choice of one alternative among many alternatives represents the decision outcome. Varying the value of attributes changes the decision outcome to the degree that actors use the attributes in their decision making.

To study the effect of a new attribute, e.g., consumer rating, researchers may compare decisions made by several people without consumer rating to decisions made by several people with consumer rating. If these groups of people are randomly established and the only difference in the multiattribute task is the existence of the consumer rating, then any difference in the decision choices from one group to the next may be attributable to use of consumer rating in the decision process. When the addition of a new variable does not lead decision makers to make a new decision, then it is said that the decision makers were *complacent* to the new variable. When the addition of a new variable leads to a new decision, the decision makers were not complacent with respect to the new variable. If the decision makers were not complacent but there was a scattering of multiple new first choices, then there is less *consensus*. If the decision makers change their order of rankings, then there is less *consistency*.

To operationalize *decision outcome* we employ three measures of the impact of DQI (Chengalur-Smith et al. 1997, 1998, 1999). These are: *Complacency* (a measure of the lack of impact of DQI), *consensus* (a measure of agreement on the top choice in the presence of DQI), and *consistency* (a measure of the degree to which the overall rankings are not affected by DQI). These three variables are defined in the context of individual decision making involving decision problems with multiattribute alternatives. Complacency and consensus consider changes in the top-ranked alternative, whereas consistency considers changes in the ranking of all alternatives. The three measures explore different aspects of the impact of DQI, but are not independent.

Complacency is the proportion of people in one group who choose a specific alternative for their first choice as compared to the proportion of people in a second group who choose the same alternative for their first choice. To measure complacency, we identify the top-ranked alternative (say Alternative B) for

1-3

a group without DQI and record its frequency. We then count the number of subjects that identified the same alternative (here B) among a group with DQI. Because we are examining proportions and dealing with categorical data, the chi-square test of homogeneity is appropriate (Sheskin 2000). A significant chi-square indicates that the groups differed due to the influence of DQI and thus were not complacent, a desirable outcome (Table 1).

Lack of complacency indicates that the people with DQI used it. However, people may use DQI differently, which brings us to consensus. Consensus is similar to complacency in that it compares proportions of people in two groups as to their most-preferred alternative. Consensus differs from complacency in that the most-preferred alternative may be different in each group. Differences in the number of times the top-ranked site is selected by members of the two groups are compared using chi-squared statistics. A significant chi-squared value indicates a change in consensus, implying that DQI either detracted from or enhanced a group's ability to reach a decision. Or, a significant chi-square indicates a change in the level of agreement, which could be a result of increased uniformity or decreased uniformity. Thus, complacency alone does not provide a complete measure of the impact of DQI. Note that complacency and consensus may be in a hierarchical relationship to each other; i.e., consensus should only be considered after non-complacency is established.

In some cases the focus may not be on just the top-ranked alternative. Decision consistency refers to the rankings of all alternatives from the most preferred to the least preferred. Thus, consistency can be considered an extension of complacency because it considers the entire set of rankings instead of just the top-ranked alternative. Consistency indicates that DQI did not influence the decision. To measure consistency, a correlation is performed between the two groups of average rankings for each alternative, with and without DQI. A significant correlation between one group's rankings and another group's rankings implies consistency between the groups' results. Low correlation implies that DQI caused a difference in the overall rankings.

Table 1 Ideal Values for Measures of DQI

Measure	Ideal	Implication
Complacency	Low (high chi-square)	DQI was used—different first choice
Consensus	High (low chi-square)	DQI did not change level of agreement
Consistency	Low (low correlation)	DQI was used—rankings varied

Experience Level

Gilliland et al. (1994, p. 406) state that there has been "relative lack of attention to the study of prior knowledge in the decision-making literature." One would expect that experience is an important variable to study in decision making, but there are conflicting possibilities regarding its significance. Some researchers state that experience may improve performance in decision making, while others state the opposite.

Experience may improve performance because it increases alertness to errors (Klein et al. 1997), sensitivity to omissions (Sanbonmatsu et al. 1992), use of relevant information (Sanbonmatsu et al. 1992), adaptation to subtle contextual differences (Payne et al. 1993), ability to identify important features of a problem (Mackay and Elam 1992), ability to organize the information better (Mao and Benbasat 2000), ability to attend to greater amounts of knowledge (Mao and Benbasat 2000) and process it more extensively (Sanbonmatsu et al. 1992). The benefits of experience may be attributable to domain-specific knowledge (Morrow et al. 1992). Mao and Benbasat (2000) stated that domain-specific knowledge is a critical factor in *reading-comprehension* studies, while Shaft and Vessey (1995) said domain knowledge aids computer programming comprehension. For these reasons it appears that experts would make more use of DQI than novices.

However, there are potential dangers in assuming that an intuitive feel for the data is always positive. Experience may not influence accuracy (Paese and Sniezek 1991), prior experience influences beliefs and expectations about data (Klein et al. 1997), and may truncate the decision process early (Dukerich and Nichols 1991, Hall 1991). A novice may be more attentive to new information (such as DQI) than an expert (Yates et al. 1991). For example, in a business

1-4

relocation task, people without task-specific experience gave more accurate ratings than did experienced people (Gilliland et al. 1994).

Earlier studies did not have a uniform definition of experience. Chengalur-Smith et al. (1999) found little impact of DQI in complex task scenarios with undergraduate students, and therefore called for research using more experienced people as subjects. Klein (1997) used a mix of graduate and undergraduate students. Mackay and Elam (1992) used professional employees only, and Yates et al. (1991) and Gilliland et al. (1994) used a combination of novices and domain experts. Given the mixed results of these prior studies, we investigated multiple levels of experience, including novice (college freshmen) vs. expert (working MIS professionals); experts working less than or equal to 10 years vs. experts working greater than 10 years; domain-specific experience vs. nondomain-specific experience; and managerial vs. nonmanagerial experience.

Time

Researchers typically study decision making without time constraints (Ordonez and Benson 1997, Payne et al. 1993). Some researchers (e.g., Ahituv et al. 1998, Morrow et al. 1992) have studied "time pressure" but measured "time pressure" by simply allocating specific time to perform a task. We distinguish between time constraints and time pressure. A time constraint is a specific allotment of time for making a decision, while time pressure is a subjective reaction to the amount of time allotted. Time pressure is experienced whenever the time available for the completion of a task is *perceived* as being shorter than normally required for the activity (Svenson and Edland 1987). Some people may *feel* pressure in a long time constraint while others may not *feel* time pressure in a short time constraint (see Figure 5 in this paper).

While time pressure affects decision processes (Payne et al. 1993), there are some mixed results as to the effects of time on decision making. Some say that time pressure decreases decision accuracy (Zakay and Wooler 1984), while others say that increasing time pressure may increase quality in software development projects (Austin 2001). Time constraints may have more impact on decision making for novices than for the sophisticated decision makers (Dukerich and Nichols 1991). Ahituv et al. (1998) found that time pressure impaired the performance of middle-level field commanders more than it affected top-level commanders. Payne et al. (1993) found that an increase in domain-specific experience level under time pressure improved performance. The time factors, constraint or pressure, have not been studied in relationship to DQI and experience levels until this present study.

In Figure 1, the rectangles represent constructs, the double ovals represent measures (variables), and the ovals represent the values of the measures or constructs. It should be noted that many of the independent variables displayed in Figure 1 take on multiple values; e.g., the variable pressure has two values (felt time pressure and did not feel time pressure). For the sake of simplicity we did not display all of these. Our primary focus is on the exploration of the impact of DQI, together with various facets of time and experience on decision outcome. As seen in Figure 1, we place time and experience in the context of task complexity and also consider three demographic variables: Age, gender, and education.

Hypotheses

The hypotheses that we explore are motivated by the research described earlier and focus on the time and experience aspects of the research model given in Figure 1. Experience and time are investigated within the context of task complexity. Later, we briefly discuss the role of demographics.

HYPOTHESIS 1 (EXPERIENCE). *For a given value of experience, there will be no differences in decision outcome between subjects with and without DQI.*

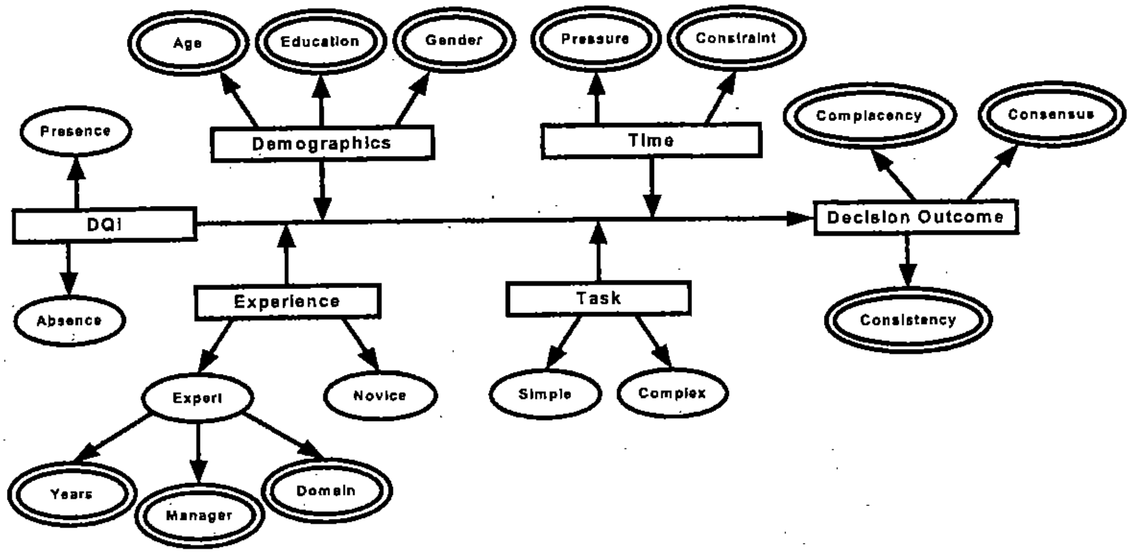
In the case of experts, this null hypothesis will be tested using the different variables describing expertise that are shown in Figure 1.

HYPOTHESIS 2 (TIME). *For a given measure of time, there will be no differences in decision outcome between subjects with and without DQI.*

This null hypothesis will be tested for different values of the measures of time shown in Figure 1.

1-5

Figure 1 Research Model



Research Method

We performed two experiments to explore the effect of providing DQI in multiattribute decision-making tasks of varying complexity, using subjects with various experience levels and assigned to different time constraints. For the DQI we used interval numbers ranging from 0 to 1, where 1 implies the highest-possible quality and 0 implies no quality.

Pilot

To test our procedures and determine appropriate time constraints, we conducted pilot tests and recorded the amount of time required to complete the tasks. The pilot studies provided feedback as to the usability and clarity of the questionnaire instruments, and the intelligibility and consistency of the tasks and procedures.

The pilot study for the simple task included 17 undergraduate students. They took a mean time of 11.2 minutes and a median of 12 minutes to complete this task. The standard deviation was three minutes. All subjects in the pilot finished in less than 40 minutes. To create a short time constraint, we employed the shorter of the Median Method (van Bruggan et al. 1998) or the Mean Method (Ordonez and Benson 1997). The subtraction of one standard deviation from the mean yielded a short time of eight minutes. The short time constraint for the simple task was set at eight minutes and the long time constraint at one hour.

The pilot study for the complex task was conducted with 11 graduate students in a systems design course. Eight of the subjects were full-time employees within the community, while three were full-time students who completed at least one semester of internship and one full year of graduate school. The mean time required to complete the complex task was 24.2 minutes and the standard deviation was eight minutes. The longest completion time was 35 minutes. Because one of the primary goals of this research was to explore the effect of time, we employed three different time constraints with the expert subjects. The short time constraint for the complex task was set at 15 minutes, the medium time constraint at 25 minutes, and the long time constraint at 45 minutes. The subjects reported that there were no ambiguities in either the task or the questionnaire.

Experiment 1

Subjects. We used two groups of subjects: Novices and experts. There were 118 novices who were students majoring in computer science, information systems, or information technology and enrolled in a freshman seminar course. There were 21 females and

1–6

97 males. There were 38 experts who were professional employees in an information systems organization at a major international service company. This random group had 19 males and 19 females; 27 had managerial experience while 11 did not; 6 had high school education, 1 had two years of college, 25 had a bachelor's degree, and 6 completed a master's degree. For the context of our experiments there is face validity to our assertion that the IS professionals had more experience than the freshmen. This assertion was confirmed in our questionnaire (see Question 7, Appendix B) relating to the apartment selection task. Most of the IS professionals had experience in choosing their own apartments, but only a handful of freshmen had ever chosen an apartment on their own.

All subjects were volunteers and received no pay or credit for this activity.

Task. A simple task with only 20 cells required the subjects to select an apartment from among four alternative apartments, based on 5 criteria (Payne et al. 1993, Chengalur-Smith et al. 1999). The five criteria were weighted and scored for each apartment; the weighted scores were provided for each attribute, but were not summed for each alternative. Two forms of the simple task were used, one without data quality information and one with data quality information. All subjects involved with a particular task were given the same numbers for each of the cells in the alternative-criteria matrix. Those with the DQI, however, had additional information as to how reliable some or all of those attributes were. The numbers in the cells and the DQI values were specified so that at an intuitive level at least the inclusion of DQI would be likely to result in different rankings of the alternatives. For instance, Apartment B is very attractive without DQI but much less attractive with DQI. Thus, if the rankings did not change, then the DQI clearly was ignored.

Procedure. Experiment 1 was conducted at two different locations, the national MIS headquarters of an international service company, and on a college campus. Novices and experts were randomly divided into two time groups, short and long, at their respective locations. Within each time constraint, approximately half of the subjects received tasks with no DQI and the other half received tasks with DQI. As each subject completed the task, the moderator collected demographic data on a questionnaire (Appendix B). The procedure was strictly controlled so that the subjects did not view the questionnaire until finished with the task.

Experiment Design—Experiment 2

Subjects. There were 69 experts who were employed in the information systems department of an international service company. All subjects had at least one full year of work experience, while 34 had greater than 10 years of experience and 35 had less than or equal to 10 years of work experience; 17 were less than or equal to 30 years old, while 51 were greater than 30 years old. There were 28 females and 41 males; 40 reported having at least one year of managerial experience, while 29 did not have managerial experience; 24 had experienced at least one job change that required a household move, while 44 did not have a job change that required a household move. All subjects were volunteers and received no pay or credit for this activity.

Task. A "job transfer task" was developed for this study. The goal was to create a complex task that was deemed to be real and interesting to a group of experienced professionals. In addition, we needed to have a common problem with which the professionals had varying degrees of personal experience and knowledge.

The job transfer task is a complex multiattribute decision task with seven alternatives, each described by nine attributes (Appendix A). Subjects are asked to rank job alternatives from most desirable to least desirable, based on criterion attributes that are weighted (prioritized) and scored. Two forms of the job transfer task were used, one without data quality information and one with data quality information. The nine criteria were weighted and scored for each job; the weighted scores were provided for each attribute but were not summed for each alternative. If a subject used the weighted additive decision process, Alternative B would be the preferred outcome without DQI; with DQI, Alternative G would be preferred and Alternative B would be towards the bottom of the list (see Appendix A).

1-7

Procedure. The 69 experts were divided randomly into three time-control groups. There were 21 people in the short time-constraint group, 23 people in the medium time-constraint group, and 25 in the long time-constraint group. Each of these three groups was randomly subdivided into two groups: Those who received tasks with no DQI and those who received tasks with DQI. The task description informed those subjects with DQI of the presence of DQI but did not suggest how to use the DQI. For tasks without DQI there was no mention of DQI.

A posttask questionnaire was used to obtain information to group people by domain-specific experience, years of work experience, management experience, perceptions of time pressure, age, gender, and education.

Results

The primary purpose of this study was to explore the effects of experience and time on the use of DQI in decision making. We first explore the effects of experience on the use of DQI and later explore different combinations of experience and time. Finally, we address some demographic questions.

Experience

Table 2 shows the results of testing Hypothesis 1 based on the simple task. The first row establishes that, in the absence of DQI, novices and experts arrive at similar results when performing the simple task. The nonsignificant chi-square statistics imply that the two groups chose the same alternative and did not differ in their level of consensus. Finally, the significant correlation for consistency suggests that the overall rankings given to the sets of alternatives by the novices and experts were essentially the same.

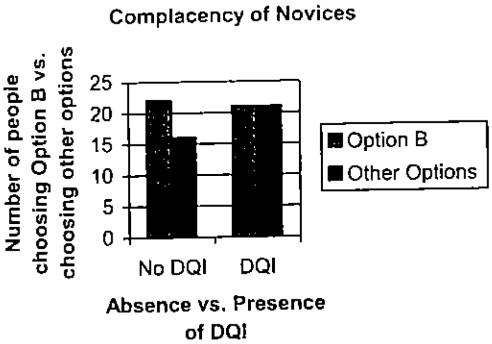
Table 2 Simple Task: Novices Versus Experts

Comparison Groups	Sample Sizes	Complacency	Consensus	Consistency
Novices (no DQI)*	38	$\chi^2 = 0.4$	$\chi^2 = 0.4$	Corr = 0.98
Experts (no DQI)	20	(ns)	(ns)	($p < 0.05$)
Novices (no DQI)	38	$\chi^2 = 1$	$\chi^2 = 1$	Corr = 0.99
Novices (DQI)	42	(ns)	(ns)	($p < 0.01$)
Experts (no DQI)	20	$\chi^2 = 10.9$	$\chi^2 = 5.4$	Corr = 0.94
Experts (DQI)	18	($p < 0.01$)	($p < 0.05$)	(ns)

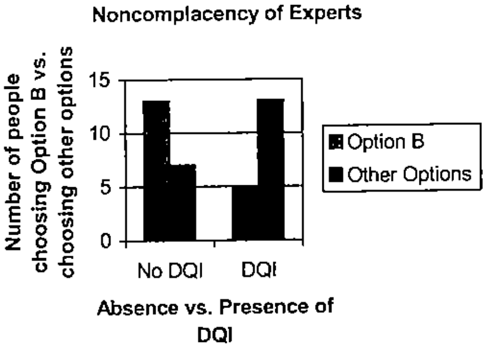
Note. *Technically this is not a complacency measure because neither group had DQI.

The complacency and consistency statistics in the second and third rows show that novices did not use the DQI information in making the apartment selection, while the experts did use the DQI. For a graphical comparison of the complacency of novices versus the noncomplacency of experts, please see Figures 2A and 2B. The first column in each set indicates how many people in that set chose Apartment B as their first choice, while the second column in each set indicates how many people chose one of the other apartments, A, C, or D. The first choice for novices with or without DQI was Apartment B (see Figure 2A). Recall that if DQI was given, Apartment B should not be the right choice. Figure 2B shows that the experts switched from Apartment B to a different apartment when provided with DQI.

Figure 2 (A) Complacency of Novices; (B) Noncomplacency of Experts



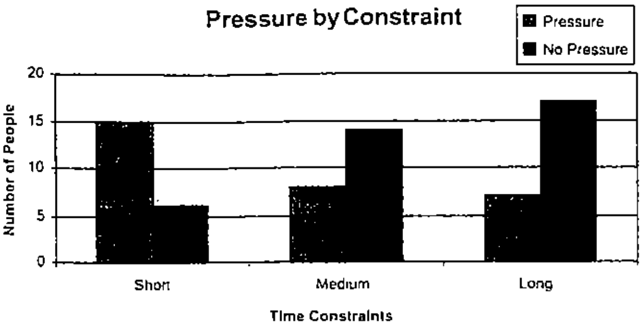
(a)



(b)

1–8

Figure 5 Perception of Time Pressure Within Time Constraints



As seen in Table 8, the feeling of time pressure had more impact on the results than did a particular time-constraint group. The first row shows that the groups with and without time pressure behaved similarly when no DQI was available. However, when DQI was available there were differences. Whether or not subjects felt time pressure, they were noncomplacent, but also had difficulty coming to a consensus. A direct comparison of the top choices made by the groups who felt time pressure and those who did not yielded a chi-square of 9.7 ($p < 0.005$) when both groups were provided with DQI. This, coupled with the larger chi-square for complacency for the group who felt time pressure, shows that those with DQI while under time pressure made the most use of the available DQI. Time pressure was found to affect the overall ranks as well.

Demographic Variables

Using categories based on the posttask questionnaire for the complex task, we continued to explore the degree of complacency based on age, education, and gender. Both male and female experts were not complacent to DQI. We found no differences between males and females when using DQI, as shown by a nonsignificant $\chi^2 = 0.23$. We found some difference based on age. The younger (aged 30 or below) group complacency measure had a $\chi^2 = 12$, while the greater than age 30 group had a $\chi^2 = 367$. A direct comparison of the two groups yielded a $\chi^2 = 6.8$ with $p < 0.005$.

Finally, education was a factor. The high-school-only graduates had the lowest noncomplacency at $\chi^2 = 7$ ($p < 0.05$), the master's degree subjects had the next level of non-complacency at $\chi^2 = 23.4$ ($p < 0.005$), and the bachelor's degree subjects had the highest degree of noncomplacency at $\chi^2 = 231$ ($p < 0.0001$). This leads to the conclusion that college graduates made more use of DQI than high school graduates or postgraduates, even though all were experienced professionals in the business world.

Table 8 Time Pressure (Experts—Complex Task)

Comparison Groups	Sample Sizes	Complacency	Consensus	Consistency
Time pressure (no DQI)	14	$\chi^2 = 0.22$	$\chi^2 = 0.22$	Corr= 0.97
No time pressure (no DQI)*	22	(ns)	(ns)	($p < 0.01$)
Time pressure (no DQI)	14	$\chi^2 = 157$	$\chi^2 = 63$	Corr= 0.3
Time pressure (DQI)	18	($p < 0.001$)	($p < 0.001$)	(ns)
No time pressure (no DQI)	22	$\chi^2 = 86$	$\chi^2 = 0.82$	Corr= 0.89
No time pressure (DQI)	15	($p < 0.001$)	($p < 0.001$)	($p < 0.01$)

Note. *Technically this is not a complacency measure because neither group had DQI.

