# Comparison for Three Internet Search Tools

## (Yahoo, AltaVista, Lycos)

Ming  Hou

April 1998

# ABSTRACT

The objective of this project is to evaluate the usability and effectiveness of information retrieved by three Internet search tools: Yahoo, AltaVista, and Lycos. An 8x2x3 factorial experiment was conducted to compare the precision on relevance for the first ten retrieved links, the sensitivity d' and decision criterion bias beta for finding useful information, and subjective ranking for response time and interface output format. Three-way and two-way analysis of variance (ANOVA) and signal detection analysis (SDA) were used to analyse subjects' responses. The main findings were that there was no significant difference in precision for the retrieved references, and all these web search engines were sensitive to the query question and had significant effects on the sensitivity and the decision criterion. Among three Internet search engines, Yahoo was found to be most suited for current technology searching and obtained the highest score in all feature tests, followed by Lycos and then AltaVista.

# INTRODUCTION

Today, searching on the World Wide Web (WWW) seems to become part of our routine life. The web has even become a necessary tool for collecting information. Undoubtedly, the web provides people the convenience, but this sea of information made any query in this huge information reservoir extremely difficult, even people thought it as a chaotic repository (Lynch 1997). In order to solve this problem, more than two dozen commercial companies and academic institutions have developed search aids (search tools) on the Internet so far, such as Yahoo, Infoseek, AltaVista, and Lycos etc. Obviously, you cannot use all of them at the same time. Facing so many search tools, people can get confused quite easily. Which one is the best ? Which one should I use? The present study is trying to answer these questions by comparing three search tools: Yahoo, AltaVista, and Lycos.

However, to answer these questions, the first thing should be clarified is what you want to search and which feature you prefer for searching: the amount of information or the searching efficiency? The second thing should be understood is how a search tool works, just like an operator should be familiar with the machine used by himself/herself. Whatever your preference is when you do searching, clarifying the work mechanism and thus familiarising yourself with search tools are essential.

## Classification of Search Tools

To answer how a search tool works, it is necessary to know the classification of current search tools. Demoss (1996) divided the search aids to four categories:

Search Engines, Directories, Software Search and All-in-one-search. According to how a search aid works, other researchers classified these services into two basic types: Directory and Search Engine (Liu 1996, Richard 1997). which are briefly introduced as below.

**Directory** A directory can be regarded as a manual catalogue of sites on the internet. This means that people actually create categories and assign sites to a place within a structured index. One of typical directory is Yahoo which screens all relevant information by its staff and assigns these information to its relevant address. Yahoo also orders sites so that the most relevant or comprehensive in each category appear first on the list. This search feature can help people quickly find targeted information on more general topics.

**Search Engine** Actually, the main difference between directories and search engines is that a directory is built by people, but the search engine's database is created by software known as spiders or robots, this is why some people call search engine as robot-driven search engine or robot Wanderer, or Spider, Harvest etc. (Liu 1996, Westera 1996). These spiders attempt to crawl through the web, collect and index resources in a more automatic fashion, and put these information into a database by using their own specific algorithm, thus it does not need extensive human intervention. Searching, instead of browsing, is the main feature of this type of tools. The other component is the query module. Users search the index through a predefined query module, an interface specific to each engine. Currently, AltaVista, Infoseek, Excite, and Lycos are all popular search engines. If you search for the keyword **University of Toronto** using one of these tools, it will return a list of sites that have something to do with "University of Toronto"--and that is where another difference between directories and search engines becomes apparent.

The advantage of search engines is that they are very nearly comprehensive, often including thousands of sites in the results listed. Certainly, it is also useful when you are searching for a specific topic that may not be found in a directory. The disadvantage is that you often have to weed through piles of irrelevant sites to find what you are looking for. Although search engines attempt to list sites in order of relevance, this relevance is determined by a mathematical formula that is not only far from perfect, but also the main reason why there is so many differences among these search engines.

This study is designed to determine which of several interactive search tools----Yahoo, AltaVista, and Lycos, is best in terms of usability and the quality of retrieved web site for various query questions. The evaluation of these search tools is based on empirical user judgement of the ease of use, the relevance of the hits returned, and the quality of the hits. The data of hit numbers and the subjective ranks for the ease of use were collected and the quantitative analysis was also conducted in this experiment by using analysis of variance (ANOVA) and signal detection analysis (SDA) respectively.

The hypothesis is that there will not be much more difference for the usability of these three search engines. It is also expected that there will be some difference between expert and novice users for the experiment and the query suite will not have different effect on each search engine.

## RELATED STUDIES

Since search engines came out only in 1994 (Chu and Rosenthal 1996), there was not much prior comparison associated with the performance of these search tools. The comparisons available were descriptive reviews that were highly depending on the individual experience, thus the results are variable and some of the reported findings do not appear to agree with one another. These evaluations were performed according to some of main factors that determine the success of a search engine, such as the size, content of the database, the speed of searching, update frequency, the availability of search features, the interface design and ease of use (Willis 1996, McDonald 1996, Richard 1997, Lynch 1997).