Discussion

Because incorporating DQI into a database is both time consuming and expensive, it is important to know the characteristics of users who would benefit from having access to such. Our experiments lead us to some preliminary conclusions.

Experience

The results from the experiments provide strong evidence that experts use DQI substantially more than do novices. For the simple task in Experiment 1, experts used the DQI and novices ignored the DQI. For the complex task in Experiment 2, the chi-square statistics are indicative of near certainty that DQI influenced the decision making of experts. However, there was not much difference in the use of DQI based on general level of experience when measured by number of years. However, when we categorized the professionals by type of experience, we found that managers were more likely to use DQI, as compared to nonmanagers. This has implications for data warehouses that are designed to support ad hoc decision making.

While managerial experience increased use of DQI, domain experience did not increase use of DQI. Those

1-12

without domain-specific experience made more use (were less complacent) and better use (had more consensus) of DQI than did those with domain-specific experience. These findings suggest that domain-specific experience may inhibit the use of DQI in decision making. This is consistent with past studies (Gilliland et al. 1994, Yates 1991). Our findings indicate that DQI would be most useful for managers who have little domain-specific experience.

Chengalur-Smith et al. (1999) performed a study using college seniors to perform a simple apartment selection task and a complex restaurant selection task. The college seniors can be said to have an intermediate level of experience—more experience than the current study's freshmen but less experience than the MIS professionals. That study found that for the college seniors "complacency varied dramatically across the research design" (Chengalur-Smith et al. 1999) and that the seniors were not complacent on the simple task but were complacent on the complex task. Our study found novices were completely complacent to DQI, while experts were not complacent on any tasks. An emerging pattern indicates increasing usage of DQI when experience progresses through the stages defined as novice, intermediate, and professional.

The knowledge gained from these findings could be used to guide the selection of people for training in the use of DQI within an organization. As a person progresses from novice status to an experienced professional, there is most likely some cutoff point at which he or she begins to pay attention to DQI. However, if one considers complacency only, one might wrongly conclude that people who are beyond the cutoff point might not benefit from training in the use of DQI.

While complacency illustrates that DQI influences decision making of experts, the consensus measures show that experts do not use DQI in the same way; that is, experts who use DQI will make a variety of choices for the first choice. The consensus chi-squared statistics in most cases showed a decline in consensus when DQI was considered. It may not be desirable for an organization to have a situation where different people make different decisions based upon the same data. For an example of lack of consensus that also brings in the concept of domain-specific experience vs. general experience and time pressure, see the USS Vincennes Case (Fisher and Kingma 2001). Because DQI is such a new concept, we feel that training and/or planning sessions concerning the use of DQI by experts may be very worthwhile.

Time

Two facets of time were considered. The first was time constraint, in which people were put into groups that were allotted a fixed amount of time to complete the tasks. The second facet of time was time pressure, which reflects how people felt about the time they are given to complete a task. Some people in the long time-constraint group felt time pressure, while others in the short time-constraint group did not feel time pressure.

The simple task, when performed by experts given a short time period, led to general consensus. However, providing DQI along with a large amount of time to perform the simple task led to divergent choices or a decrease in consensus. The experts generally were not complacent in the presence of DQI, but time constraints were not a factor for the experts performing the complex task. We found no differences based on actual time constraints, but our data revealed that perceived time pressure did make a difference. Based on self-reports of time pressure, providing DQI led to significant differences in the decision choices between those who experienced time pressure and those who did not. Our data indicates that decision makers who feel time pressure would benefit from having DQI available, as the availability of DQI has a stronger impact on their decisions than for those who do not feel time pressure. This has implications for those who need to make decisions in crisis-type environments (Fisher and Kingma 2001).

Demographics

From the posttask questionnaire we collected information about gender, age, and education of the subjects in an attempt to find what effect, if any, these demographic characteristics might have on the subjects' reaction to DQI. Among the experts performing the complex task, we found that age was a factor. We found that both the older and younger groups

1-13

used DQI, but the degree of noncomplacency was much greater in the older group as compared to the younger group, revealing that older people paid more attention to DQI than younger people.

Among the IS professionals (experts) performing the complex task, education was a factor. Again, based on the questionnaire, we created three post hoc groups. All three education levels considered (high school, bachelor's degree, and master's degree) were not complacent. Interestingly, the level of noncomplacency was the lowest for the high-school-only graduates and the highest for the bachelor's degree subjects. The higher chi-square for those with bachelor's degrees as compared to the group with master's degrees could indicate that those with a more generalized background would use DQI more and use it more effectively. This is consistent with the greater use of DQI by managers as opposed to specialists.

Future Research and Concluding Remarks

At this point we are still investigating whether a decision maker will use DQI or not, and under what conditions. Further work needs to be done to determine if there is in fact a correct way to use DQI. How the DQI should be defined and weighted are key questions and may vary among organizations and problem types. Hence, actual case studies would contribute much to this area. This should be followed up with research on format of DQI for different types of problems. DQI may be formatted as numeric interval data between 0 and 1, as we have done. Alternatively it may be formatted as ordinal data with values such as good, poor, and so forth. It is well known that there are many decision processing strategies with two major categories, compensatory and cutoff (Payne et al. 1993). If cutoff techniques are used with DQI, that may significantly change the relative weight of the DQI; i.e., an alternative with the word "poor" listed as the DQI for one of its attributes may be rejected, whereas it would not have been rejected with interval data and compensatory techniques. This may advise database designers as to the desirability of using one or the other format.

Note that in our study we did not constrain the subjects to using a particular decision-making strategy. Through the posttask questionnaire we attempted to discern the strategy that was used, but found that most subjects used a combination of strategies. Although the focus of this study was *whether* DQI was used, future research might investigate the issue of *how* DQI is used when it is used.

Although the idea of incorporating data tags into databases is not new, it is not clear under what circumstances it would be most beneficial. Based on our findings, DQI should be incorporated into those data sets used by management. A direct comparison of those with management and nonmanagement backgrounds showed that the former were more influenced by DQI than were the latter. This is consistent with the finding that those without domain-specific experience used DQI more than those with domain-specific experience. This agrees with the earlier findings that too much domain-specific experience may prevent objective use of all available information (Gilliland et al. 1994, Yates 1991). One could hypothesize that those with the most experience regarding a situation, although they may indeed use DQI, are not as influenced by it on account of prior experience with the issue and similar data. The overall conclusion is that DQI should be made available to management not as familiar with the problem at hand.

Organizations wishing to begin a program of using DQI should be aware of the fact that there was a lack of consensus when experts were presented with DQI. The lack of complacency and consistency among the experts is deemed to be positive as it illustrates that the experts will use the DQI. However, the lower than expected consensus levels indicate that the experts used the DQI differently. We can predict that the addition of information about data quality to a database is likely to change the decision made, but we cannot predict what that new decision may be. Simply put, different experts, given a common task with the same data quality information, reached different decisions. It would be extremely beneficial for organizations to conduct seminars and DQI education prior to beginning a reliance on DQI if the organizations expect to have consensus in their decision making.

1-14

Appendix A. The Job Transfer Task Description[1]

J. Doe's job is being downsized and his company is allowing him to transfer to one of seven jobs. Unfortunately, J. Doe is sick on the day that he is supposed to submit his choices in order of preference. At a previous time J. Doe began the decision process of examining the jobs. First he identified nine characteristics and indicated which characteristics were most important to him. He reflected these in weights from 1 (most important) to 0.2 (least important). Next he rated each job as to the attractiveness of each individual characteristic on a 100-point scale, where the higher number is more desirable. For example, a rating of 90 for job content is more desirable than a rating of 25. Finally, he multiplied the weight times the ratings to obtain a weighted score for each job characteristic for each job. However, because he became ill, he was unable to finish ranking the jobs. He asked you to review his work and submit his choices ranked in order of preference from the most desirable job (Rank 1) to the least desirable job (Rank 7).

The job characteristics[2] and preferences are: J. Doe's number one priority is *job security* because, due to his health, he cannot risk losing his job and benefits. His second highest priority is to maintain his *current salary*. His third priority is *school quality*. J. Doe hopes to obtain a job that he likes, thus his fourth priority is *job content*. His

Job Alternative A

Criterion	Reliability	Rating	Weight	Weighted Scores
JOB CONTENT	0.8	84	0.7	58.8
CAREER GROWTH		24	0.3	7.2
CURRENT SALARY	0.8	80	0.9	72
FUTURE SALARY		16	0.5	8
LOCATION		56	0.4	22.4
CLIMATE		50	0.6	30
JOB SECURITY	0.5	54	1	54
SCHOOL QUALITY	1	42	0.8	33.6
COST OF LIVING		22	0.2	4.4

RANK = ____

Explanation:

Job Alternative B

Criterion	Reliability	Rating	Weight	Weighted Scores
JOB CONTENT	0.2	90	0.7	63
CAREER GROWTH		84	0.3	12.6
CURRENT SALARY	0.7	70	0.9	63
FUTURE SALARY		82	0.5	41
LOCATION		60	0.4	24
CLIMATE		68	0.6	40.8
JOB SECURITY	0.2	90	1	90
SCHOOL QUALITY	0.2	80	0.8	64
COST OF LIVING		50	0.2	10

RANK = ____

Explanation:

Job Alternative C

Criterion	Reliability	Rating	Weight	Weighted Scores
JOB CONTENT	0.7	50	0.7	35
CAREER GROWTH		20	0.3	6
CURRENT SALARY	0.8	16	0.9	14.4
FUTURE SALARY		48	0.5	24
LOCATION		30	0.4	12
CLIMATE		32	0.6	19.2
JOB SECURITY	0.8	24	1	24
SCHOOL QUALITY	1	20	0.8	16
COST OF LIVING		60	0.2	12

RANK = ____

Explanation:

[1]This task was developed by drawing on one of the author's experiences with job transfers and physical relocations as a middle-level manager at a major computer company. Additionally, four other professionals with varied experience were interviewed to determine the criteria for a relocation decision. They included a retired professional who is now an information systems consultant, a director of career services at a small college, a management consultant (team productivity), and a corporation quality manager. A Delphi process led to the final set of nine attributes for the "job transfer task."

[2]The job characteristics are indicated in italic letters.

1-15

Job Alternative D Criterion	Reliability	Rating	Weight	Weighted Scores
JOB CONTENT	0.8	30	0.7	21
CAREER GROWTH		52	0.3	15.6
CURRENT SALARY	0.6	48	0.9	43.2
FUTURE SALARY		54	0.5	27
LOCATION		26	0.4	10.4
CLIMATE		54	0.6	32.4
JOB SECURITY	0.8	80	1	80
SCHOOL QUALITY	0.8	30	0.8	24
COST OF LIVING		52	0.2	10.4

RANK = ____

Explanation:

Job Alternative E Criterion	Reliability	Rating	Weight	Weighted Scores
JOB CONTENT	0.8	50	0.7	35
CAREER GROWTH		76	0.3	22.8
CURRENT SALARY	0.7	24	0.9	21.6
FUTURE SALARY		30	0.5	15
LOCATION		56	0.4	22.4
CLIMATE		44	0.6	26.4
JOB SECURITY	0.8	18	1	18
SCHOOL QUALITY	1	48	0.8	38.4
COST OF LIVING		56	0.2	11.2

RANK = ____

Explanation:

Job Alternative F Criterion	Reliability	Rating	Weight	Weighted Scores
JOB CONTENT	0.2	30	0.7	21
CAREER GROWTH		24	0.3	7.2
CURRENT SALARY	0.8	18	0.9	32.4
FUTURE SALARY		20	0.5	10
LOCATION		56	0.4	22.4
CLIMATE		18	0.6	10.8
5 JOB SECURITY	0.5	82	1	82
SCHOOL QUALITY	0.2	72	0.8	57.6
COST OF LIVING		44	0.2	8.8

RANK = ____

Explanation:

Job Alternative G Criterion	Reliability	Rating	Weight	Weighted Scores
JOB CONTENT	0.8	50	0.7	35
CAREER GROWTH		90	0.3	27
CURRENT SALARY	1	82	0.9	73.8
FUTURE SALARY		60	0.5	30
LOCATION		24	0.4	9.6
CLIMATE		64	0.6	38.4
JOB SECURITY	0.8	52	1	52
SCHOOL QUALITY	0.8	48	0.8	38.4
COST OF LIVING		22	0.2	4.4

RANK = ____

Explanation:

fifth priority is to avoid a colder *climate.* He is moderately interested in salary increases or *future salary.* Whatever his new *location* is he would like to minimize a commute to work. He is not interested in *career growth* opportunities. His last priority is *cost of living* since he has acquired most things he needs and can avoid unnecessary expenses.

The relative weights for these characteristics (criterion weights) are: *Job content* = 0.7; *career growth* = 0.3; *current salary* = 0.9; *future salary* = 0.5; *location* = 0.4; *climate* = 0.6; *job security* = 1; *school quality* = 0.8; and *cost of living* = 0.2. The job alternatives from A to G with the criterion weights, ratings, and weighted scores (weight × rating) are shown on the next two pages.

1-16

Your task is to rank the jobs to meet J. Doe's needs given his weights and ratings. Rank the jobs from 1, being the best choice, to 7 being the last choice. However, you realize that the data he obtained may not be completely accurate. For instance, his information on *job content* came from someone who never worked at the new locations. Also, *job security current* and *future salary*, and *career growth*, are dependent on a volatile market. *School quality* information may be unreliable if presented by real estate people only interested in selling particular houses. *Location* commute time may be based on single trips at 2 pm at some locations but on many trips during rush hour at other locations. The reliability of the information about the job characteristics came from different sources and may vary from job to job.

You decide to incorporate this uncertainty into your decision-making process by using a 0-1 reliability measure where a score of 1 indicates perfectly reliable data and 0 scores imply completely unreliable data. You were only able to estimate reliability for four of the nine criteria.

The job alternatives from A to G with the criterion weights, ratings, and weighted scores (weight X rating) are shown on the next two pages. In addition, you have included a "reliability" column to indicate the 0-1 reliability measure for each criterion for each alternative. Remember that reliability refers to the data and not to the weights. Next to each job description write its rank, along with a brief explanation of exactly how you arrived at the rank.

Appendix B. Post Questionnaire[3]

Number: ________

The following information will not be used to identify individuals in any way. Information is recorded based on the random number assigned to your questionnaire. No attempt will be made to correlate these random numbers with any actual identity.

1. Female ________ Male ________
2. My Education is: (Please indicate the highest level that you have achieved)
   High School ________
   Bachelors Degree ________
   Masters Degree ________
   Post Masters Degree ________ (Specify: ____________)
3. My Occupation may be described as: Professor ________; Professional Educator ________; Full time graduate Student ________; Engineer ________; Programmer ________; Administrative ________; Accountant ________; Entrepreneur ________; Business ________; Other: ________.
4. My age is 17–20 ____; 21–30 ____; 31–40 ____; 41–50 ____; 51–60 ____; greater than 60 ____.

[3] (Fisher 1999, pp. 204–211)

5. The number of years that I have lived in my own apartments or homes (i.e., not my parents') is
   0 ____; 1–5 ____; 6–10 ____; 11–15 ____; 16–20 ____; 21–25 ____; 26–30 ____; greater than 30 ____.
6. I am currently a manager or supervisor: Yes ____; No ____.
7. I have selected and lived in ____ (how many) apartments or homes.
8. In the apartment selection task what data was most useful to you?

   ____

9. In the apartment selection task what data would you like to have had that you did not have?

   ____

10. I am confident that my apartment selection choices are correct:
    : ________ : ________ : ________ : ________ : ________ :
    Strongly Agree | Agree | Neither Agree/Disagree | Disagree | Strongly Disagree
11. The factors that contribute to my degree of confidence (or lack of) in the apartment selection task are:

    ____

    ____

12. For the apartment selection task: Did you compare alternatives two at a time and then pick the best one and then compare that one to the next one and so on until only one was left standing?
    Always ____ Sometimes ____ Seldom ____ Never ____
13. For the apartment selection task: Did you focus on single characteristic (attribute) and compare across all alternatives?
    Always ____ Sometimes ____ Seldom ____ Never ____
14. For the apartment selection task: Did you tend to compute a sum of all attribute values multiplied by their weights and derive a single score for each alternative?
    Always ____ Sometimes ____ Seldom ____ Never ____
15. For the apartment selection task: Did you establish minimal acceptable values for each attribute of each alternative and then see if each alternative, one by one, met that "cutoff?"
    Always ____ Sometimes ____ Seldom ____ Never ____
16. For the apartment selection task: Did you use a combination of the above techniques?
    Always ____ Sometimes ____ Seldom ____ Never ____
17. I experienced time pressure to complete the Apartment Selection Task.
    : ________ : ________ : ________ : ________ : ________ :
    Strongly Agree | Agree | Neither Agree/Disagree | Disagree | Strongly Disagree
18. The number of years that I have been a full-time employee is:
    0 ____; 1–10 ____; 11–20 ____; 21–30 ____; greater than 30 ____.
19. In the job relocation task what data was most useful to you?

    ____

1–17

20. In the job relocation task what data would you like to have had that you did not have?

____

21. I am confident that my job relocation choices are correct:

: ______ : ______ : ______ : ______ : ______ :

Strongly Agree	Agree	Neither Agree/ Disagree	Disagree	Strongly Disagree

22. The factors that contribute to my degree of confidence (or lack of) are:

____

____

23. How many times have you transferred jobs within a location?

0 ____ : 1–3 ____ : 4–6 ____ : 7–9 ____ : 10 or more ____

24. How many times have you transferred jobs to a new (e.g., change in commute) that did not require a household move?

0 ____ : 1–3 ____ : 4–6 ____ : 7–9 ____ : 10 or more ____

25. How many times have you transferred jobs that required a household/apt move?

0 ____ : 1–3 ____ : 4–6 ____ : 7–9 ____ : 10 or more ____

26. For the job relocation task: Explain the approach that you used in reaching a conclusion. How did you determine the rankings of the alternatives?

27. For the job relocation task: Did you compare alternatives two at a time and then pick the best one and then compare that one to the next one and so on until only one was left standing?

Always ____ Sometimes ____ Seldom ____ Never ____

28. For the job relocation task: Did you focus on single characteristic (attribute) and compare across all alternatives?

Always ____ Sometimes ____ Seldom ____ Never ____

29. For the job relocation task: Did you tend to compute a sum of all attribute values multiplied by their weights and derive a single score for each alternative?

Always ____ Sometimes ____ Seldom ____ Never ____

30. For the job relocation task: Did you establish minimal acceptable values for each attribute of each alternative and then see if each alternative, one by one, met that "cutoff?"

Always ____ Sometimes ____ Seldom ____ Never ____

31. For the job relocation task: Did you use a combination of the above techniques?

Always ____ Sometimes ____ Seldom ____ Never ____

32. I experienced time pressure to complete the Job Transfer Task.

: ______ : ______ : ______ : ______ : ______ :

Strongly Agree	Agree	Neither Agree/ Disagree	Disagree	Strongly Disagree

References

Agarwal, R., V. Sambamurthy, R. M. Stair. 2000. The evolving relationship between general and specific computer self-efficacy—An empirical assessment. *Inform. Systems Res.* 11(4) 418–430.

Ahituv, N., M. Igbaria, A. Stella. 1998. The effects of time pressure and completeness of information on decision making. *J. Management Inform. Systems* 15(2) 153–172.

Austin, R. D. 2001. The effects of time pressure on quality in software development: An agency model. *Inform. Systems Res.* 12(2) 195–207.

Ballou, D. P., H. L. Pazer. 1995. Designing information systems to optimize the accuracy-timeliness tradeoff. *Inform. Systems Res.* 6(1) 51–72.

——, G. Tayi. 1999. Enhancing data quality in data warehouse environments. *Comm. ACM* 42(1) 54–57.

Chengalur-Smith, I., Pazer, H. 1998. Decision complacency, consensus and consistency in the presence of data quality information. *Proc. Conf. Inform. Quality.* Cambridge, MA, 88–101.

——, D. P. Ballou, H. Pazer. 1997. The impact of data quality tagging on decision complacency. *Proc. Conf. Inform. Quality.* Cambridge, MA, 209–221.

——, ——, ——. 1999. The impact of data quality information on decision making: An exploratory analysis. *IEEE Trans. Knowledge and Data Engrg.* 11(6) 853–864.

Dukerich, J. M., Mary Lippitt Nichols. 1991. Causal information search in managerial decision making. *Organ. Behavior and Human Decision Processes* 50 106–122.

Fisher, C. W., 1999. An empirically based exploration of the interaction of time constraints and experience levels on the data quality information (DQI) factor in decision making. Doctoral dissertation, State University of New York at Albany, NY. *Dissertation Abstr. Internat.* (UMI No. 9954037).

——, B. R. Kingma. 2001. Criticality of data quality as exemplified in two disasters. *Inform. and Management* 39 109–116.

Gilliland, S. W., L. Wood, N. Schmitt. 1994. The effects of alternative labels on decision behavior: The case of corporate site selection decisions. *Organ. Behavior and Human Decision Processes* 58 406–427.

Grether, D. M., A. Schwartz, L. L. Wilde. 1986. The irrelevance of information overload: An analysis of search and disclosure. *Southern California Law Rev.* 59 277–303.

Haisten, M. 1995. Planning for a data warehouse. *InfoDB* (February) 12–21.

Hall, R. H. 1991. *Organizations: Structures, Processes and Outcomes.* Prentice Hall, Englewood Cliffs, NJ.

Kingma, B. R. 1996. *The Economics of Information: A Guide to Economic and Cost-Benefit Analysis for Information Professionals.* Libraries Unlimited, Englewood, CO.

Klein, B. D., D. L. Goodhue, G. B. Davis. 1997. Can humans detect errors in data? Impact of base rates, incentives and goals. *MIS Quart.* 21 169–194.

Mackay, J. M., J. J. Elam. 1992. A comparative study of how experts and novices use a decision aid to solve problems in complex knowledge domains. *Inform. Systems Res.* 3(2) 150–172.

Mao, J., I. Benbasat. 2000. The use of explanations in knowledge-based systems: Cognitive perspectives and a process-tracing analysis. *J. Management Inform. Systems* 17(2) 153–179.

March, J. G., H. A. Simon. 1958. *Organizations.* John Wiley & Sons. New York.

1–18

Morrow, D. G., V. O. Leirer, P. A. Altieri. 1992. Aging, expertise, and narrative processing. *Psych. Aging* 7 376–378.

Ordonez, L., L. Benson III. 1997. Decisions under time pressure: How time constraints affect risky decision making. *Organ. Behavior and Human Decision Process* 71 121–140.

Paese, P. W., Janet A. Sniezek. 1991. Influences on the appropriateness of confidence in judgment: Practice, effort, information and decision-making. *Organ. Behavior and Human Decision Processes* 48 100–130.

Payne, J. W., James R. Bettman, Eric J. Johnson. 1993. *The Adaptive Decision Maker*. Cambridge University Press, Cambridge, MA.

Sanbonmatsu, David M., Frank R. Kardes, Paul M. Herr. 1992. The role of prior knowledge and missing information in multiattribute evaluation. *Organ. Behavior and Human Decision Processes* 51(1) 76–91.

Shaft, T. M., I. Vessey. 1995. The relevance of application domain knowledge: The case of computer program comprehension. *Inform. Systems Res.* 6(3) 286–299.

Sheskin, David J. 2000. *Handbook of Parametric and Non-parametric Statistical Procedures*, 2nd ed. Chapman & Hall, Washington, D.C.

Simon, H. A. 1957. *Administrative Behavior: A Study of Decision Making Processes in Administrative Organizations*. Collier/MacMillan, New York.

Svenson, O., A. Edland. 1987. Changes of pressure under time pressure: Choices and judgments. *Scandinavian J. Psych.* 28 322–330.

Van Bruggan, G., A. Smidts, B. Wierenga. 1998. Improving decision making by means of a marketing decision support system. *Management Sci.* 44(5) 645–658.

Wand, Y., Richard Y. Wang. 1996. Anchoring data quality dimensions in ontological foundations. *Comm. ACM* 39(11) 86–95.

Wang, R. Y., S. E. Madnick. 1990. A Polygon model for hetergeneous database systems: The source tagging perspective. *Proc. 16th Internat. Conf. Very Large Databases*. Brisbane, Australia, 519–538.

——, D. Strong. 1996. Beyond accuracy: What data quality means to data consumers. *J. Management Inform. Systems* 12 5–34.

——, Veda C. Storey, Christopher P. Firth. 1995. A framework for analysis of data quality research. *IEEE Trans. Knowledge and Data Engrg.* 7(4) 623–639.

Yates, J. F., Linda S. McDaniel, Eric S. Brown. 1991. Probabilistic forecasts of stock prices and earnings: The hazards of nascent expertise. *Organ. Behavior and Human Decision Processes* 49 60–79.

Zakay, D., S. Wooler. 1984. Time pressure, training, and decision effectiveness. *Ergonomics* 27 273–284.

1–19

編號: 2

1 INTRODUCTION

THE traditional process applied to Web searching is to consider that the semantics of a Web page can be computed solely from the contents of the page. Typical queries submitted to Web search engines are short lists of keywords given by the user. There has been a significant research effort, mainly by Google [5], to find a way of ordering the results of such queries, in order to significantly decrease the time spent browsing irrelevant URLs. The *PageRank* algorithm [5] takes into account various other parameters, such as the link structure of the World Wide Web, in order to give an "importance" ranking to pages retrieved by a query. We believe, and we show in this paper, that we can derive information about the semantics of a page by analyzing the links that point to a given page. It is in this sense that World Wide Web documents are different from a simple collection of documents. We insist on the fact that links convey simple yet robust information. Such information deserves to be taken into account when answering queries. Our objective is to construct thematic subsets of World Wide Web documents called THESUs (Thematic Subsets of the World Wide Web). The semantic proximity of the pages in a THESU is not only derived by the pages' content, but by the semantics of the links pointing to this page; this is what we call **link semantics**.

- *I. Varlamis and M. Halkidi are with the Department of Informatics, Athens University of Economics & Business, Patision Street 76, Athens, 10434, Greece. E-mail: {varlamis, mhalk}@aueb.gr.*
- *M. Vazirgiannis is with the Department of Informatics, Athens University of Economics & Business, Patision Street 76, Athens 10434, Greece, and INRIA Futura, 4 rue Jacques Monod, Orsay Parc Club, 91893 Orsay, France. E-Mail: mvazirg@aueb.gr and michalis.vazirgiannis@inria.fr.*
- *B. Nguyen is with INRIA Futurs, 4 rue Jacques Monod, Orsay Parc Club, 91893 Orsay, France. E-mail: benjamin.nguyen@inria.fr.*

Manuscript received 20 Mar. 2003; accepted 16 July 2003.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-0013-0303.

Thus, the important feature of the World Wide Web on which we capitalize is the following:

Link semantics enrich Web pages' semantics and can be valuable when searching for similarities in the Web.

The contributions of this paper are summarized in the following:

- The main contribution is a *model* and a *language* to manage the extraction of pages inlinks and construct a semantic characterization of each page by enriching these keywords using an ontology.
- A clustering mechanism that exploits a similarity measure to organize a set of Web documents into groups of similar pages. The similarity measure deals with weighted sets of terms in a hierarchy. The similarity between sets is not based on exact matching of terms in the two sets, but on the combined similarity between all terms of the first set and all terms of the second set. Two distance (similarity) based clustering algorithms have been adapted to the system requirements, to perform Web document clustering.
- A fully implemented client/server system, called THESUS. Using the THESUS link management language, World Wide Web pages are collected, based on an initial list of keywords. These pages are processed and a short list of more precise descriptive keywords are extracted (mainly from the incoming and outgoing links) and mapped to an ontology provided by the user. A clustering scheme is then applied to this enriched data and the results are stored in a relational database. The THESUS client allows users to perform SQL queries over the stored data and enables searching for interesting documents or document sets based on connectivity and semantics.

2-1

We also demonstrate a set of services, which make use of the incoming or outgoing link semantics. They allow users:

- to provide a set of URLs and get the terms that frequently appear in their incoming or outgoing hyperlinks,
- to provide two sets of URLs (A and B) and get the terms that frequently appear in the hyperlinks from URLs in A to URLs in B, and
- to provide a set of URLs and an ontology and have the URLs characterized and clustered.

All these services are available at our Web site http://www.db-net.aueb.gr/thesus/.

1.1 Motivation for Collective Link Use

Web content is not always self-descriptive (i.e., multimedia contents). Generally, when answering queries such as looking for a car advertisement, looking for a picture, etc., there will be some added value if we also take into account what *others* think about the page. This extra information concerns the type or quality of the targeted page and is not always limited to the contents inside the <A HREF>...</A> tag. We claim that the semantics of a document U can be derived from the semantics of the incoming links that point to it. Assume a document V points to U, using the term "computer science" in the neighborhood of the hyperlink. This is a strong indication that U is characterized by the semantics the term "computer science" conveys. This indication becomes stronger when links from many different documents pointing to U bear the same semantics. This is, in a way, a collective semantics.

1.2 Motivation for Ontology Use

Web documents are usually characterized by extracted keywords and are assigned a grade of importance that takes into account link structures [5]. Finding the similarity between documents is based on exact matching between the keywords and the terms of the users' query. For instance, a document d_1 characterized by the keyword list $d_1 = \{\text{snake}, \text{desert}\}$ would be judged irrelevant to a document d_2 characterized by the list $d_2 = \{\text{adder}, \text{Sahara}\}$, yet it is arguable that the two lists of keywords (and, thus, the documents) are in fact related since an "adder" is a "snake" and the "Sahara" is a "desert," therefore, d_2 deals with the same concepts as d_1 because they are just more specialized. By replacing keywords with concepts in a hierarchy, a more flexible document matching process than binary matching can be achieved, handling both specializations and generalizations of senses.

In the current implementation, the ontology terms are connected using only one type of relation (a term X "is a subclass of" term Y, which denotes that X is less general than Y). This hierarchy of concepts forms a taxonomy, a simplified form of an ontology. Replacing extracted keywords with semantics from an ontology does not aim to identify the objective semantics of hyperlinks, but the subjective semantics in a certain domain of interest as those are perceived by the creators of the domain ontology. The main advantage of mapping extracted keywords to an ontology is that a large set of extracted keywords is mapped to a smaller set of concepts, which are organized in a hierarchy and represent the domain of interest. Users' queries are keyword sets that can also be mapped to concepts. This mapping reduces the dimensionality of the problem of matching documents to queries and, moreover, allows approximate matching based on concepts instead of "exact keyword matching."

THESUS[1] is a system that enhances the semantic organization of a set of pages and guides the user within it. It is envisioned as a personal service that will assist the user in the creation and querying of a rather compact high quality thematic collections of pages. The concepts developed here are expandable to the full World Wide Web; we bore scalability in mind during the creation and the implementation of THESUS. In order to give a real feeling of what sort of results can be achieved by the system, we also compare it to some popular search engines, even though it is NOT a search engine.

1.3 Paper Organization

The organization of the paper is the following: In Section 2, we refer to related work. Section 3 provides the THESUS information model and introduces the notion of "link semantics." Section 4 contains the definitions of the THESUS language operators divided into three main categories: crawling, semantics extraction, and link analysis operators. The architecture of the system follows in Section 5. In Section 6, we give some examples and results of our system on real data, proving it returns good results. The final section contains conclusions and discusses further work.

The appendices (which can be found on the Computer Society Digital Library at http://computer.org/tkde/archives/htm) contain the following information: Appendix A summarizes the definitions of data types and operators of the THESUS Model, while Appendix B provides the algorithms of the main operators in pseudocode.

2 Related Work

This section refers to related work mainly from the areas of link information analysis and querying of the World Wide Web. Additionally, since we capitalize on Web documents semantics and clustering based on semantics, we will present similar efforts in these areas. On the general THESUS topic, we refer to [17], [26], which focus more precisely on the impact of the similarity measure on the clustering schemes of THESUS and the subsequent experimentation. Readers familiar with these articles may want to skip Section 5.4.

2.1 Hyperlink Information and Semantics

The main issue addressed in this paper is that of managing keywords that are extracted from the links of Web pages. This is not a novel idea, since Phelps and Wilensky [28] introduce the idea of "robust" hyperlinks that contain "descriptive" information on the target document which

1. THESUS is the name of the system and the language. A THESUS is a thematic subset, and the plural is THESUs. We refer indifferently to a document, a page, or as we see later, a node, in order to define the basic entities stored in our system.

2-2

can be limited to five words, and some experiments have already been run, such as [7] where an "anchor window," the text neighboring the hyperlink, is defined to be 50 characters. In [32], a structure for hyperlink information with rich semantics emanating from a domain specific conceptual hierarchy is proposed. They also introduce a system to increase the Web searching capabilities by attaching information to a document, concerning its concept and its hyperlinks semantics.

2.2 Web Querying

WebSQL [3] is a language for extracting information from the Web. It models the Web as a relational database composed of two (virtual) relations, Document and Anchor, and provides an expandable set of Java classes, for querying the two virtual relations. WebSQL queries serve either information retrieval or Web maintenance tasks. Google's [14] advanced search allows inclusion and exclusion of terms that appear in the title or URL of a page. The ranking algorithm, PageRank [5], gives higher rank to most important pages, but does not group pages of similar topic. However, the algorithm exploits information of incoming links to enhance content information of pages. WebWatcher [2] detects the presence of certain keywords inside the hyperlink and neighboring text and ranks the links according to the user's interests. The set of keywords is produced using a training set but is limited to a few hundred terms. The ARC system [7] uses the hyperlink neighboring text to describe the contents of the pointed page and to enhance the hub and authority weight on a certain topic. This kind of ranking is called connectivity-based ranking [19].

2.3 Link Analysis and Web Document Clustering

Web document clustering is usually based on connectivity between documents [8] (Web structure) and not on conveyed semantics. Web content mining techniques mainly perform text mining on the whole document while ignoring Web structure. It would be very useful to consider both hyperlinks of pages and their contents when clustering large collections of Web documents. Zamir and Etzioni in [35] propose a suffix tree-clustering (STC) algorithm that uses phrases shared between documents to create the clusters. Haveliwala et al. [20] propose a clustering approach in which the similarity is based on words found in the documents along with their occurrence frequencies and the term matching is exact. Some works relate to the importance of links in promoting semantics in a hypermedia network. Kleinberg [22] states: "The link structure of a hypermedia environment can be a rich source of information about the content of the environment (...) But, for the problem of searching in hyperlinked environments such as the World Wide Web, it is clear from the prevalent techniques that the information inherent in the links has yet to be fully exploited."

There is a consensus that clustering techniques should be applied to the results of a query rather than on the whole Web space, in order to discover groups of relevant documents. An interesting server that uses hyperlink's structure to group interconnected results is Kartoo [23]. Vivísimo [32] proposes a clustering approach for Web document organization, making use of the short descriptions returned by other search engines. Northern Light [25] classifies all the documents of the entire collection into predefined subjects and, at query time, selects the subjects that best match the search results.

Our working hypothesis in THESUS is that Web page characterization is more valid when external information is used instead of the page's contents. As a result, hyperlink information can refer to the semantics of the target as they are provided by the source of the hyperlink. Hopefully, external authors act independently of the page's author and their opinions can be added to the ranking of the page; yet, our system could also be able to detect "malicious" links by finding that they have no term in the ontology that sets them close to the ones that appear in other links, for instance, and a scheme could be devised to filter them out. In our system, we rely on the use of an ontology and WordNet, in order to compute similarities between words. For more details on ontologies, you can refer to [15] and for information on WordNet [33], [1].

多媒体
超文字
之網頁

3 THESUS' Model

In this section, we provide an informal introduction of the THESUS model that motivates the formal specifications of the basic data types that will be detailed in Section 4.

We model the World Wide Web using the following concepts:

- *Pages* or, more precisely, URIs that provide a unique way of accessing some information on the Web. A page can be empty of textual content (for instance, a picture) yet have other pages pointing to it, with meaningful semantics. Henceforth, we will interchangeably use the terms "page," "node," and "document" for representing World Wide Web pages.
- *Links*, uniquely defined by source and target nodes. We believe that it is the links that carry the semantics of the Web page they point to. Therefore, we are interested in the union of the semantics of all the links pointing from a given page to another one. The exact location of link anchors in the two pages is also interesting to consider.

3.1 Link Semantics

Assume two pages S (source) and T (target) and the set of links $\{l_i\}$ that point from S to T. It is quite common for authors of a page to link it to another page by using a small set of keywords to describe the target page. These keywords usually appear in the hyperlink source tag, i.e., *<A HREF=...>a page on computer science</A>*, or in a short area around the hyperlink, i.e., *you will find <A HREF=...> here</A> information on computer science*. We call this set of keywords $\{k_j\}$ **link keywords**. Semantics is the study of meaning in language, and refers to the concepts to which extracted keywords map to a given ontology and a mapping mechanism from keywords to concepts. Given an ontology, which is the representation of the domain of interest and is the bearer of semantics in our system and by using the WordNet thesaurus, we are able to map extracted

2-3

keywords to ontology concepts, thus converting link keywords to what we call **link semantics**.

3.2 Page Characterization

As mentioned before, page T is effectively characterized by the semantics of its incoming links. Therefore, the set of keywords (and the corresponding set of concepts) that is derived from $\{l_l, \{k_j\}\}$ semantically defines the classification of page T seen from page S. In other words, the union of the sets of keywords $\{k_j\}$ is the keyword characterization that T assigns to S. This keyword characterization is the basis for producing the semantic properties that T assigns to S. Taking into account all the links to page T from different source pages $\{S_i\}$, we get an aggregate characterization, a collective view of how the set of pages $\{S_i\}$ characterizes T. In the case of a set of target pages $\{T_l\}$, the result of processing the links from $\{S_i\}$ to $\{T_l\}$ will be the collective characterization that $\{S_i\}$ assigns to $\{T_l\}$.

3.3 THESUS Model Datatypes

In this section, we define the entities that constitute the nucleus of the THESUS language design. Our reference space is the World Wide Web, perceived as a collection of documents (`w_docs`) connected with links (`w_links`). For a thematic area, we assume a subset of the documents and links of the World Wide Web (`docs` and `links`, respectively) form a thematic subset (THESU). For completeness, we assume the data types `URL` (the unique identifier of a page in the World Wide Web context), `keyword` (a string of characters characterizing a page), and `ontology_term` (a concept from the domain ontology). We also consider sets of `URL`, `keyword`, and `ontology_term` as types in our system. We will now define the fundamental entities necessary for THESUs' creation and manipulation.

3.3.1 THESUS Model Entity Definitions

Definition 1. *A* collection of documents, `docs` *is a set of vectors of the form* (`URL`, `{keyword}`, `{ontology_term}`, `other info`)*, where*

1. `URL` *is the identifier of the document as it appears on the World Wide Web,*
2. `{keyword}` *is a set of keywords characterizing the document,*
3. `{ontology_term}` *is the set of corresponding terms of the ontology that semantically describe the document, and have been constructed from the set of keywords, and*
4. `other info` *are additional informative facts that can be assigned to the document, such as the content of the page perceived as a string of characters (text) or the last date of change of the document (modification date).*

The documents in such a collection constitute a THESU.

Definition 2. *A* collection of links, `links`, *is a set of vectors:* (`URLS`, `URLT`, `{keyword}`) *where:* `URLS` *is the URL of the document from which the link emanates,* `URLT` *is the URL of the document to which the link points,* `{keyword}` *is a set of keywords characterizing the document, and* `{ontology_term}` *is a set of terms from the ontology that semantically define the document.*

3.3.2 World Wide Web Entities

In order to construct `docs`, we need to extract information from pages of the Web. The following collections, `w_docs` and `w_links`, are used during construction. In a nutshell, these entities are our vision of the World Wide Web as a virtual entity. We are able to materialize portions of it by combining Web services, such as search engines and Web crawling techniques.

Definition 3. *The* World Wide Web collection of documents, `w_docs`, *is a set of vectors of the form* (`URL`, `text`)*, where 1)* `URL` *is the identifier of the document as it appears on the World Wide Web and 2) text is content perceived as a string of characters.*

Definition 4. *The* links in the World Wide Web, `w_links`, *is a set of vectors of the form* (`URLS`, `URLT`)*, where 1)* `URLS` *is the URL of the document from which the link emanates and 2)* `URLT` *is the URL of the document to which link points. Here, we assume that all the links between two pages are aggregated in one vector in the* `w_links` *entity.*

4 THESUS Language

In this section, we introduce THESUS language that enables thematic selection of World Wide Web subsets and subsequent enrichment by extracting semantics from the links pointing to the pages of the subset. The THESUS language defines operators that combine connectivity features and related semantics, yielding meaningful results that are otherwise not obtainable from existing search mechanisms. The model's datatypes, the language's operators along with necessary extensions, and auxiliary functions are summarized in Appendix A (which can be found on the Computer Society Digital Library at http://computer.org/tkde/archives.htm).

When designing a language or a set of predicates, we bear in mind the following issues:

- Minimality: We are searching for a minimal set of operators, which have the least possible overlap of semantics, are simple in their design, and can be combined in constructing richer operations. The operators are grouped in distinct categories (crawling, link semantics extraction, and link analysis) aimed at different tasks. Nonetheless, there is some overlap between some of them since certain operators could be expressed in terms of the others.
- Expressiveness: The set of operators should be as high level as possible, easily understandable by the potential users, and with clear semantics. The operators are semantically rich and are directly applicable to the domain considered (i.e., traversal of the World Wide Web graph and semantics extraction).
- Closure: The sets of types and operators on them are closed. This implies that any simple or complex operator returns a type that is already defined in our system.

Let us note that the elements of type `ontology_term` are used as a specialization of the `keyword` type in some of the operators defined later on. In order to avoid repetition,

2-4

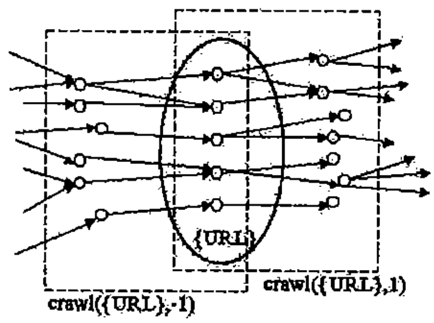


Fig. 1. Crawl operator.

in the definition of the language operators, we use only the keyword type. However, operators based on the semantics of the extracted keywords can be defined in a similar way. The transformation from one to the other is done by calling the function getSemantics({keyword}) (see Section 5.3).

In this section, we define the basic operators as introduced above and define additional operators on top of them, which convey richer knowledge about a set of World Wide Web pages. The following list of operators is neither complete nor exhaustive, but it gives an indication of the potential of the fundamental set of THESUS operators in terms of finding rich semantics in the World Wide Web. These higher level operators can all be expressed based on the previously introduced ones.

4.1 Crawling Operators

In order to create and characterize THESUs, it is necessary to be able to traverse the World Wide Web graph in order to 1) find the target pages of a page's outgoing links and 2) to find the source pages of a page's incoming links. This crawling function can be carried out in several levels in the World Wide Web graph starting from some node under consideration. We introduce the following "crawling" operators.

Definition 5: fetch(URL). *Given a URL, it returns either its textual content or null in case the link is broken.*

Definition 6: groupMatch({URLexpr}). *Given a set of URL expressions,* `{URLexpr}`*, the operator returns a set of URLs that match any of the expressions. The expression is a partial URL string and all returned URLs should begin with the URL expression.*

Example. For instance, the operation

`groupMatch({http : //www.nasa.gov/)`

returns all the pages under http://www.nasa.gov/.

Definition 7: crawl({URL}, N). *Given a set of URLs (*`{URL}`*) and an integer N, the operator returns a set of URLs, consisting of the pages that are being pointed to/from each page in* `{URL}` *in each level of the World Wide Web graph until level N. Positive values of N imply that we crawl the World Wide Web in a "forward" manner by following the outgoing links of pages in {URL}, whereas negative values of N denote "backward" crawling through the incoming links of pages in {URL}. The example in Fig. 1 shows that for N=1, the crawl operator returns the documents in {URL} and all the documents pointed by them, whereas for N=-1, it returns the documents in {URL} and all the documents that point to them.*

4.2 Links' Semantics Operators

Here, we propose a set of auxiliary functions to be used in the definition of this category of operators. In order to extract the semantics of links, we need to be able to locate the position of the source of the link (i.e., <A HREF='...'>) in a page and, subsequently, process its neighborhood in order to extract keywords. This is performed using a set of auxiliary functions.

Function 1: getpos (text, URLT). Given the text of a page, text, and the string corresponding to the target URL of the link, `URLT`, the function returns the set of positions—{pos}—in the text, where the string of URLT occurs.

Function 2: process (w_docs.TEXT, pos, 100). It is called for each occurrence of URLT, processes the hyperlink's neighboring area (i.e., 100 characters) and returns a set of keywords—{keyword}—extracted from the area that represents the keyword characterization of the specific link.

Function 3: getSemantics ({keyword}). We use this function that returns a set of terms from the ontology—`{ontology_terms}`. The function maps the extracted set of keywords to the closest concepts and is detailed in Section 5.3.

THESUS' language works not only with specific keywords contained in the pages, but also with ontology terms and, thus, truly represent "semantics" as an understanding of what a page deals with.

Definition 8: linkKeywords (URLS, URLT). *Given two pages identified by URLs,* `URLS` *and* `URLT`*, the operator returns the union of keywords that are extracted from all the links emanating from* `URLS` *targeting to* `URLT`*.*

Definition 9: groupKeywords ({URLS}, {URLT}). *Given two sets of URLs,* `{URLS}` *and* `{URLT}`*, the operator returns the keywords extracted from all the links that point from any page in* `{URLS}` *to any page in* `{URLT}`*. The sets of URLs {URLS} or {URLT} in groupKeywords may contain a single URL, or can be empty. When {URLS} is empty, operator* `groupKeywords` *examines every incoming link to pages in {URLT}. This information indicates how pages in {URLT} are described; in other terms, "what the world thinks" for pages in {URLT}. When {URLT} is empty, the operator examines every outgoing link of pages in {URLS}. The extracted information indicates how pages in {URLS} describe the pages they point to, in other words, how pages in {URLS} "perceive the world."*

Definition 10: thematicCrawl ({URL},{keyword},N). *The operator combines Definitions 7 and 8. Given a set of URLs (*`{URL}`*), a set of keywords (*`{keyword}`*), and an integer N, the operator returns a set of URLs. This set of URLs consists of the pages that are being pointed to/from each page in* `{URL}` *in each level of the World Wide Web graph until level N, having at least one keyword from the* `{keyword}` *set in the link keywords.*

Example. In the case of thematicCrawl (`{U1}`, `{keyword1,keyword2}`, `2`), starting from page `U1` the operator returns all the URLs that are targeted by links starting from `U1`, containing either keyword1 or

2-5

keyword2 and that are reachable after at most 2 hops.

The operator is called thematicCrawl since, when it is used with keywords that fall in the same thematic area, it collects URLs that have a high probability to be on similar subjects. The operator can also defined as thematicCrawl ({URL}, {ontology_term}, N), where only the links with specific semantics (i.e., terms of the ontology or similar words) are followed.

4.2.1 Advanced Link Semantics Operators

The definition of the groupKeywords operator assumes that the set of keywords returned, is the union of the keywords that appear in each link without counting keyword occurrences. However, if we take into account keyword occurrences, the modified groupKeywords definition conveys three different meanings based on the way keyword occurrences are aggregated.

In the definitions that follow, we use three auxiliary functions:

Function 4: getkeys(WKEYS). It takes a set (WKEYS) of keywords and number of occurrences pairs (KEY, TIMES)—as an input and returns the set of keywords {KEY}.

Function 5: getTimes(WKEYS,K). It takes a set of (KEY,TIMES) pairs and a keyword K as an input and returns the number of occurrences (TIMES) of keyword K.

Function 6: update(WKEYS,K,N). It takes a set of (KEY,TIMES) pairs, a keyword K, and an integer N as an input and increases the TIMES value of K by N (updates pair (K,T) of WKEYS to (K,T+N)).

Definition 11: weightedGroupKeywords ({URLS},{URLT}). *Given two sets of URLs,* {URLS} *and* {URLT}*, the operator returns a set of pairs {(KEY, TIMES)}, where KEY represents a keyword that appears in the links that point from any page in* {URLS} *to any page in* {URLT} *and TIMES the number of occurrences of keyword KEY.*

Definition 12: weightedTargetKeywords ({URLS},{URLT}). *Given two sets of URLs,* {URLS} *and* {URLT}*, the operator returns a set of pairs {(KEY,TIMES)}, where KEY represents a keyword that appears in the links that point from any page in* {URLS} *to any page in* {URLT} *and TIMES the number of target pages characterized by keyword KEY.*

Definition 13: weightedSourceKeywords ({URLS}, {URLT}). *Given two sets of URLs,* {URLS} *and* {URLT}*, the operator returns a set of pairs* {(KEY,TIMES)}*, where* KEY *represents a keyword that appears in the links that point from any page in* {URLS} *to any page in* {URLT} *and* TIMES *the number of source pages that use keyword* KEY *in their links to* {URLT}.

The examples in Section 6.2.2 illustrate the different semantics of the various implementations of groupKeywords.

4.3 Link Analysis Operators

Apart from the core set of fundamental operators, we considered defining specific operators that are of high added value. Moreover, potential users of the system can easily define other operators in terms of the fundamental ones. We extend the definitions of hub and authorities [22] and those of cocitations [29] and couplings [21]. We exploit and enrich the above concepts with semantics ε from the links. We introduce the following concept:

- *Thematic Hubs.* We enhance the semantics of a by considering the semantics of the outgoing link is obvious that a hub with many outgoing links similar semantics is thematically focused and potentially more interesting. Kleinberg in [22] provides a recursive definition of hubs and authorities; however, for simplicity, we use the one step definition.
- *Thematic Authorities.* We enhance the semantics of an authority by considering the semantics of the incoming links. For instance, if an authority node is pointed to by links of type "databases," it is very likely that this is a page that contains reference material on databases.
- *Thematic Cocitations and Couplings.* If nodes B and C contain links to node A, then A is "cocited" by B and C. Moreover, if the links have similar semantics, then the semantics of A are strengthened and we have an indication that B and C have some similarity (thematic cocitation). On the other hand, if document C points both to documents A and B and the two links bear similar semantics, we have an indication that A and B are similar (thematic coupling).

5 THESUS System Architecture

This section presents the architecture of the THESUS system. The system components, their input and output, their operation, and the innovative aspects they introduce follow.

The THESUS system is fully implemented. The system's components include (see Fig. 2):

- The "information acquisition" module gathers a set of URLs that appear relevant to the topic under consideration.
- The "information extraction" module extracts keywords within an expanded hypertext area around a hyperlink.
- The "information enhancement" module enhances extracted hyperlink information with semantic information by mapping extracted keyword sets to sets of concepts in an ontology.
- The "clustering module" that partitions the set of URLs into semantically coherent subsets based either on the extracted keywords or on the respective ontology concepts. Similarity between pages is computed using a novel similarity measure for sets of ontology terms. The measure is based on the distance of terms in the ontology hierarchy, instead of exact term matching.
- The "query engine" enables searching in the collection. It takes advantage of the clusters that are closer to the user's query and ranks the results accordingly.

Information concerning the documents is stored in a relational database. Knowledge from the WordNet 1.6 database [33] and the ontology is also exploited. The system has been evaluated with experimental sets of the order of 10^4 Web documents, a reasonable size for a thematic Web repository.

2-6

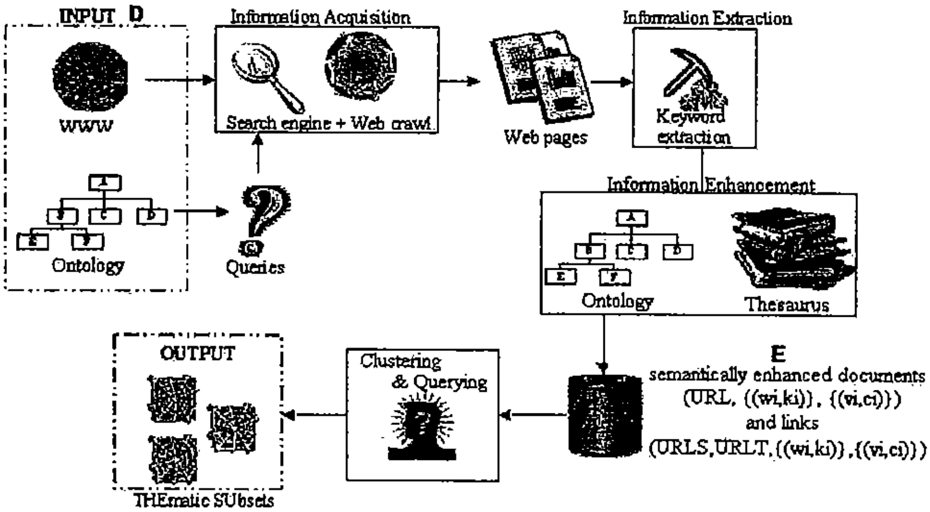


Fig. 2. THESUS functional architecture.

It is reasonable that the acquisition and processing of a large corpus of Web documents is a resource demanding task. As a consequence, the first three modules of the system are mainly used as offline tools for building a semantically enriched repository of Web documents and the clustering module assists in organizing this repository into meaningful subsets. The query module operates as an end-user tool that allows for browsing and searching the enriched information repository. See Section 6 for details on running times for the various tasks.

The crawling system was developed as a multithreaded Java application. Information extracted from Web pages and related semantics are stored in a relational DBMS (Microsoft-SQL Server®). The access to the database is performed using a JDBC native driver. The client application remotely connects to the database server using JDBC and allows users to perform the THESUS language operations on the collected data. The Web services are implemented as Java classes.

5.1 Information Acquisition

We assume an initial set of documents $D = \{d_i\}$, that are related to a certain thematic domain of interest, and an ontology O on the same domain (i.e., Arts, Music, Technology, etc.). The crawler generates the core set of URLs depending on the existence of an initial set of documents D:

1. If D is given, the crawling starts from this set without using any Web search engine services. Expansion is performed as multiple iterations of the "root-set generation" algorithm of HITS [22], limited to those links that display on-topic keywords in an expanded window of hypertext. In opposition to the full focused crawling [6], this crawling is totally unsupervised and is not based on exact matching between found keywords and ontology terms. Expansion follows hyperlinks that are *"characterized"* by at least one concept in the ontology (the anchor-text area around and inside the hyperlink contains a keyword that is mapped to the specific concept in the ontology). This process stops after a certain number of crawling levels have been completed or after all documents suggested for crawling in a step have already been collected. Due to the weak connectivity of the World Wide Web documents (only 28 percent of the Web is strongly connected according to [4]), the crawling process for a small ontology (i.e., less than 20 concepts) and a small initial set stops in a few steps, whereas, for an ontology with a few hundred concepts and a larger initial set stops after many crawling steps.
2. If D is not given, we consider no prior knowledge on the specific domain, so we use a Web search engine to locate documents that possibly relate to the domain. The only input is the ontology O describing the user's interests formed as a hierarchy of concepts. The system generates the paths from the root to all the nodes in the ontology, queries a Web search engine for all the generated paths ($node_1$ AND $node_2$ AND... AND $node_K$) and collects the top answers returned for each query. If the ontology contains N concepts, then N paths and, respectively, N queries are generated. If the search engine returns a maximum of M answers, the total number of URLs in D will be no more than $M \times N$. Experimentation shows that the results are usually highly overlapping (in terms of URLs) so the final number of documents in D after removing duplicates is significantly less than $M \times N$.

5.2 Information Extraction

In order to enrich the information on the core set of URLs, we crawl one level backwards and collect the incoming links information. The crawl({URL},-1) operator requires the use of a Web service that gives the top N incoming links of a page. The documents are processed as follows to enhance their classification quality. The first step is to process the

2-7

```
<p><a href="http://www.lib.unc.edu/music/">University of North Carolina, Chapel Hill
- Music Library</a></p><p>[illegible]
[illegible] <br>&quot; one of the largest
university music collections in the United States has over 250,000 volumes of
books, periodicals,
```

Fig. 3. Hyperlink to the Music Library of the University of North Texas homepage.

incoming links to D and extract keywords from the respective source pages. Calling the weightedGroupKeywords operator for the documents {b} that point to each document d_i in the set we get a set of pairs of keyword/ times of occurrences $\{(k_{di}, t_{kdi})\}$ that characterize d_i.

In Section 6, we experimentally verify the potential of the approach.

The issue of extracting keywords from the links is important in the context of this paper. Taking into account the works of [28], [7], in THESUS, keywords are extracted from 1) the link anchor (i.e., the string between the <a> and tags) and 2) from two text strings (each of 100 characters long), one preceding the starting link tag and another following the ending link tag. The "window" is trimmed whenever certain html tags appear, such as <a>, <li>, <tr>, <td>, etc., because such tags usually signify the logical end of the hyperlink neighboring area. The window size has been chosen *after* experimentation so that the resulting mean number of keywords is approximately five [28]. The procedure is summarized in Fig. 4.

For example, in the HTML fragment shown in Fig. 3, keywords are extracted from a hyperlink to the Music Library of the University of North Texas homepage.

The phrases extracted from the highlighted URL are:

- From the hyperlink area (dark color, white text), the phrase:

  "University North Texas Music Library"

  after removing frequent keyword "of" and character "-".
- From the 100 characters preceding (gray area) the hyperlink: The area is trimmed where appears, tags </p> and <p> are then removed, so no characters remain.
- From the 100 characters following (gray area) the hyperlink: The last partially selected number is ignored,
, " and punctuation marks are removed, stop-words are removed, too. The remaining phrase is:

  "largest university music collection United States".

The set of (keyword, occurrences) pairs that are extracted for this target URL is {(university, 2), (north, 1), (texas, 1), (music, 2), (library, 1), (largest, 1), (collection, 1), (united, 1), (states, 1)}. The system will then try to map this set of keywords to relevant terms in the ontology. "music" and "library," for instance, will probably be mapped as such, whereas "north" might not be mapped at all.

5.2.1 Amendments on Keyword Extraction Process

Some of the incoming links of a page are the navigation links that usually provide abundant and possibly irrelevant information for the targeted page (i.e. "back," "next," "home," etc.). Additionally, intrasite hyperlinks (links that point to pages in the same host) may carry misleading information on purpose, and attempt to affect Web agents that extract information from hyperlinks. A common solution to the first problem is to ignore all the navigation links when performing link analysis. One could also argue that it is better to disregard intrasite links since they also could provide biased information about pages. However, in our system, keywords extracted from such links do not affect the characterization of a Web page. Keywords that are commonly used for navigation purposes (i.e., back, next, etc.) are included in the stop-words list and, consequently, ignored by the keyword extraction module. Nevertheless, keywords that are not included in the stop-words list, but are irrelevant to the domain of interest (as it is expressed by the ontology), will not be mapped to a node in the ontology, while keeping links that are relevant.

The system also provides a solution for malicious hyperlinks, which are deliberately added in Web pages to

```
1)  Find the limits of the hyperlink area (start, end position)
2)  Find the limits of the pre-hyperlink area (start-100, start)
3)  Find the limits of the post-hyperlink area (end, end+100)
4)  Search for image tags in the inside hyperlink area.
   -Extract keywords from the alt attribute of image tags
   -Remove image tags
   -Remove punctuation, stopwords, small words
   -Extract keywords from the remaining text
5)  Search for special html tags (table row, table column, hyperlink, list item
      etc) in the pre-hyperlink or post-hyperlink area. Since such tags denote a
      change on the document structure, they should limit the hyperlink's
      neighboring area.
   -Remove complete html tags inside the area. For example <B>, <I>, <IMG...> etc.
   -Remove punctuation, stopwords, small words
   -Extract keywords from the remaining text
```

Fig. 4. The procedure of keyword extraction from hyperlinks.

2-8

ontology. Thus, it is slow and tends to provide wrong mappings.

The first step to improve mapping is to reject the irrelevant senses of the terms in the ontology, as described previously. In order to further improve mappings, it is essential to reduce the number of senses examined for each keyword k by removing senses that are completely irrelevant in the scope of the ontology. To do this, for each keyword k in $\{k_j\}$, we compute the similarity of each of its senses to the senses of all other keywords in $\{k_j\}$ and keep the set of senses that gets the highest similarity score. This set is characterized by high self-correlation and gives significantly better results.

If the keyword set contains n keywords, we first create all the possible sets of senses that contain one sense for each keyword (cardinality n). Since the Wu and Palmer similarity measure can be employed for pairs of senses, we compute the average Wu-Palmer similarity for all the different pairs of senses in each set. For example, for the words *guitar, flute,* and *wind,* WordNet provides 1, 3, and 8 senses, respectively. For all the 24 triplets of senses, the average Wu-Palmer similarity for all pairs of senses ($\frac{n\cdot(n-1)}{2}$ combinations, 3 for each triplet) is computed. The keyword set (*guitar, flute, wind*) gives a score of 0.8 to the triplet of meanings (*guitar, flute/transverse flute, wind instrument/wind*) and less than 0.5 to any other combination. Similarly, the set (*storm, cloud, wind*) gives a score of 0.8 to the triplet of meanings (*storm/violent storm, cloud, wind/air moving*) and lower scores to any other combination. These are indications that "*wind*" is closer to the sense of "*wind instrument*" when it appears in the set (*guitar, flute, wind*), whereas it is closer to the sense of "*wind as weather phenomenon*" in the set (*storm, cloud, wind*). As a result, the first set will be mapped to the set of ontology terms (*string instrument, wind instrument, wind instrument*) using an ontology on music, whereas the second set will not be mapped at all to a set of terms related on music.

Since the most appropriate senses for each keyword have been selected, each keyword in $\{k_j\}$ is mapped to the closest ontology term as previously shown. As a result, each k_j is mapped to a term t_i with a similarity s_j. It is expected that more than one keyword in the set of keywords $\{k_j\}$ are mapped to the same ontology term, or to no term at all, so the cardinality of $\{t_i\}$ is usually smaller than that of $\{k_j\}$. The weight assigned to each term t_i is the average similarity of all keywords k_i mapped to t_i, taking into account the respective weight n_j of each keyword k_j. The weight is computed using the formula:

$$r_i = \sum_{k_j \to t_i} (n_j \cdot s_j) \Big/ \sum_{k_j \to t_i} n_j. \quad (1)$$

So, each document d_i is represented as $(\mathrm{URL}_i, \{(t_i, r_i)\})$, where $r_i \in [0, 1]$ since $s_j \in [0, 1]$.

5.4 Clustering Module

In this phase, the documents are fed into the clustering module. To be able to run the clustering algorithm, we need a similarity measure for documents. A lot of research has been conducted in this field (see [16]). One contribution of THESUS is the introduction of a novel similarity measure between sets of keywords.

5.4.1 Similarity Measure

The similarity measure introduced in [17] is used for clustering the documents and for answering queries. Both documents and users' queries are seen as weighted sets of terms and this measure is employed to evaluate the similarity between Web documents or between a document and a query. As a result, queries may retrieve documents that do not contain a certain keyword but a synonym, hypernym, or hyponym of it, and that are nonetheless very relevant.

Let us not forget that we aim at computing similarity between sets of *weighted* words (the ontology terms), and not simply words. There has been very little research on similarity measures between sets of elements (see [13], [11], [24]). The similarity measure employed in THESUS is a generalization of Wu and Palmer's [34] measure and is defined in the following. A detailed discussion on the properties of this measure can be found in [17].

Notations:

1. Let Ω represent the ontology (set of terms in a hierarchy).
2. We use **cursive capitals** $\mathcal{A}, \mathcal{B}$ to represent sets of weighted words, such as: $\mathcal{A} = \{(w_i, k_i)\}$ and $\mathcal{B} = \{(v_i, h_i)\}$, with $k_i, h_i \in \Omega$, and $w_i, v_i \leq 1$.

We define:

$$\varsigma(\mathcal{A},\mathcal{B}) = \frac{1}{2}\left[\left(\frac{1}{K}\sum_{i=1}^{|A|}\max_{j\in[1,|B|]}(\lambda_{i,j}\times S_{W\&P}(k_i,h_j))\right) + \left(\frac{1}{H}\sum_{i=1}^{|B|}\max_{j\in[1,|A|]}(\mu_{i,j}\times S_{W\&P}(h_i,k_j))\right)\right], \quad (2)$$

where

$$\lambda_{i,j} = \frac{w_i + v_j}{2 \times \max(w_i, v_j)},$$

$$K = \sum_{i=1}^{|A|} \lambda_{i,x(i)},$$

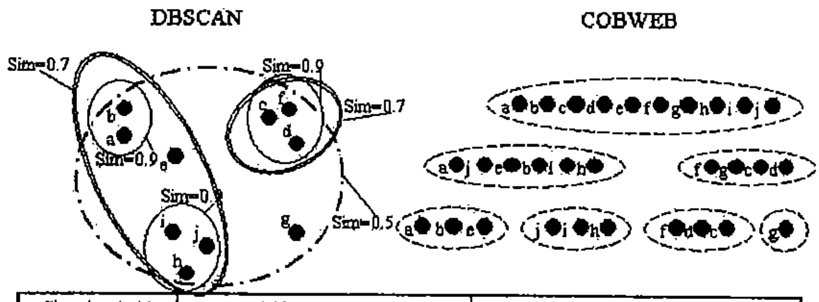
and

$$x(i) = x | \lambda_{i,x} \times S_{W\&P}(k_i, h_x) = \max_{j\in[1,|B|]}(\lambda_{i,x} \times S_{W\&P}(k_i, h_j)).$$

Simply put, K is a normalizing factor (the sum of all the $\lambda_{i,j}$ used). $\mu_{i,j}$ and H are defined similarly.

Intuitive explanation. The weight factors give less importance to terms that do not clearly describe the document. For instance, if a document is described by a term with a low weight, we do not want this weight to play too great a role in the *commonality* and the *differences* with other terms describing other documents. The *commonality* is calculated by finding the terms in document B that are closest to those in document A, and the *differences* are calculated by taking the terms in B and finding those that they are close to in A. As far as the weights are concerned, $\lambda_{i,j}$ and $\mu_{i,j}$ reduce the overall impact of terms with low weights. We also note that regardless of the values of the weights w_i and the maximal value for $\lambda_{i,j}$ (respectively, $\mu_{i,j}$) is equal to 1 and is reached for $w_i = v_j$.

2-10



Sim. threshold (DBSCAN only)	DBSCAN clusters			COBWEB clusters			
0.5	{a,b,c,d,e,f,g,h,i,j}			{a,b,c,d,e,f,g,h,i,j}			
0.7	{a,b,e,h,i,j}		{c,d,f}	{a,j,e,b,i,h}		{f,g,c,d,}	
0.9	{a,b}	{h,i,j}	{c,d,f}	{a,b,e}	{j,i,h}	{f,d,c}	{g}

Fig. 5. Clustering hierarchies produced by THESUS.

We refer to [26] for more details on the similarity measure its properties.

5.4.2 Web Document Clustering Algorithms

The problem is considerably different compared to the case of points in a metric space since there are no coordinates and ordering. We only have a similarity measure between the objects to be clustered, which are sets of (weighted) strings that correspond to concepts of an ontology. The first algorithm used in THESUS is a modification of the density-based algorithm DBSCAN [10]. Modification details and the algorithm itself are described in [17]. An incremental hierarchical algorithm, a variation of the COBWEB algorithm [12], has also been implemented.

The two algorithms operate differently and produce different final partitioning. In the modified version of DBSCAN, each document D of a cluster must contain a *minimum number of neighboring documents* (MinDoc). Neighbors of D are the documents whose similarity to D is higher than or equal to a *minimum similarity* (MinSim) threshold. DBSCAN considers as *"noise"* the documents that are not density-connected [10] with other documents. DBSCAN produces a flat clustering scheme. By decreasing MinSim, the size of clusters increases, several clusters are merged, and fewer documents are considered as noise. By repeating DBSCAN in the same document set for decreasing MinSim values, we produce a pseudohierarchy of clusters. On the other hand, COBWEB puts every document in the same cluster in the first level and in one of its children clusters in the next level. It merges or splits clusters aiming to increase the quality of the produced clusters. It produces a hierarchy of clusters, in which each level contains all the documents in the set, but the number of clusters increases from the top to the bottom levels. The two clustering schemes are depicted in Fig. 5.

5.5 Query Module

The clusters resulting from the clustering module can be exploited to answer user queries, in a more meaningful way. Let $q = \{k_i\}$ a query, where k_i are keywords defining the user's interest. The set $\{k_i\}$ is mapped to a set of ontology terms $\{t_j\}$, thus converting q to a query with ontology terms $q' = \{t_j\}$. Subsequently, the cluster(s) relevant to q' are identified by computing the similarity between q' and the clusters. Clusters whose similarity to q' surpass a threshold are considered relevant and the query. The query-processing module focuses within the relevant cluster(s) and computes the similarity of q' to the description of each document. Results are grouped by cluster and ranked within each cluster based on their similarity to q'.

In the experimental results that follow, the evaluation is focused on the "information extraction" and on the "clustering" modules. More specifically, hyperlink information extraction is evaluated for single pages and sets of pages employing the incoming and outgoing hyperlink semantics. Experiments on clustering test the ability of the similarity measure and the clustering algorithm that we plug it into, to group together pages based on incoming links semantics.

6 Testing and Experiments

In this section, we present an extensive experimental evaluation of THESUS. The section begins with some performance indications on the system. Two sets of experiments follow: one investigating the value of incoming links semantics in providing characterizations for a page or a set of pages and one testing the capabilities of the system in clustering sets of URLs by comparing to the results of clustering done by humans.

In all the tests concerning Web page characterization, we used unbiased "blind" testers to evaluate the results of THESUS, and those of other systems involved. The testers have no knowledge of the system that provided the characterization; thus, the results are not skewed in favor of our system. For the clustering test, we used a predefined (and preclassified by humans) collection of URLs.

A complementary set of experiments that test the ability of the system to characterize and, consequently, cluster large document sets can be found in [17]. In those experiments, different information sources (content, inlinks, user provided summary) and different similarity measures

2-11

are used in order to evaluate the performance of the THESUS clustering module.

6.1 System Performance

The first three modules of the THESUS architecture (information acquisition, extraction, and enhancement) take as input an ontology and, optionally, an initial set of documents and produce an extended set of documents and links, enriched with keywords, semantics, and the respective weights. This procedure creates the information pool for the THESUS clustering and querying modules.

The duration of the crawling procedure is mainly affected by the network bandwidth and the architecture of the crawling system itself as well as by the number of documents collected. In a prototype implementation, we used an ontology on musical instruments (containing 52 terms, http://www.db-net.aueb.gr/thesus/onto/instrum.rdf), we started from two directory pages on music (those of DMOZ and Google directory) and crawled for 12 levels. The crawling was performed using a single computer (a Pentium III PC with 256MBytes RAM and an IDE hard drive), in a multithreaded approach, attempting to take advantage of the available network bandwidth (10MBps). The information acquisition took place simultaneously to the crawling phase. In less than a day, we collected hyperlink information for approximately 190,000 documents. However, only 4,000 of these documents have at least two incoming links, which proves the weak connectivity of the collected document set. We may increase incoming link information by using a Web service that provides backward links for the documents in our collection (e.g., Google's backward link service, [14]).

The performance of the information enhancement module is affected by the size of the keyword set of each document and by the size of the ontology. As the number of keywords in the set increases, the number of combinations of senses that should be considered for each document increases. An increase in the number of terms in the ontology, increases the number of pairs (keyword, term) that should be computed. The processing time for the whole collection was about 3 hours.

The clustering phase is carried out offline, but is repeated many times with different parameters, until the optimal clustering schema is achieved. Clustering quality is validated using several internal measures [18]. The clustering time depends on the number of documents and on the algorithm used and can be accelerated by caching information that is calculated repeatedly. In the current implementation, we precalculate the similarity-matrix between all pairs of documents in the set and store it in main memory. For the running example, if the similarity is stored as a byte number approximately 33.6 Gigabytes $(190,000^2/2)$ of memory are needed to store the matrix. We decide to cluster only the 4,000 documents having at least 2 incoming links. The similarity matrix was computed in 38 seconds and the set was clustered, into 36 clusters in 0.8 seconds, using DBSCAN.

6.2 Keyword Extraction from Incoming Links—Evaluation

6.2.1 Page Characterization

In order to demonstrate the capabilities of the information extraction module in characterizing the contents of a Web document, we selected 50 URLs and obtained their incoming links characterization using at most 100 incoming links ([14]) for each URL. We aggregated keyword appearances by source URL (`weightedSourceKeywords`—how many source URLs use each keyword in the description of the target URL) and kept the top 10 keywords for each target URL. This description, and the descriptions of Altavista and Google for the same pages, were presented to a group of testers, which rated their quality (1 equals very bad and 5 equals very good). In more than 50 percent of the cases, THESUS' descriptions were considered the best of the three. The average rating for THESUS results was 3.7 out of 5, outscoring the other two systems (Altavista—1.9, Google—3.4). Some of the URLs used in the test, were listed by Google and Altavista's directory and described *by human editors*, whereas THESUS' descriptions were *automatically created* and were not based on the contents of the page.

6.2.2 Group Characterization

Demonstration of the group semantics operators. In order to demonstrate the use and value of the operators that assign semantics to group of pages, we performed an experiment on the homepages of certain London museums, listed in Google: http://directory.google.com/Top/Regional/Europe/United_Kingdom/England/London/Arts_and_Entertainment/Museums.

If we call the set of the homepages' URLs {U}, the set of pages that point to them {I}, and the set of pages pointed by them {O}, then we get the following results:

- weightedGroupKeywords ({I},{U}) = museum 681, london 264, home 185, freud 102, belfast 100, hms 99, national 96, maritime 92, link 87, soane 84, holmes 81, victoria 80, albert 80, etc.
- weightedTargetKeywords ({I},{U}) = london 17, museum 14, Web site 13, site 12, visit 11, home 11, Web 10, world 10, information 10, history 9, library 9, page 9, art 8, british 8, national 8, etc.
- weightedGroupKeywords ({U},{O}) = museum 56, link 40, war 24, london 17, information 16, collections 16, shop 15, education 15, imperial 14, news 13, exhibitions 13, soane 12, etc.
- weightedSourceKeywords ({U},{O}) = museum 11, exhibition 9, collections 9, shop 9, exhibitions 9, online 8, home 7, news 7, events 7, visit 7, information 7, etc.

The first operator counts the number of keyword occurrences in the links of the first 100 pages that point to {U}. The second operator counts the number of pages that point to {U} using a certain keyword. In order to find the incoming links of a page, we used Google's backward links service. The third operator counts the number of times a keyword appears in all the outgoings links of pages in {U}. The last operator counts the number of pages in {U} pointing to a page using each keyword.

Comparing the two former sets of keywords, we see that keywords with a high value of TIMES in the first set do not appear in the second set. These are keywords that appear many times but only in links to the same page (e.g., the names of each museum appears in most of the links to its homepage). Although, *weightedGroupKeywords* assigns high values to these keywords, *weightedTargetKeywords* counts

2-12

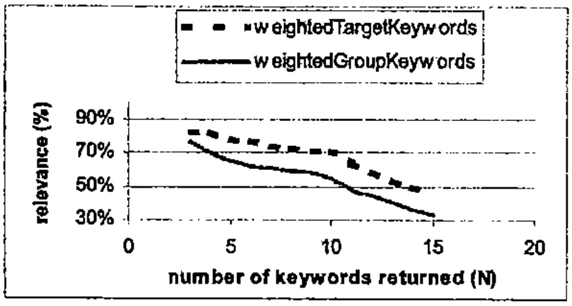


Fig. 6. Relevance percentage of group characterization.

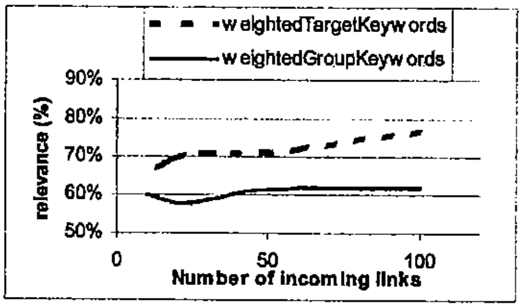


Fig. 7. How relevance is affected by the number of incoming links.

them only once, since they characterize only one page of the target set. The contents of the second set characterize the group of URLs {U} as a whole. The two latter sets show the Museums' interest as those are expressed in the outgoing links of their homepages. The words *exhibition*, *collection*, *online*, and *shop* represent things that are usually found on the Web site of a Museum. In order to perform more thorough semantic operations, we need to construct an ontology on *culture* for instance.

Efficient characterization of groups of pages. In this experiment, we extract the keywords characterizing sets of pages listed under the same topic in Google's directory and measure their relevance to the topic. *WeightedGroupKeywords* and *weightedTargetKeywords*operators are employed for the characterization. The experiment is performed on 30 Google topics. For different values of N (number of keywords extracted), we calculate the percentage of keywords provided by THESUS that are on-topic; that are contained in the path description provided by Google. The results are depicted in Fig. 6.

The examples justify the intuition, that THESUS operators allow the characterization of a set of pages based on incoming links to a very satisfactory degree compared to characterization assigned by humans (which is about 80 percent, i.e., eight out of 10 keywords provided by THESUS are included on the human characterizations). Moreover, aggregation of the results per target page (operator *weightedTargetKeywords*) yields better results than aggregation per keyword (*weightedGroupKeywords*). The above results indicate that an increase on the number of keywords used for the characterization of a page decreases the total relevance and suggests keeping at most 10 keywords to characterize the set.

It is noteworthy that, in the results presented above, a maximum of 100 incoming links was used to characterize each page in a set. Visiting 100 documents in order to characterize one is not efficient, even when the documents have been prefetched or preprocessed. In another test, we characterize the same sets of pages, using fewer incoming links. We take into account the top seven keywords (having 75 percent relevance according to results of Fig. 6). Fig. 7 illustrates how the number of incoming links affects the relevance of the extracted characterization to the topic. The figure shows that a reduced amount of incoming links (i.e., 20 links) gives comparable results to the ones achieved by taking into account the 100 incoming links (less than a 10 percent loss in accuracy in both operators).

6.3 Testing THESUS Ability to Discover Thematic Subsets

In order to test THESUS ability in clustering a set of pages, we cluster with THESUS, a set of URLs that have been manually clustered. Then, we measure the quality of the clustering scheme using an external clustering quality measure, ***entropy*** [30] as defined in [31].

The core set of pages contains the URLs suggested by DMOZ directory [27] under a few categories in the Music → Styles path. The categories are: *hip-hop, experimental, world, blues, Indian music, pop, dance, filk, electronic, electronica, country, dance, polka, early music, new age, lounge, rhythm and blues, folk, classical, gamelan, opera, bluegrass, rock, and easy listening* and contain 481 URLs in total. We characterized documents by extracting keywords from incoming links and enhancing information with semantics using a manually created ontology on music (with 177 different terms/concepts) and WordNet.

We measure the effectiveness of the two clustering algorithms (DBSCAN and COBWEB). The minimum value of ***entropy***, which is considered to be the best, is 0 and can be achieved when each cluster contains documents of one category only. The worst (maximum) value of entropy is log(N), where N is the number of different categories, when each cluster contains one document from each different category. The maximum entropy is compared to the entropy of each clustering scheme (with DBSCAN and COBWEB) for different values of N. The goal is to decrease entropy by putting documents of the same category to the same cluster, thus increasing the homogeneity of clusters.

In Tables 1, 2, and 3, we evaluate the resulting clustering schemes. The table columns present the total number of documents to be clustered in each case, the number of different categories they fall into, the value of the minimum similarity parameter (for DBSCAN only), the number of documents assigned in a cluster (DBSCAN only), the value of entropy for the resulting clustering scheme, the maximum entropy for the set of documents, and the decrease ratio we achieved using our system.

6.3.1 Evaluation of the Results

We clustered documents using DBSCAN and semantics first (Table 1) and then DBSAN and extracted keywords. Clustering using the semantics is significantly faster than clustering using keywords since the similarity between ontology concepts is precalculated once for all the 177×177 combinations (using Wu and Palmer's measure on the

TABLE 1
Comparison of the DBSCAN Results for Various Input Parameters (minDocs=3) Using Ontology Terms

Documents	Categories	MinSim	Documents Clustered	Clusters	Entropy	Maximum entropy	Entropy decrease
400	23	0.55	115	8	0.96	1.36	29%
400	11	0.8	24	3	0.65	1.04	38%
191	22	0.55	74	8	0.76	1.34	43%
191	7	0.8	16	2	0.47	0.85	44%
457	27	0.65	132	6	1.13	1.43	21%
457	5	0.8	10	2	0.47	0.70	33%
182	23	0.6	71	7	0.79	1.36	42%
						Average decrease	36%

TABLE 2
Comparison of the DBSCAN Results for Various Input Parameters (minDocs=1) Using Keywords

Documents	Categories	MinSim	Documents Clustered	Clusters	Entropy	Maximum entropy	Entropy decrease
396	8	0.95	18	7	0.16	0.90	82%
396	13	0.85	31	4	0.67	1.11	40%
295	12	0.95	17	4	0.60	1.08	45%
295	20	0.8	48	6	0.90	1.30	31%
						Average decrease	49%

TABLE 3
COBWEB Results (Using Concepts) for an Increasing Number of Documents

Documents	Categories	Clusters	Entropy	Maximum entropy	Entropy decrease
25	2	7	0.11	0.30	63%
37	3	6	0.27	0.48	43%
55	4	12	0.22	0.60	64%
78	5	7	0.45	0.70	35%
100	10	33	0.58	1.00	42%
160	24	56	0.57	1.38	59%
244	26	85	0.59	1.41	58%
				Average decrease	52%

ontology), whereas the similarity between keywords must be calculated for every new pair of keywords. The latter computation significantly slows down the process since it involves accessing WordNet. In both cases, we notice a significant decrease in the entropy of the system. The decrease is higher when using keyword characterization (49 percent of the maximum entropy, Table 2) rather than using the respective semantics (36 percent, Table 1).

In COBWEB, documents are clustered based on their semantics. The entropy is again lower than the maximum entropy (a decrease of 52 percent). As can be deduced from the results in Table 3, after a certain number of documents (approximately 100) have been clustered, entropy remains constant. This means that the clustering algorithm needs many documents as input in order to perform well.

6.4 Conclusion

When comparing our results on various data sets (ranging from hundreds to thousands of pages) to those provided by Web search engines or Web directories, we come to the following conclusions:

1. Querying links semantics yields correct information on the semantics of pages, at least comparable to those established with full text querying, as proved in Sections 6.2 and 6.3.
2. Defining a measure of similarity between Web pages, which is based on semantics conveyed by the links between pages, lets us discover clusters of pages that have similar semantics and have similar connectivity features.

7 Conclusions

7.1 Summary—Contributions

Searching in the World Wide Web is a task of very high importance, in social and financial terms, since hundreds of millions of users worldwide, with diverse profiles, are searching for pages relevant to their requests. Currently, this task is mainly carried out by submitting keyword

2-14

queries to search engines. The search criteria are based on the pages' contents ignoring the additional semantics emanating from links. The results contain pages that exactly match the majority of terms in the query. Information retrieval algorithms rank the results based on the distance of the query to the document contents. However, document authors may add misleading contents to their documents aiming to manipulate ranking algorithms and affect search engines results. With hyperlink information, the introduction of such bias becomes harder.

In this paper, we capitalize on this observation. We present THESUS model, language, and prototype system, which can be applied for selecting, processing, and querying thematic subsets of the World Wide Web. The salient features of THESUS language are 1) it extracts and exploits semantics from links and 2) it enables querying and organization of pages based on aggregate connectivity features and link semantics. We also present the implemented THESUS system that collects URLs based on a set of keywords, extracts keywords from the collection's incoming and outgoing links (by processing link's neighboring text in the source URL), maps extracted keyword to semantics, and populates a relational database with all this information. A clustering module is able to detect semantically cohesive groups of the initial document set that have similar semantics, assigned by incoming links.

7.2 Further Work

Further work will be devoted to the following;

- Proprietary Web crawling based on semantics. The thematic crawler developed so far follows links to pages only when the extracted link information exactly matches the set of keywords of interest. It would be useful to be able to enhance the initial keyword set with semantics and follow links based on semantics and not on keywords. Current work on the Web crawler also aims on providing an initial estimation on pages' importance using semantics and connectivity.
- Composite ranking of results. The ranking of the results should take into account two different issues: 1) the importance of a page, as it is defined by a Page Rank [5] and 2) the number of matching keywords between the query and the page's description emanating from incoming links.
- Link position in the page. The location of the source is interesting information. For instance, many similar links are often located in a <LI></LI> list.
- Test other clustering algorithms. The tests performed with DBSCAN gave interesting results. However, a lot of pages were perceived as noise and, consequently, not clustered. Additionally, since DBSCAN is not an incremental algorithm, the clustering process should be repeated every time the initial document set changes. The experiments conducted with the COBWEB algorithm [12] performed better compared to the manual classification. The ultimate goal is to be able to discover cohesive groups of pages that are semantically close and form semantic clusters, in other words, to discover THESUs rather than constructing them.
- Design of a query module, giving the user the ability to perform advanced queries for thematic hubs, authorities, cocitations, and couplings. This is for the moment feasible only at keyword level. However, we are working on employing link semantics in the search for strongly interconnected pages with similar semantics. This is the first step on integrating semantic information and link structure in the clustering algorithm.

ACKNOWLEDGMENTS

"The greatest help we can give others is not to share our riches with them, but to discover theirs," by Louis Lavelle, *L'Erreur de Narcisse*, 1939.

REFERENCES

[1] R. Al-Halami, R. Berwick, *WordNet, An Electronic Lexical Database*, C. Fellbaum and G. Miller, eds. Cambridge, Mass.: MIT Press-Bradford Books, 1998.
[2] R. Armstrong, D. Freitag, T. Joachims, and T. Mitchell, "Web Watcher: A Learning Apprentice for the World Wide Web," *Proc. AAAI Spring Symp. Information Gathering from Heterogeneous, Distributed Environments*, pp. 6-13, Mar. 1995.
[3] G. Arocena, A. Mendelzon, and G. Mihaila, "Applications of a Web Query Language," *Proc. Sixth Int'l World Wide Web Conf.*, Apr. 1997.
[4] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph Structure in the Web," *Proc. Ninth Int'l World Wide Web Conf.*, May 2000.
[5] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Proc. Seventh Int'l World Wide Web Conf.*, Apr. 1998.
[6] S. Chakrabarti, M. Berg, and B. Dom, "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery," *Proc. Eighth Int'l World Wide Web Conf.*, May 1999.
[7] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan, "Automatic Resource List Compilation by Analyzing Hyperlink Structure and Associated Text," *Proc. Seventh Int'l World Wide Web Conf.*, 1998.
[8] S. Chakrabati, B. Dom, D. Gibson, J. Kleinberg, S. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Mining the Link Structure of the World Wide Web," *Computer*, vol. 32, no. 8, pp. 60-67, Aug. 1999.
[9] The DARPA Agent Markup Language Ontology Library, http://www.daml.org/ontologies/, 2004.
[10] M. Ester, H.P. Kriegel, J. Sander, and X. Xu, "A Density Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *Proc. Second Int'l Conf. Knowledge Discovery and Data Mining ACM-SIGKDD*, 1996.
[11] T. Eiter and H. Mannila, "Distance Measures for Point Sets and Their Computation," *Acta Informatica J.*, vol. 34, 1997.
[12] D. Fisher, "Knowledge Acquisition Via Incremental Conceptual Clustering," *Machine Learning*, vol. 2, pp. 139-172, 1987.
[13] A. Gionis, D. Gunopulos, and N. Koudas, "Efficient and Tunable Similar Set Retrieval," *Proc. ACM SIGMOD Int'l Conf. Management of Data*, pp. 247-258, May 2001.
[14] The Google.com search engine, http://www.google.com/, 2004.
[15] N. Guarino, "Formal Ontology and Information Systems," *Proc. First Int'l Conf. Formal Ontologies in Information Systems (FOIS '98)*, pp. 3-15, June 1998.
[16] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On Clustering Validation Techniques," *J. Intelligent Information Systems (JIIS)*, vol. 17, nos. 2-3, pp. 107-145, 2001.
[17] M. Halkidi, B. Nguyen, I. Varlamis, and M. Vazirgiannis, "THESUS: Organizing Web Document Collections Based on Semantics," *Very Large DataBases J.*, special edition on semantic Web, to appear.
[18] M. Halkidi and M. Vazirgiannis, "A Data Set Oriented Approach for Clustering Algorithm Selection," *Proc. Principles of Data Mining and Knowledge Discovery, Fifth European Conf. (PKDD '01)*, pp. 165-179, 2001.

2-15

ELSEVIER

JOURNAL OF OPERATIONS MANAGEMENT

www.elsevier.com/locate/dsw

編號：3

Abstract

1. Introduction

Recent reports indicate annual on-line consumer sales have increased 40% in the past year (Moore, 2002). Despite this growth in online retailing, many companies continue to struggle with the development of effective Internet-based systems. Even with the proper supporting infrastructure, many established businesses have struggled with their expansion into the on-line arena. These organizations have discovered that their sales fall far short of what they had originally expected. For many, the goal was to dramatically increase sales and profitability. Unfortunately, many have discovered that such ambitious goals represent a brass ring that is difficult to reach.

What causes some of these businesses to succeed and others to fail? There are many underlying reasons why so many businesses fail when competing in the on-line arena. Failing to pay attention to customer desires and required supporting infrastructure are two common lapses. Specifically, most companies do not comprehend the driving factors that sway their customers to use the Internet as a purchasing tool. This can lead to poor performance when customer concerns are neglected. A recent study reports that 28% of all attempted purchases failed and four out of five ordering on-line have experienced at least one failure (Boston Consulting Group, 2000).

There are several driving forces that affect consumers' use of on-line purchasing. Research suggests that end user viewpoints and perceptions have a significant impact on adoption and implementation of new technologies (Agarwal and Prasad, 1999; Davis et al., 1989). Unfortunately, scores of websites are designed to have the latest animation or other new gimmicks and do not consider what the customer actually desires. These websites are attractive, but are often slow to load and very difficult to order from. Reports suggest that consumers are not likely to use the web to purchase goods and supplies after an initial unfavorable experience (Boston Consulting Group, 2000).

The main goal of this study is to investigate how end users' viewpoints and preferences influence the use of the Internet as a purchasing avenue. This research examines data gathered from a survey of 416 customers of Office Depot, a major Internet retailer of commodity office supplies. Office Depot is a leading retailer of office supplies, with 825 office supply super stores in 46 states and US$ 11.6 billion in sales for 2000 (Office Depot, 2001). More importantly, for the purposes of this study, Office Depot is also a leader in Internet sales with both a relatively lengthy (for Internet companies) and successful track record. They have been selling online for 5 years and had annual Internet sales of US$ 849.5 million for 2000, on which they actually make a profit unlike the vast majority of Internet only startups (Office Depot, 2001). Office Depot is widely considered to be the leader in the online office supplies market (Gulati and Garino, 2000; Haddad, 2001; Troy, 1999; Warner et al., 1999). Their goal is to move 50% of the orders from its business services division to the net.

Office Depot provides an excellent opportunity to study online purchasing patterns for several reasons. First, the company is a pioneer in this area and one of the few companies actually showing a profit on Internet sales. Second, Office Depot sells to a cross-section of American business, from Fortune 500 firms to the mom and pop stores in your neighborhood. This allows us to develop a well-stratified sample designed to compare the Internet usage patterns of a variety of businesses. Third, Office Depot stocks approximately 8000 commodity type products. Their customers do not require highly specialized applications or products. Thus, the results of surveying Office Depot's customers are readily generalizable to those companies that are considering adoption of Internet purchasing. Finally, Office Depot straddles the line between B2B and B2C applications. It deals directly with large, Fortune 500 firms using customized applications that can be considered B2B sources of office supplies. It also deals with much smaller businesses with fewer employees that are more characteristic of B2C transactions. Office Depot, thus, offers a view of both the B2B and B2C sectors.

2. Background and research questions

Motivation for this research stems from three areas. First, the massive expansion of the Internet as a tool for a wide variety of businesses implies that it will act as a catalyst for radical change. Second, many corporations and businesses have experienced mixed results using the Internet to exchange data and information. Finally, end users' views and beliefs should impact how technology is adopted and implemented. Most of the companies and media in the tidal wave of e-commerce appeared to treat customers as a single homogenous group, in defiance of decades of marketing segmentation research. Now that the initial flurry of Internet activity has passed, we seek to identify groupings of customers with different approaches and attitudes to online purchasing.

Modern websites offer consumers the ability to perform a variety of tasks ranging from purchasing materials to market research. Consumers are faced with a tough decision when choosing between competing websites. Based on content alone, many of the websites appear to be virtually identical. What causes one website to be successful while others struggle and eventually fail? Research indicates that factors such as perceived ease of use, perceived usefulness and attitude impact the adoption of new technologies (Davis et al., 1989). We use these technology acceptance factors to classify individuals into groups that provide insights into their viewpoints about the Internet as a purchasing medium. In particular, our study shows differences in both contextual factors related to the adoption of Internet purchasing and outcome factors (i.e. satisfaction and improved efficiency).

The following sections develop three research hypotheses to be examined in this study. Fig. 1 provides a model of our basic research objectives—we first identify groups of companies with similar attitudes toward Internet purchasing, then we test for differences in contextual factors that lead to differences and outcomes in terms of organizational and site performance. It is important to note here that there are at least two groups of factors that might influence buying behavior: properties of the individual buyers and properties of the buying companies. Both are important, but do not necessarily match or complement each other. Thus, we will examine aspects of both individual users and companies as a whole throughout the remainder of this paper and will not explicitly differentiate whether a particular property is primarily associated with an individual or company.

2.1. *Internet purchasing acceptance factors*

Many researchers have examined the factors that influence the acceptance and usage of technology in the workforce. The theory of reasoned action (TRA) (Fishbein and Ajzen, 1975) and the Technology Acceptance Model (TAM) (Davis et al., 1989) are two prominent theoretical models that have been widely used to explain technology acceptance. The Technology Acceptance Model expanded on the theory of reasoned action by extending it to the domain of end user acceptance of the technology. Since the introduction of TAM, the model has been validated in numerous settings and shown to be a good predictor of end usage of a new information technology. The Technology Acceptance Model has been successfully applied in numerous settings with several technologies, including information technology (Davis et al., 1989), bank broker workstations (Lucas and Spitler, 1999), and voice mail systems (Straub et al., 1995). Agarwal and Prasad (1999) indicate that attitude, beliefs about usefulness, ease of use and comfort level impact a user's behavioral intentions towards the acceptance of new technologies. Extending TAM to the Internet, Jiang et al. (2000) were able to determine that one of the driving factors influencing the utilization of the Internet was experience with the technology. In addition to testing the TAM in various information technology settings the model has been tested in various organizational settings. Igbaria and Zinatelli (1997) examined the impact of the technology acceptance factors in small organizations. They were able to determine that perceived usefulness and perceived ease of use were strong predictors of system usage (Igbaria and Zinatelli, 1997). These factors are the foundation for our initial research question. We will extend this research by testing to see if groups of individuals in small organizations exist that have similar beliefs about the use of Internet purchasing. The groups will be developed using the factors of attitude, perceived usefulness, perceived ease of use and comfort level. Our initial research question is

Research question: Can users of Internet purchasing in small organizations be classified into distinct groups based on end users' viewpoints and preferences regarding the Internet?

Once we develop groups of companies with similar beliefs regarding Internet purchasing, we will then examine contextual factors that affect Internet purchasing adoption (Hypothesis 1) and outcome factors that result from the use of Internet purchasing (Hypothesis 2). Fig. 1 shows a model of this sequential approach to data analysis.

2.2. *Contextual factors that affect Internet purchasing adoption*

Numerous studies have examined the notion that individual difference plays a crucial role in the implementation of new technologies (Agarwal and Prasad, 1999; Harrison and Rainer, 1992). Gender, age, experience and personality traits have been examined in the context of end user computer skills (Venkatesh and Morris, 2000; Agarwal and Prasad, 1999). The current study explores how the variables of education level, tenure in the workforce and annual training received impact an end user's viewpoint regarding

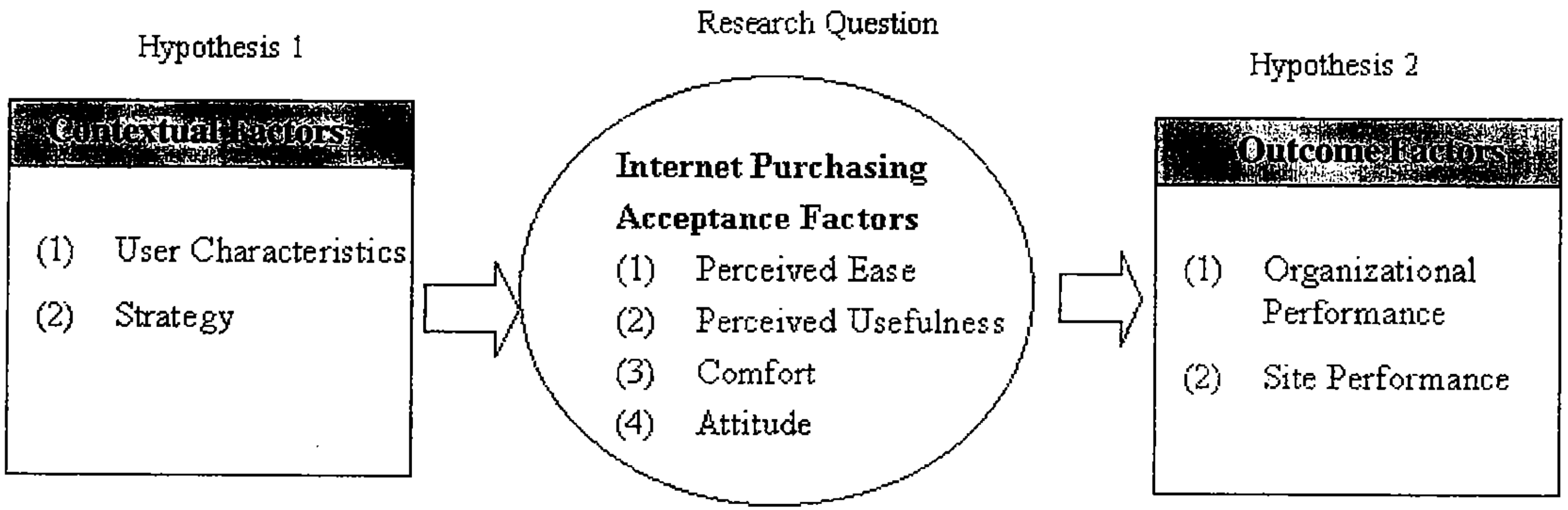


Fig. 1. A model of Internet purchasing adoption, context and outcome factors.

Internet purchasing. While these factors were not specifically tested in the original TAM, there is substantial evidence that they have significant influence on user behavior. For example, Venkatesh and Morris (2000) examined the influence of various organizational behavior factors such as income, occupation, and education levels on perceptions regarding technology acceptance.

The majority of extant studies examine how these variables impact the implementation of new technologies. In contrast, the current study explores if the variables of education level, tenure in the workforce and annual training received impact an end user's viewpoint regarding the adoption of Internet purchasing. Unlike the previous studies, we seek to discover the impact of these various factors on a specific Internet based application. We define the adoption of Internet purchasing as the placement of at least one order over the Internet and utilize the number of orders placed (with Office Depot—the company being studied) as a proxy for intensity of adoption. Furthermore, we argue that in contrast to most adoptions of IT systems that represent a binary, yes/no type decision for the purchasing organization, the adoption of both Internet purchasing in place of more traditional methods and a specific website is more a matter of degree. For example, Reichheld and Schefter (2000) find that attracting new customers in pure play Internet businesses is up to 40% more difficult than in traditional brick-and-mortar services. Hence, we posit that the TAM applies in a post-adoption case setting and is useful for explaining not only adoption, but also the degree of loyalty and repeat usage of a particular website. While there are limitations (see Section 7) associated with applying these factors in this fashion, many other studies have employed a similar approach. Recent studies examine how the TAM relates to World Wide Web and Internet usage rates (Jiang et al., 2000; Lederer et al., 2000), while other researchers have extended the TAM factors to a post adoption case to examine how these factors change over time (Venkatesh and Davis, 2000, Venkatesh, 2000). Thus, we examine the following hypothesis.

Hypothesis 1a. Internet purchasing groups will exhibit differences from each other in terms of individual characteristics of education level, training received and tenure in the workforce.

There is an extensive stream of research in operations strategy that suggests that companies pursue different competitive priorities. There is also a high degree of consensus that competitive priorities represent the key decision variable for operations managers. Competitive priorities denote a strategic emphasis on developing certain operational capabilities that may enhance a corporation's position in the marketplace. Models of operations strategy content generally include four competitive priorities: cost, quality, flexibility and delivery. Although some conceptual studies suggest innovativeness and service as additional priorities, empirical research and strategy theories consistently stress the four basic priorities (Schmenner and Swink, 1998; Ward et al., 1998). Early research in this area focused primarily on manufacturing operations, yet numerous researchers have utilized this model over the past 10–15 years to examine non-manufacturing businesses (Roth and Van der Velde, 1996; Kellogg and Nie, 1995; Butler et al., 1996). Thus, the general model applies equally well to service operations, so we will use the more general term—operations strategy.

Skinner's (1969) seminal article proposed that companies must make choices regarding which competitive priorities should receive the greatest investment of time and resources. Companies make trade-offs between various priorities, based on their relative importance. Managers must choose a manufacturing priority, and then allocate their scarce resources accordingly (Garvin, 1993; Hayes and Wheelwright, 1979).

The current study assesses three competitive priorities related to the Internet purchasing: (1) reducing cost or improving convenience/delivery, (2) improving the administration of purchasing transactions and (3) improving the delivery accuracy, service and security. Krause et al. (2001) find that delivery accuracy, speed, and dependability are all key measures when assessing competitive priorities related to the purchasing. Since this study is examining Internet purchasing, we must also take into the account the influence of the electronic medium. Many studies that have examined the use of EDI and other data exchange technology cite administrative savings as one of the primary benefits of switching to the new system (Massetti and Zmud, 1996; Mukhopadhyay et al., 1995; Carter and Frendall, 1990). Thus, our expectation is that specific groups of companies will have differing strategic objectives regarding Internet purchasing and will also differ in terms of adoption rates. We, therefore, test the following hypothesis:

Hypothesis 1b. Internet purchasing groups have different strategic objectives.

2.3. Outcome factors resulting from Internet purchasing adoption

We assess two different dimensions of organizational performance. First, transactional performance measures the ease and accuracy with which day-to-day purchasing transactions are conducted by employing the Internet in place of traditional purchasing methods. Second, system performance provides a broader assessment of the cost of purchasing activities and the accuracy/availability of billing and supplies. Measuring performance at two different levels is a common approach in operations management research, since operations typically have a fairly direct effect on daily transactions (i.e. transactional performance) and a positive, but more indirect effect on business level or system wide performance (Boyer et al., 1997; Miller and Roth, 1994). We will, therefore, test if the various groups developed in our initial research question perceive differences in the performance impact of Internet purchasing.

Hypothesis 2a. Internet purchasing groups differ from each other in terms of performance outcomes—specifically, organizational performance is enhanced by the application of Internet purchasing.

Site performance measures refer to those items that have an impact on the individual productivity of the user of the website. Specific features of a website that we believe affect individual productivity levels include the ease of use, the accuracy of information on the site and the reliability of transactions. These items will eventually affect the end users' choice to use the website or that of a competitor. The scales used to assess the various components of Hypotheses 1–3 are described Section 4.

Hypothesis 2b. Internet purchasing groups perceive differences in website performance.

3. Research methods

3.1. Data sampling

Data was collected from customers of Office Depot that had placed at least one order using Office Depot's website within the prior year. Our initial database of contacts consisted of approximately 65,000 customers from Office Depot's Business Services Division. Since they had fairly extensive internally collected data regarding their larger customers, Office Depot asked us to focus on smaller companies of 100 or fewer employees. The original database represented a very high quality sample since all of the companies had current mailing information and a fairly substantial amount of descriptive information, including the total dollar volume of business done with the retailer over the past year, the total dollar volume of Internet orders, the number of orders and the number of orders placed over the Internet.

We used a two-part strategy to stratify our database:

(1) We computed an Internet usage variable (percent Internet orders) by dividing the number of Internet orders over the past year by the number of all orders during that year.
(2) Next, we randomly chose firms in various categories for the target sample. Table 1 shows how the initial database was stratified across the various categories.

- Category A represents customers who have tried Internet ordering once, but have primarily used traditional ordering by phone or fax for the bulk of their orders.
- Category B represents customers who have *only* ordered over the Internet, with no orders placed via traditional methods. This group is unique because it consists of single orders with no repeat purchases. There were over 30,000 companies in this category. The large amount of customers in category B occurred because of Office Depot's many promotions encouraging the use of their Internet site. Many Internet start-ups have discovered the hard way that this category of customers is not profitable since the costs of acquiring them are not recouped through repeat purchases.
- Category G represents customers who conduct business exclusively over the Internet. There were

Table 1
Initial contact sample

Category	Number of companies in initial database	Number of companies in contact sample	Percentage of contact sample	Responses per category (%)	Criteria used in selection
A	4,238	300	15.0	70 (46.6)[a]	Single order over the Internet Multiple orders with Office Depot >US$ 100 per order
B	32,210	300	15.0	51 (34)	Single Internet orders No other orders with Office Depot
C	2,461	300	15.0	75 (50)	10–25% Internet orders
D	2,211	300	15.0	62 (41.2)	25.1–50% Internet orders
E	3,864	300	15.0	44 (29.4)	50.1–75% Internet orders
F	2,985	250	12.5	51 (34)	75.1–99% Internet orders
G	17,438	250	12.5	49 (39.2)	100% Internet orders More than 1 order
Total	65,407	2000	100.0	416 (39.8)	

Values shown in the parentheses are in percentages.

[a] The column with number of companies in contact sample shows the number in each category that were pre-emailed. Actual surveys were then mailed to half of this group. For example, 300 companies in category A were pre-contacted by email, of those that did not object to being sent the survey, 150 were sent the actual survey. This represented 15% of our total sample. Within this category, we received 70 responses, or a 46.6% response rate.

approximately 17,000 of these companies that had multiple orders and conducted all of their business with Office Depot over the Internet.

- Categories C–F represent customers that have ordered a percentage of their orders over the Internet with each category based on quartiles. The sampling plan was not constructed solely on quartiles for two reasons: (1) the individuals who ordered just once on the website and those who ordered all of their supplies on the website would greatly skew the data; and (2) we specifically wanted to capture this information to observe differences in these individuals.

In all cases, when choosing companies to contact, we first sorted by last transaction date and then randomly selected firms that had placed orders within the last few months. Note the selection of seven categories and number of contacts within each category is subjective in nature. The goal was to prevent any one particular group from overly biasing the data. Notice that category B contains 30,000 individual; we wanted to avoid a proportionate representation of this group because Office Depot does not consider them to be their core customers.

Office Depot then pre-notified our target group to ask permission to conduct the study. Those who responded no were deleted from the list. This was done to ensure their customers that no private information would be disclosed, since many Internet consumers are quite sensitive about their personal information being sold or given to other organizations. Once Office Depot gave us the final contact list, they were no longer involved in the data collection or analysis phases. Office Depot and the respondents of the survey were provided summary data only.

The last step prior to data collection was to pre-test the instrument. We contacted three companies that agreed to fill out the pre-test and participate in a short interview regarding its readability and provide suggestions for improvement. Several useful suggestions were obtained regarding wording of questions and presentation of the survey. Overall, very minor modifications were made to the survey.

3.2. Data collection

A total of 1045 customers were contacted in our first round of mailings. The data were collected using both a printed survey and an electronic survey. The survey was distributed to those individuals who placed the order on the website within the last calendar year. Several factors helped to increase the response rate of the survey. Office Depot offered a US$ 15 rebate to all survey respondents and provided a letter stating their interest in the study and explaining that the authors were acting as independent, non-biased third party researchers. Two follow-up letters were then sent out in 2-week-windows. Approximately 5% of the mailings were returned due to incorrect addresses or the contact person having left the company.

The final response rate was 39.8%. A total of 416 usable surveys were collected out of the initial sample of 1045. Interestingly, the response rate for the printed survey was 41.4% and for the computer-based survey was 37.4%. We had expected the response rate for the computer survey to be lower since it has a higher entry barrier (computer access and skills). The overall response rate is higher than that seen in similar studies (Boyer et al., 1997; Duray et al., 2000; Kathuria, 2000). After we collected the data, companies with total employment greater than 1000 individuals were removed from the database. The final database used for analysis contained 395 individuals.

Consistent with previous studies we assessed non-response bias by conducting chi-square tests on the proportion of positive responses for three variables: the percent of respondents that received computer-based versus printed surveys, the proportion of respondents within each of the seven categories shown in Table 1 and industry membership (17 categories, self-typed by customers when ordering) (Johnson et al., 2002; Boyer et al., 1996). None of the chi-square tests revealed a significant difference. In combination with the high response rate, we conclude that there is no evidence to indicate non-response bias. Table 2 shows the breakdown of respondents across industries (the industries were defined by Office Depot and self-typed by customers). Table 3 provides an overview of the respondents.

4. Scale development

This section describes the scales used in the study. We used a combination of existing scales and newly developed scales in our study. Since we are using old and new scales, we tested the inter-item reliability of the items used in the scales. Each of the factors used in scales is described below. The data were tested to see if any significant differences existed between the two data collection methodologies. Based on Cronbach's alphas, response rates or scale means/variabilities, we found no significant difference between the two data sets. The analysis for the remainder of this paper combines both the print and electronic data. Table 4 gives the means, standard deviations and Cronbach's alphas.

Table 2
Industry membership

Industry	Number	Percentage
Education/schools	4	1.0
Business services	37	9.0
Membership/organizations	19	4.6
Wholesale trade	24	5.8
Accounting	3	0.7
Social services	7	1.7
Medical/health services	7	1.7
Retail/restaurants	21	5.1
Transportation/communications/utilities	12	2.9
Real estate	12	2.9
Legal	14	3.4
Government	6	1.5
Engineering/architecture/consulting	41	10.0
Manufacturing/printing	26	6.3
Insurance	11	2.7
Construction/contractors	12	2.9
Finance	14	3.4
Missing/other	126	34.3
Total	395	100.0

Table 3
Profile of survey respondents

	Mean	Standard deviation	Median
Total employment	43.27	99.43	12.50
Purchasing employment	2.23	8.16	1.00
Years in workforce	16.95	11.03	15.00
Years with current company	6.00	5.90	4.00
Years in current position	4.87	4.85	3.00
Computer hours per week	27.15	11.34	
Training hours per year	8.44	24.79	
Business located in:			
Single location	276	70.6[a]	
Multiple locations	115	29.1[a]	
Total	401	100.0[a]	
Education level			
High school	131	33.2[a]	
2 years degree	102	25.7[a]	
4 years degree	114	28.3[a]	
Graduate degree	54	12.9[a]	
Total	401	100.0[a]	

[a] Values shown are in percentage.

Table 4
Descriptive and reliability data for scales

	Mean	Standard deviation	Cronbach's alpha
Strategic factors			
GOALS	5.74	1.12	0.63
ADMINISTRATIVE	5.04	1.37	0.79
DELIVERY	5.70	1.28	0.80
Site performance			
SITE EASE	5.36	1.18	0.89
SITE ACCURACY	5.38	1.07	0.87
TRANSACTION	5.34	1.27	0.65
Organizational performance			
INTERNET IMPROVEMENT	4.57	1.54	0.90
COST REDUCTION	4.14	0.81	0.42[a]
PURCHASING EFFICIENCY	4.52	0.88	0.50[a]
Internet purchasing acceptance factors			
PERCEIVED USEFULNESS	4.94	1.49	0.95
PERCEIVED EASE	5.36	1.32	0.93
ATTITUDE	5.44	1.33	0.89
COMFORT	5.14	1.25	0.90

[a] Represents a scale with only two items. Cronbach's alpha is not applicable, so these are correlations.

The Cronbach's alpha scores for the scales ranged from a low of 0.63 to a high of 0.90, indicating that they have a sufficiently high degree of inter-item reliability. Cronbach alpha scores greater than 0.70 for existing scales and 0.60 for new scales indicate a high degree of inter-item reliability (Nunnally, 1978; Flynn et al., 1990). The individual items included in each scale are shown in Appendix A. All final scales are formed by computing the mean of the items comprising that scale.

4.1. Technology acceptance factors

The first step in the data analysis is to identify groups using cluster analysis based on four scales adopted from the Technology Acceptance Model (Davis et al., 1989). Perceived ease of use (PERCEIVED EASE) was originally defined as the extent to which the target technology's use or implementation is free from undue effort on the part of the end user. Perceived usefulness (PERCEIVED USEFULNESS) is defined as a potential user's subjective views of the new technology as offering benefits relative to alternative methods of performing the same task. ATTITUDE measures more general feelings regarding the technology. COMFORT refers to their prior experience and overall level of familiarity with computer systems.

4.2. Contextual factors

We developed three scales to measure the reasons companies chose to use the website. GOALS consists of three items measuring the importance of cost, convenience and delivery speed. ADMINISTRATIVE consists of four items assessing the importance of various standard purchasing administrative functions. DELIVERY includes three items relating to order accuracy, delivery reliability and security of the system.

4.3. Outcome factors

4.3.1. Organizational performance

Performance is measured using three scales. INTERNET IMPROVEMENT is a scale developed to assess the degree that the Internet improves individual purchasing transactions such as the time to place an order, completeness of order documentation and the reliability of delivery. The other two scales measure system level performance. For example, COST PERFORMANCE assesses the degree to which the Internet improves general purchasing cost and personnel training costs. Note that this variable is reverse-scaled and the mean shown in Table 4 has already been reversed. PURCHASING EFFICIENCY measures items such as billing accuracy and availability of supplies and materials. The Cronbach's alpha for INTERNET IMPROVEMENT (0.90) exceeds the accepted threshold value, while the inter-item correlations between the two items comprising COST PERFORMANCE and PURCHASING EFFICIENCY are quite high (0.42 and 0.50, respectively), suggesting a high degree of inter-item reliability.

4.3.2. Site performance

We developed a list of 17 items that referred to specific aspects of a particular Internet site. An exploratory factor analysis was run to assess whether these items were uni-dimensional. The factor analysis showed eigen values of 6.07, 1.80 and 1.75 for the first three factors, so we ran a varimax rotation for the three factors. We then retained items that had loadings of 0.40 or higher on a single factor. We labeled the resulting factors as SITE EASE, with items relating to the ease to navigate and load the site; SITE ACCURACY, with items relating to current prices, promotions, in-stock status, etc. and TRANSACTION which includes items on billing and order placement. There were three items that did not load on a specific factor and were, thus, dropped from further analysis. Each of the three scales has a Cronbach's alpha above 0.60.

5. Statistical approach

5.1. Cluster development

To explore the initial research question, we conducted a hierarchical cluster analysis using Ward's method. This analysis technique has been used in previous studies to examine configurations of low contact services (Verma and Young, 2000), approaches to mass customization (Duray et al., 2000) and patterns of investment in firms employing advanced manufacturing technologies (Boyer et al., 1996). For our purposes, although the Technology Acceptance Model has been widely studied, it has not been examined for patterns of users that utilize the Internet as a purchasing tool. In this study, we are attempting to provide insight as to how various groups of respondents view Internet purchasing.

The primary challenge in cluster analysis is determining the appropriate number of clusters to use in the analysis since this choice is primarily subjective. Previous researchers have suggested various heuristics to choose the number of clusters. Our subjective assessment suggested that a six cluster solution provided a good blend of parsimony (fewer clusters) and accuracy (more clusters retains more data, but is harder to describe intuitively). We also employed the suggestion of Milligan and Cooper (1985) that the number of clusters be chosen such that a further reduction in clusters results in a substantial drop in R^2. The semipartial R^2 for seven clusters is 0.004, while the semipartial R^2 for six clusters is 0.02 and the semipartial R^2 for five clusters is 0.06. This indicates that decreasing from seven to six clusters results in a very small decrease in explanatory power (0.004), while a decrease from six to five clusters reduces explanatory power more substantially (0.02). Hence, we chose the six cluster solution shown in Table 5 with a R^2 of 0.624.

After choosing the number of clusters to use in the study, a one-way ANOVA was conducted to assess differences between the various groups. A Scheffe's pairwise comparison procedure is used to test the statistical significance between the individual differences in the clusters (Miller and Roth, 1994; Boyer et al., 1996). This test is the most conservative pairwise comparison test and, thus, the least likely to result in alpha errors.

6. Results

6.1. Internet purchasing clusters

Table 5 presents the results of the cluster analysis performed on the four individual preferences listed in Fig. 1. Each of the groups presented in the six-cluster model represents a viewpoint regarding the adoption of Internet purchasing. One-way ANOVA results for each of the four clustering variables were significant at the $P < 0.01$ level. Based on a Scheffe's pairwise comparison, all of the clusters varied from at least one of the other groups. Together, these results give support to our initial research question, which states that individuals who use Internet purchasing can be grouped based on end users' viewpoints and preferences regarding the Internet.

Approximately 14.1% of the individuals were classified into cluster 1 (Ambivalent Users). These appear to be the individuals who are indifferent towards using the Internet. They are neither overly in favor of the Internet for purchasing or are strictly opposed to using the Internet.

Approximately 26.7% of the individuals were classified into cluster 2 (Technology Acceptors). These are the individuals who find using the Internet to be easy and are very comfortable with using technology. However, in general they find the Internet only somewhat useful (PERCEIVED USEFULNESS has only the 4th highest ranking) when it comes to ordering goods and supplies. They accept using the technology, but perhaps do not find it to be the most useful purchasing medium.

Table 5
Internet purchasing clusters

	(1) Ambivalent Users ($n = 56$)	(2) Technology Acceptors ($n = 106$)	(3) Enthusiastic Users ($n = 62$)	(4) Slow Adopters ($n = 32$)	(5) Pro-Technology ($n = 110$)	(6) Reluctant Users ($n = 29$)	F-statistic
PERCEIVED EASE	(2, 3, 5, 6)	(1, 4, 5, 6)	(1, 4, 5, 6)	(2, 3, 5, 6)	(1, 2, 3, 4, 6)	(1, 2, 3, 4, 5)	
Mean	4.35	5.55	5.78	4.02	6.53	2.81	166.31
Standard deviation	0.55	0.89	0.64	1.04	0.50	1.11	$P < 0.01$
Rank	**4**	**3**	**2**	**5**	**1**	**6**	
PERCEIVED USEFULNESS	(3, 4, 5, 6)	(3, 4, 5, 6)	(1, 2, 4, 5, 6)	(1, 2, 3, 5)	(1, 2, 3, 4, 6)	(1, 2, 3, 5)	
Mean	4.79	4.56	5.40	2.76	6.44	2.38	183.87
Standard deviation	0.58	1.04	0.80	0.90	0.52	1.13	$P < 0.01$
Rank	**3**	**4**	**2**	**5**	**1**	**6**	
COMFORT	(2, 5, 6)	(1, 2, 4, 6)	(2, 5, 6)	(2, 5, 6)	(1, 3, 4, 6)	(1, 2, 4, 5)	
Mean	4.15	6.01	4.16	3.70	5.88	4.75	86.13
Standard deviation	0.94	0.61	0.82	1.05	0.89	1.25	$P < 0.01$
Rank	**5**	**1**	**4**	**6**	**2**	**3**	
ATTITUDE	(3, 4, 5, 6)	(3, 4, 5, 6)	(1, 2, 4, 5, 6)	(1, 2, 3, 5, 6)	(1, 2, 3, 4, 6)	(1, 2, 3, 4, 5)	
Mean	5.18	5.25	5.98	4.41	6.53	2.47	133.03
Standard deviation	0.85	1.01	0.65	0.80	0.46	1.27	$P < 0.01$
Rank	**4**	**3**	**2**	**5**	**1**	**6**	

Cluster analysis performed on $n = 395$ complete responses. Numbers in parentheses indicate the group numbers from which his group was significant according to Scheffe's pairwise comparison procedure. F-statistics and associated P-values are derived from one-way ANOVAs.

3-11

Approximately 16.3% of the individuals were classified into cluster 3 (Enthusiastic Users). These are the individuals who find Internet purchasing very useful and have a positive attitude towards the Internet. However, members of this group are just relatively comfortable (COMFORT is ranked 4th) with using the technology. They perhaps still feel more comfortable using more traditional methods, but are eager to learn and use the Internet for purchasing.

Approximately 8.0% of the individuals were classified into cluster 4 (Slow Adopters). This group is characterized by people who are generally unenthusiastic users of the Internet. This groups ranks either last or second to last on each of the scales shown in Table 5.

Approximately 27.6% of the individuals were classified into cluster 5 (Pro-Technology). In general, this group appears to have the most positive attitude towards the use of Internet technology. The Pro-Technology group has the highest ranking or second highest ranking for all four constructs shown in Table 5. These individuals on average are very comfortable with using the Internet and have a very positive attitude about the impact of the Internet on their organization. The findings may indicate that they are positive towards the use of the Internet and possible new technologies in the workforce. It is interesting to note that many of the companies that started ventures into web technology believed that the majority of people were in this group. This may help to explain why so many dot-com organizations are going out of business. The wide range of differences in the cluster analysis clearly indicates that Internet retailers must be more cognizant of the different needs, Goals and desires of customer groups.

Finally, approximately 7.3% of the individuals were classified into cluster 6 (Reluctant Users). Members of this cluster generally do not want to accept new technology. While many of the individuals in this group have to use the Internet for their job functions they would rather use more traditional methods. They have a negative attitude towards using Internet purchasing and do not believe that it adds value. It is interesting to note here that these individuals also find it difficult to understand and use the Internet. This frustration helps to explain their poor attitude and why they do not find the Internet useful.

6.2. Impact of contextual factors—Hypothesis 1

Based on Hypothesis 1a, we expected that an individual's background including education, tenure in the workforce and annual training received would have an impact on their beliefs about using the Internet as a purchasing medium. This hypothesis is tested using one-way ANOVA followed by Scheffe's pairwise comparisons to indicate which mean cluster scores varied from each other for the contextual variables. Table 6 indicates somewhat surprising results. The one-way ANOVAs were not significant for education level, tenure in workforce or annual training received. Interestingly, the lowest nominal score for education is for the Pro-Technology group (average is 2.09, where 1: HS, 2: 2 years degree, 3: 4 years degree, 4: graduate degree). Note that we computed an average value for the categorical values for Education level (as shown in Table 6) for ease of presentation only—higher values indicate higher education levels. The more statistically appropriate test is a non-parametric chi-square test, which is also non-significant. While the follow-up Scheffe's pairwise comparison tests indicated that none of the individual clusters were different from one another, this is attributable to the more conservative nature of a Scheffe test relative to the one-way ANOVA. It is somewhat contradictory that the two groups with the highest training levels (Slow Adopters and Pro-Technology) represent essentially opposite approaches to Internet purchasing. Perhaps this is because the Slow Adopters have been forced to increase levels of training in order to support Internet usage—in short the training overcomes their natural resistance to the technology. In comparison, the Pro-Technology group is most likely comfortable with the technology, partly due to their increased training. These results raise more questions than they answer, but do offer an interesting area for further examination. It is certainly in the best interests of Internet retailers to better understand the relationship between training and Internet purchasing adoption. Overall, these results do not provide support for Hypothesis 1a. They indicate that an individual's background may not have much impact on an end user's viewpoint or preferences towards Internet purchasing.

We believe this helps to dispel popular belief that the real supporters of the Internet are those individuals that are well educated. Our results give support

Table 6

Contextual factors by Internet purchasing groups

	(1) Ambivalent Users ($n = 56$)	(2) Technology Acceptors ($n = 106$)	(3) Enthusiastic Users ($n = 62$)	(4) Slow Adopters ($n = 32$)	(5) Pro-Technology ($n = 110$)	(6) Reluctant Users ($n = 29$)	F-statistic
Hypothesis 1a: individual differences							
EDUCATION							
Mean	2.20	2.31	2.26	2.23	2.07	2.14	0.61
Standard deviation	1.07	1.04	1.14	1.04	1.04	1.03	$P = 0.70$
Rank	4	1	2	3	6	4	
ANNUAL TRAINING							
Mean	5.39	5.42	5.64	13.65	12.48	6.28	1.52
Standard deviation	9.56	10.80	10.77	53.45	30.91	9.85	$P < 0.18$
Rank	6	4	5	1	2	3	
TENURE IN WORKFORCE							
Mean	19.85	17.09	14.54	20.28	15.02	16.51	
Standard deviation	11.88	10.41	10.90	10.47	10.82	9.87	2.70
Rank	2	3	6	1	5	4	$P = 0.25$
Hypothesis 1b: strategic goals							
GOALS		(5)	(5)	(5)	(2, 3, 4)		
Mean	5.66	5.53	5.59	5.27	6.23	5.53	6.97
Standard deviation	1.15	1.21	1.00	1.39	0.79	1.20	$P < 0.01$
Rank	2	4	3	6	1	4	
ADMINISTRATIVE	(5)	(5)	(4, 6)	(3, 5)	(1, 2, 4, 6)	(3, 5)	
Mean	4.88	4.69	5.28	4.14	5.85	4.12	18.12
Standard deviation	1.17	1.38	1.18	1.50	1.09	1.25	$P < 0.01$
Rank	3	4	2	5	1	5	
DELIVERY		(5)		(5)	(2, 4)		
Mean	5.48	5.53	5.75	5.24	6.14	5.51	4.49
Standard deviation	1.17	1.41	1.30	1.51	0.96	1.43	$P < 0.01$
Rank	3	3	2	6	1	3	

Numbers in parentheses indicate the group numbers from which his group was significant according to Scheffe's pairwise comparison procedure. F-statistics and associated P-values are derived from one-way ANOVAs.

3-13

to the notion that regardless of an individual's education level, there may be either a positive or negative association with Internet purchasing. The data also suggest that tenure in the workforce has very little impact on an individual's opinion of the Internet as a purchasing medium. This finding parallels that of Agarwal and Prasad (1999), who had similar results in situations when new information technologies were being adopted by organizations. Finally, the data suggest that hours of computer training received each year also have little impact on how an individual perceives Internet purchasing. Two possible reasons for this outcome include: (1) the question did not directly ask how many hours training hours a person receives on the Internet or more specifically Internet purchasing, or (2) amount of training on computers may have little bearing on how individuals perceive Internet purchasing.

Table 6 also provides data on the strategic goals of our six Internet purchasing groups. There are clear differences in strategy, with more proactive companies (Pro-Technology and Enthusiastic Users) placing the greatest emphasis on all three strategy constructs. The most interesting group is the Reluctant Users, who place a high degree of importance on DELIVERY relative to ADMINISTRATIVE or GOALS. The relatively heavy emphasis on delivery factors such as SITE ACCURACY and DELIVERY suggests that this group has fairly clear goals in adopting Internet purchasing.

6.3. *Outcomes of Internet purchasing—Hypothesis 2*

Hypothesis 2 was tested in the same manner as Hypothesis 1—one-way ANOVA and Scheffe's pairwise comparison results are presented in Table 7. The F-tests were significantly different at the $P < 0.01$ level for each of the purchasing performance and site performance measures.

The first conclusion that can be drawn from Table 7 is that the groups with higher acceptance of Internet purchasing generally have the highest purchasing performance. The Pro-Technology and Enthusiastic User groups have the highest and second highest rankings for all three performance measures. The lowest technology groups—Slow Adopters and Reluctant Users—have the lowest rankings for the three performance measures. In short, the measures of purchasing performance are highest for groups with high levels of Internet purchasing.

The second set of outcome factors included in Table 7 relates to site performance, or the relative reliability, ease and site accuracy of Office Depot's website. In general, the ranking of measures follows a pattern similar to that seen for purchasing performance, with a couple notable exceptions. Companies with a higher level of purchasing acceptance generally rank Office Depot's website higher, and vice versa. However, it is interesting to note that the Slow Adopters have the second highest ranking for SITE ACCURACY. Apparently these consumers believe the site to be very accurate with respect to inventory and information on products, but are less impressed with the ease with which the site can be utilized (TRANSACTION and SITE EASE). In contrast, the Enthusiastic Users believe the site to be fairly easy to use, but rate it lowest in terms of SITE ACCURACY.

The final set of outcome measures is shown in Table 8 and Fig. 2. The data indicate that the more technologically proactive groups place more Internet orders and have a higher percentage of Internet orders (Internet orders placed with Office Depot divided by all orders placed). The differences in the absolute number of Internet orders placed are nominally large, but not statistically significant. However, the ANOVA test for Percentage of Internet Orders is significant ($P < 0.01$). The Pro-Technology, Enthusiastic Users and Technology Acceptor groups all have greater than 50% of their orders placed over the Internet. This suggests that the Internet has become these companies' preferred ordering method. This is an important finding since Office Depot estimates that Internet orders are generally less expensive to process than phone-in or fax orders. This savings has been estimated at 1% of sales. Clearly, Office Depot would like to find ways to increase the percentage of orders placed over the Internet for all companies. Yet, it is interesting that even the technology "laggards" place a substantial portion of orders on the Internet.

The story when it comes to the total dollars spent with Office Depot is a complete opposite, as shown in Fig. 2 and Table 8. It is the companies with the smallest percentage of Internet orders that spend the most money (the Slow Adopters and Reluctant Users). This suggests that while the retail company (Office Depot) wants to encourage companies to purchase over the web because of its greater efficiencies, they must be careful to support and not risk alienating companies

Table 7
Outcome factors by Internet purchasing groups

	(1) Ambivalent Users ($n = 56$)	(2) Technology Acceptors ($n = 106$)	(3) Enthusiastic Users ($n = 62$)	(4) Slow Adopters ($n = 32$)	(5) Pro-Technology ($n = 110$)	(6) Reluctant Users ($n = 29$)	F-statistic
Hypothesis 2a: purchasing performance							
INTERNET IMPROVEMENT	(5, 6)	(3, 5, 6)	(2, 4, 6)	(3, 5)	(1, 2, 4, 6)	(1, 2, 3, 5)	
Mean	4.35	4.16	5.00	3.56	5.53	3.04	26.12
Standard deviation	1.54	1.51	1.24	1.36	1.08	1.38	$P < 0.01$
Rank	**3**	**4**	**2**	**5**	**1**	**6**	
COST REDUCTION					(6)	(5)	
Mean	4.02	4.10	4.29	3.89	4.36	3.73	4.52
Standard deviation	0.55	0.63	0.89	0.35	1.03	0.88	$P < 0.01$
Rank	**4**	**3**	**2**	**5**	**1**	**6**	
PURCHASING EFFICIENCY	(5)	(5)		(5)	(1, 2, 4, 6)	(5)	
Mean	4.38	4.39	4.54	4.06	4.97	4.09	10.47
Standard deviation	0.70	0.76	0.91	0.59	1.00	0.67	$P < 0.01$
Rank	**3**	**3**	**2**	**5**	**1**	**5**	
Hypothesis 2b: site performance							
TRANSACTION	(5, 6)	(5, 6)	(6)	(5)	(1, 2, 4, 6)	(1, 2, 3, 5)	
Mean	5.16	5.18	5.61	4.80	5.94	4.03	15.78
Standard deviation	1.02	1.19	1.32	1.38	1.05	1.20	$P < 0.01$
Rank	**3**	**3**	**2**	**5**	**1**	**6**	
SITE ACCURACY	(5)	(5)	(6)	(5)	(1, 2, 4, 6)	(3, 5)	
Mean	4.96	5.14	5.50	5.17	6.00	4.68	15.38
Standard deviation	0.86	1.04	1.11	1.08	0.83	1.22	$P < 0.01$
Rank	**4**	**2**	**6**	**2**	**1**	**5**	
SITE EASE	(3, 5, 6)	(3, 5, 6)	(1, 2, 4, 6)	(3, 5, 6)	(1, 2, 4, 6)	(1, 2, 3, 4, 5)	
Mean	4.85	5.18	5.93	4.81	6.09	3.70	39.22
Standard deviation	0.91	1.13	0.79	0.88	0.85	1.35	$P < 0.01$
Rank	**4**	**3**	**2**	**4**	**1**	**6**	

Numbers in parentheses indicate the group numbers from which his group was significant according to Scheffe's pairwise comparison procedure. F-statistics and associated P-values are derived from one-way ANOVAs.

3-15

Table 8
Purchasing and percent Internet orders by Internet purchasing groups

	(1) Ambivalent Users ($n = 56$)	(2) Technology Acceptors ($n = 106$)	(3) Enthusiastic Users ($n = 62$)	(4) Slow Adopters ($n = 32$)	(5) Pro-Technology ($n = 110$)	(6) Reluctant Users ($n = 29$)	*F*-statistic
INTERNET ORDERS							
Mean	3.41	3.80	4.79	2.41	4.07	1.90	
Standard deviation	4.99	8.38	5.97	2.55	5.46	0.90	1.31
Rank	**4**	**2**	**1**	**5**	**3**	**6**	$P < 0.26$
PERCENTAGE INTERNET ORDERS		(6)	(4, 6)	(3, 5)	(4, 6)	(2, 3, 5)	
Mean	0.41	0.49	0.61	0.35	0.60	0.24	7.19
Standard deviation	0.37	0.37	0.38	0.36	0.38	0.24	$P < 0.01$
Rank	**4**	**3**	**1**	**5**	**1**	**6**	
PURCHASE		(4)		(2)			
Mean (US$)	2315	1527	1943	3389	1732	2358	2.82
Standard deviation (US$)	2905	1888	2907	4899	2153	2624	$P < 0.05$
Rank	**3**	**6**	**4**	**1**	**5**	**2**	

Numbers in parentheses indicate the group numbers from which his group was significant according to Scheffe's pairwise comparison procedure. *F*-statistics and associated *P*-values are derived from one-way ANOVAs.

3-16

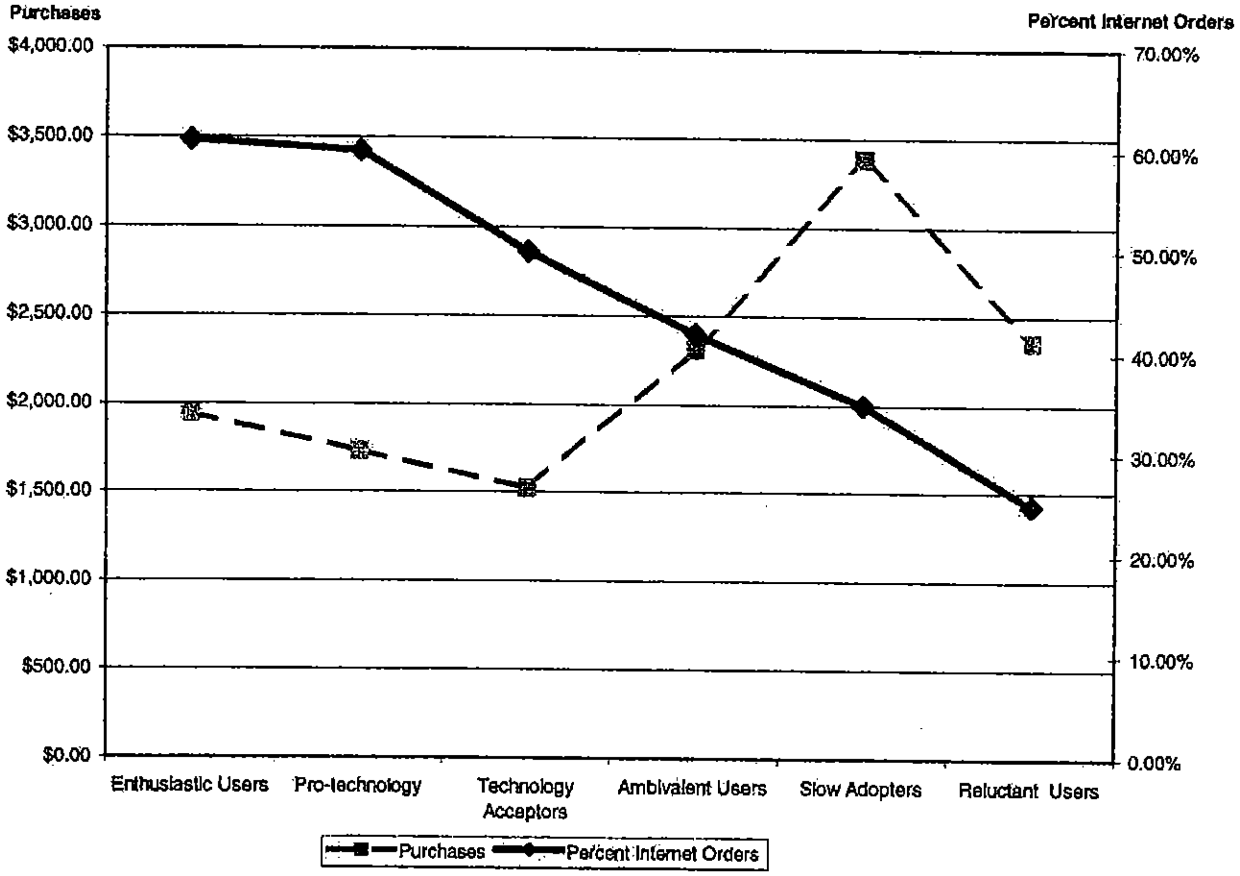


Fig. 2. Purchases and percent internet orders by internet purchasing group.

that do a substantial amount of business via more traditional methods. With an average purchase volume of US$ 3389 during the previous 12 months, the Slow Adopters represent a valuable market segment that needs to be carefully nurtured. What our data does not provide insight on is the degree to which the Internet increases total purchases per company. Many retailers believe that offering more channels for consumers to purchase through generally increases the average volume per company. In order to address this question, we would need to gather data on companies that have not purchased over the Internet.

7. Managerial insights

The Internet has been the one of the more recent tools to revolutionize the way organizations conducts business. As with the invention of any revolutionary technology, many individuals were able to capitalize on the expanding market and get rich quick. However, simply the concept of streamlined transactions facilitated by the Internet is no longer enough. Organizations are increasingly being forced to offer consumers value with their on-line services. A critical component in the future development of effective websites is the need for companies to pay attention to a wide variety of consumer desires while simultaneously pursuing their own internal goals.

Managers of companies involved with Internet initiatives on the sales side can learn some powerful lessons from this data. First, the notion that the Internet is only for the well educated and new generation of workers is simply not true. Our results indicate that education level, tenure in the workforce or annual training received, might have either a positive or negative impact on an individuals' view of Internet purchasing. This wide range of responses suggests that the audience for a website focusing on the sale of commodity type products is very broad. Potential poor results may result if website developers are too focused on marketing to leading edge, highly educated consumers in their approach to creating the on-line system.

Second, organizations that are creating on-line retailing venues must consider that there are groups of individuals that have very different viewpoints and preferences. In the ideal world, when a company offers a new service they expect to get customers who want and desire to use their product. The reality is that groups of individuals will use the service for widely different reasons. Evidence in our study suggests that three general attitudes towards the use of Internet purchasing exist: (1) those who are very positive towards the idea; (2) those who are indifferent towards the idea; and (3) those who are uncomfortable with or dislike yet may have to use the technology for their jobs. Fortunately, only 15% of our sample contained those who indicated they disliked the Internet. Generally speaking, no matter what an organization does they will not be able to cater to these customers' needs. The remaining 85% of our sample fit roughly into the other two categories. Among the outcomes for this group were improvements in purchasing efficiency (INTERNET IMPROVEMENT—the degree to which Internet purchasing increases the accuracy, speed of billing and order placement) and a better perception of the particular website being utilized (SITE EASE—the reliability, speed and ease of navigation for a site).

The so-called middle market respondents (individuals who are indifferent towards the Internet and simply view it as part of their everyday job function) provide some interesting insights. We classify this group as the meat and potatoes group—to be successful these are the individuals that a company *must* pay attention to. In contrast, the proactive, technology vanguard (the Pro-Technology and Enthusiastic User groups) would appear to be easy to sell on new technologies. Thus, it is compelling that the indifferent group of Internet purchasers found benefits in purchasing over the Internet. Purchasing performance typically was better for the Internet (both COST PERFORMANCE and PURCHASING EFFICIENCY had high scores compared to traditional methods). The message here is quite clear, purchasing companies are realizing value through the Internet. However, sites that are hard to navigate, unreliable or difficult to use, can turn customers away easily. In short, the data suggest that it is critical not to give customers unnecessary headaches. The overriding theme is that customers want a website that is able to accurately take their order and not cause them to duplicate their work.

We believe that this study has demonstrated that considering end user's viewpoints and preferences will have a significant impact toward an organization being successful in on-line purchasing. Future research needs to explore the differences between commodity-based products and more specialized products. There is also a need to examine non-adopters—those companies who have not even tried to purchase over the Internet. In addition, it is important to examine both market leaders and market laggards in terms of Internet retail activities. Our results are based on a survey of customers of an acknowledged leader, it is reasonable to expect that results would be different for companies without Office Depot's history or success. Future research should examine a wide variety of Internet retailers to capture lessons from several market segments. Our results show that purchasing companies are finding significant benefits, thus, Internet purchasing is here to stay, but there is always room for improvement.

8. Limitations and directions for future research

This study provides an initial examination of how the human element affects the utilization of Internet purchasing. Like any exploratory study, several limitations exist that restrict the ability to generalize the results. As the research evolves in this area, researchers will be able to refine both their research models and instruments. Future research needs to continue the development of sound theoretical models and instruments.

Based on the data the researchers feel that the study is generalizable to the population of small organizations who order products on-line. However, this is an analysis of one organization's customers and it should be recognized that some of the results might be specific to Office Depot customers. Future research should be conducted to verify that these results are applicable across other industry groups, and various sized organizations.

Due to the nature of the study a positive response bias is possible. The original TAM model was specifically designed to measure a pre-adoption of technology scenario. In a similar fashion to Lederer et al. (2000) and Jiang et al. (2000) we applied the TAM factors in a post adoption case. However, the results indicate that distinctions still exist among the respondents. Future research needs to focus on how the factors used in the study change over time. Specifically examining if usage and adoption level affects the rate at which the factors change over time.

One final limitation is that we did not control for who took the survey. One reason for not controlling for this variable is that we focused specifically on smaller organizations. In many instances people who work for smaller organizations have many job functions and activities and the line between senior manager and staff becomes blurred. In addition, we felt that it was imperative that the people who were using the technology were the individuals responsible for filling out the questionnaire. As the research extends to other organization structures differences in responses based on job function should be examined. Despite these limiting factors, the present research is the first step in the examining the factors that influence the utilization of Internet purchasing. Future studies that consider these limitations should be undertaken before other generalizations are made.

Acknowledgements

This grant was partially funded by APICS Research and Educational Foundation, E-Business: Managing the Integrated Business Enterprise track. The grant was titled, "Internet Supply Chain Management, An Exploratory Study of Buyer–Supplier Benefits." We appreciate the many helpful suggestions provided by three reviewers and the associate editor.

Appendix A. Internet purchasing acceptance, context and outcome scales

A.1. Internet Purchasing Acceptance Factors

The following questions are rated on a seven-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree).

A.1.1. PERCEIVED USEFULNESS

- Using Internet purchasing enables me to accomplish tasks such as order placement, order estimating and order tracking more quickly.
- Using Internet purchasing improves my job performance.
- Using Internet purchasing gives me greater control over my work.
- Using Internet purchasing improves the quality of the work I do.
- Using Internet purchasing improves my productivity.
- Using Internet purchasing enhances my effectiveness on the job.
- Using Internet purchasing makes it easier to do my job.
- Overall, I find Internet purchasing technology useful in my job.

A.1.2. PERCEIVED EASE

- It is easy for me to remember how to perform tasks using Internet purchasing.
- It is easy to get Internet purchasing to do what I want it to do.
- My interaction with Internet purchasing is clear and understandable.
- Overall, I believe that Internet purchasing is easy to use.

A.1.3. ATTITUDE

- I like using Internet purchasing.
- Internet purchasing is fun to use.
- I dislike using Internet purchasing (R).
- Internet purchasing provides an attractive working environment.

A.1.4. COMFORT

- I am knowledgeable about personal computer usage.
- I am comfortable and experienced with the Internet.
- I am proficient at fixing glitches when working on the computer.
- I am good at resolving problems with computers.

A.2. Contextual Factors—Strategy

How important were the following factors in the decision to use Office Depot's on-line ordering system?

(Likert scale ranging with 1 (not important), 4 (somewhat important) and 7 (very important).

A.2.1. GOALS

- Cost.
- Convenience.
- Delivery speed.

A.2.2. ADMINISTRATIVE

- Ability to track inventory.
- Reduces paperwork.
- Faster access to information.
- Flexibility in order size.

A.2.3. DELIVERY

- Order site accuracy.
- Delivery of system.
- Security of system.

A.3. Outcome Factors

A.3.1. ORGANIZATIONAL PERFORMANCE

A.3.1.1. Internet improvement. Please rate the degree of improvement when the following tasks/events are conducted using the Internet for purchasing products through Office Depot in comparison to traditional methods (order by phone, fax or mail): 1 (small improvement), 4 (moderate improvement), 7 (large improvement).

- The time to place an order.
- The delivery time from when an order is placed to receipt of all items.
- The thoroughness of order documentation.
- The ease of interpretation for documentation.
- The reliability of delivery.

Please rate the impact of the Internet ordering system on the company. Each of the following items was rated on a seven-point Likert scale with 1 (greatly decreased), 4 (remained the same) and 7 (greatly increased).

A.3.1.2. Cost reduction.
- Cost of activities associated with purchasing (R).
- Cost of training new personnel associated with ordering systems (R).

A.3.1.3. Purchasing efficiency.
- Site accuracy of your billing.
- Availability of supplies and materials.

A.3.2. SITE PERFORMANCE

Please rate the following aspects of Office Depot's Internet site: from 1 (strongly disagree) to 7 (strongly agree).

A.3.2.1. Site ease.
- I can get on the site when I want to.
- The site loads quickly.
- The site is easy to navigate.
- The site has a logical sequence of pages.
- Office Depot website is easy to search.

A.3.2.2. Site accuracy.
- Contents on the web page are current with respect to *Price*.
- Contents on the web page are current with respect to *New Items*.
- Contents on the web page are current with respect to *In-Stock Items*.
- Contents on the web page are current with respect to *Promotions*.
- Office Depot has my products in stock when I place an order.

A.3.2.3. Transaction.
- I experience difficulties *placing an order* when using the on-line ordering system (R).
- I experience *web page navigation* (i.e. page would not upload or server time was expired) problems when using the on-line ordering system (R).
- I experience *billing* problems when using the on-line ordering system (R).

Items that did not load on any scale:

- Office Depot sends order confirmation in a timely fashion.
- The company delivers the items when they promised delivery.
- Office Depot provides good technical support.

Note: (R) signifies a reverse coded item.

References

Agarwal, R., Prasad, J., 1999. Are individual differences germane to the acceptance of new information technologies? Decision Sciences 30, 361–391.

Boston Consulting Group, 2000. Five Key Lessons for Online Retailers.

Boyer, K.K., Ward, P.T., Leong, G., 1996. Approaches to the factory of the future: an empirical taxonomy. Journal of Operations Management 4, 291–313.

Boyer, K.K., Leong, G., Ward, P.T., Krajewski, L., 1997. Unlocking the potential of advanced manufacturing technologies. Journal of Operations Management 15, 331–347.

Butler, T.W., Leong, G.K., Everett, L.N., 1996. The operations management role in hospital strategic planning. Journal of Operations Management 14, 137–156.

Carter, J., Frendall, L., 1990. The dollars and sense of electronic data interchange. Production and Inventory Management 31, 22–25.

Davis, F., Bagozzi, R., Warshaw, P., 1989. User acceptance of computer technology: a comparison of two theoretical models. Management Science 35, 982–1003.

Duray, R., Ward, P., Milligan, G., Berry, W., 2000. Approaches to mass customization: configurations and empirical validation. Journal of Operations Management 18, 605–626.

Fishbein, M., Ajzen, I., 1975. Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research. Addison-Wesley, Reading, MA.

Flynn, B.B., Sakakibara, S., Schroeder, R.G., Bates, K.A., Flynn, E.J., 1990. Empirical research methods in operations management. Journal of Operations Management 9, 250–284.

Garvin, D.A., 1993. Manufacturing strategic planning. California Management Review 35, 85–106.

Gulati, R., Garino, J., 2000. Get the right mix of bricks & clicks. Harvard Business Review 78, 107–114.

Haddad, C., 2000. Office Depot's E-Diva. Business Week. 6 August 2001.

Harrison, A.W., Rainer, R.K., 1992. The influence of individual differences on skill in end-user computing. Journal of Management Information Systems 9, 93–111.

Hayes, R.H., Wheelwright, S.C., 1979. Link manufacturing process and product life cycles. Harvard Business Review 57, 133–140.

Igbaria, M., Zinatelli, N., 1997. Personal computing acceptance factors in small firms: a structural equation model. MIS Quarterly 21, 279–302.

Jiang, J., Hsu, M., Klien, G., Lin, B., 2000. E-commerce user behavior model: an empirical study. Human Systems Management 19, 265–276.

Johnson, P., Klassen, R., Leenders, M., Fearon, H., 2002. Determinants of purchasing team usage in the supply chain. Journal of Operations Management 20, 77–89.

Kathuria, R., 2000. Competitive priorities and managerial performance: a taxonomy of small manufacturers. Journal of Operations Management 18, 627–642.

Kellogg, D.L., Nie, W., 1995. A framework for strategic service management. Journal of Operations Management 13, 323–337.

Krause, D., Pagell, M., Curkovic, S., 2001. Toward a measure of priorities for purchasing. Journal of Operations Management 19, 497–512.

Lederer, A., Maupin, D., Sena, M., Zhuang, Y., 2000. The Technology Acceptance Model and the world wide web. Decision Support Systems 29, 269–282.

Lucas, H.C., Spitler, V.K., 1999. Technology use and performance: a field study of broker workstations. Decision Sciences 30, 291–311.

Massetti, B., Zmud, R.W., 1996. Measuring the extent of EDI usage in complex organization: strategies and illustrative examples. MIS Quarterly 20, 331–345.

Milligan, G.W., Cooper, M.C., 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika 2, 159–179.

Miller, J., Roth, A., 1994. A taxonomy of manufacturing strategies. Management Science 40, 285–304.

Moore, L., Online Sales Rise, But Many Buyers Wary. crmdaily.com, 23 April 2002.

Mukhopadhyay, T., Kekre, S., Kalathur, S., 1995. Business value of information technology: a study of electronic data interchange. MIS Quarterly 19, 137–156.

Nunnally, J., 1978. Psychometric Theory. McGraw-Hill, New York.

Office Depot 2000. Annual Report, 2001.

Reichheld, F., Schefter, P., 2000. E-loyalty: your secret weapon on the web. Harvard Business Review 78, 105–114.

Roth, A.V., Van der Velde, M., 1996. Operations as marketing: a competitive service strategy. Journal of Operations Management 14, 137–156.

Schmenner, R.W., Swink, M.L., 1998. On theory in operations management. Journal of Operations Management 17, 97–113.

Skinner, W., 1969. Manufacturing: missing link in corporate strategy. Harvard Business Review 47, 136–145.

Straub, D., Limayem, M., Karahanna-Evaristo, E., 1995. Measuring system usage: implications for IS theory testing. Management Science 41, 1328–1342.

Troy, M., 1999. Office supplies: clicks-and-mortar dominates—for now. Discount Store News 38, 57.

Venkatesh, V., 2000. Determinants of perceived ease of use: integrating control, intrinsic motivation, and emotion into the Technology Acceptance Model. Information Systems Research 11, 342–362.

Venkatesh, V., Davis, F., 2000. A theoretical extension of the Technology Acceptance Model: four longitudinal field studies. Management Science 46, 186–204.

Venkatesh, V., Morris, M., 2000. Why do not men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behavior. MIS Quarterly 24, 115–139.

Verma, R., Young, S.T., 2000. Configurations of low-contact services. Journal of Operations Management 18, 643–661.

Ward, P.T., McCreery, J.K., Ritzman, L.P., Sharman, D., 1998. Competitive priorities in operations management. Decision Sciences 29, 1035–1046.

Warner, M., Roth, D., Schonfeld, E., Gunther, M., 1999. 10 companies that get it. Fortune 140, 115–117.