Obviously, in order to be effective on the web, it is important to utilize the search engine most suited to people's subject domain. However, the above reviews neither include a ranking, which could help to make a decision for one specific search engine, nor did some of rankings provided have a scientific basis. Leighton (1995) considered this problem and used eight reference questions from a university library as search queries. By employing the evaluation criterion of precision, he compared Infoseek, Lycos, WebCrawler and World Wide Web Worm. But he counted only the number of relevant links, ignored the duplicates and the queries created by himself, resulting in experiment bias. Westera (1996) only used five queries to conduct precision study, and all these queries dealt with wine. These test suites were too small for statistical usefulness.

Chu and Rosenthal (1996) studied first ten precisions, took enough queries for statistical comparisons, recorded crucial information about how the searching was conducted, and performed some statistical tests. However, they did not subject their mean precision figures to any test for significance.

Gauch and Wang (1996) conducted twelve queries, studied almost all of the major search services (and even the major search services) and reported first twenty precision, but did not test for significance in the difference reported.

Tomaiuolo and Packer (1996) tested first ten precisions on two hundred queries and listed the query topics that they searched. But they used structured search expressions (using operators) and did not list the exact expression entered for each service. They reported the mean precision, but again did not test for significance in the differences. They even did not define the criteria for relevance.

Leighton (1997) conducted a new study for his Master's thesis to correct the problems presented in his early study in 1995. He tested the precision on the first twenty results returned for fifteen queries and used the Friedman's randomised block design to perform multiple comparisons for significance. To avoid the questions of bias, he used a blinding algorithm for evaluator to know from which search service the citation came by developing a PERL program. Clearly, it's impossible for a user to conduct the general evaluation by himself/herself and take several months.

Another interesting study conducted by Schilchting and Nilsen (1997), evaluated AltaVista, Excite, Infoseek and Lycos. They applied Signal Detection Theory to analyse the sensitivity (d') and how conservative or risky the search engine is (beta) for finding the useful information, but they did not conduct significance tests.

In summary, not only are the finding results different, but the evaluation criteria and methodologies used by those studies differed well. There is no generally effective methodology to compare and thus evaluate search engines so far.

## EXPERIMENT

From the above literature review, it is essential to conduct a statistical test and significance analysis for comparing the difference among search engines. This study used a factorial design to compare three search engines by evaluating their usability and qualities of information retrieved for a test suite. Noteworthy here, although each search engine claimed to have the various features, such as web size, content and currency of the database, update frequency etc., many previous studies have conducted these kinds of tests before and these features are being updated day by day. Furthermore, the end user will not care about these features and thus these tests have not been performed in this study. What they are concerned about here is the search efficiency, the precision and the quality of retrieved reference when they type in some queries on each search engine.

**Method** This a 8x2x3 factorial design, which included 8 queries in test suite, expert and novice subjects and three search tools compared. There were three independent variables and five dependent variables evaluated in this experiment and shown in Table 1. The effects between and within these three independent variables were tested with two major design concerns.

Table 1 **Experimental Variables**

| **Independent Variables** | **Dependent Variables** |
|---|---|
| Web Search Tool | Hit Rate (Precision) |
| User Expertise | Sensitivity d' |
| Test Suite | Decision Criterion Bias beta |
| | Response Time |
| | Interface format |

### 1. Concerns on Selection of Independent Variables

**Search Engines**: considering the basic classification of the search tools, the typical directory is Yahoo, which can enable user to conduct keyword and hierarchical searching. Search engines can be AltaVista developed by commercial company--Digital Equipment Corporation and Lycos developed by a faulty member in Carnie Mellon University. These typical search tools have been selected in this experiment.

**Subjects**: In order to test the user expertise effect on search results. Four subjects including two experts and two novices have been randomly chosen from the graduate students in university with different background and experience in the internet. The classification for expertise depends on the answers of a question set for subjects (see Appendix A). If a subject does interact with Internet weekly over 12 months and does know the advanced searching skill either Boolean or Truncation, he/she is defined as an expert, otherwise he/she is a novice.

**Development of the Query Suite:** there are two concerns in this part of design. First, considering the subjects' workload and their consequently performance, the number of query questions was set to be eight.

Second, Leighton (1997) mentioned the two choices about what words to use to search in his study: natural language (simple, unstructured text) and proximity and Boolean (structured text by using operators). In this study, the experimenter does not use operators in his routine searching and the simple queries can force the search engine to do more of the work, ranking results by its own algorithm rather than the constraints specified by the operators. Hence, the simple queries are selected as many as possible in this questionnaire, and only choose the necessary structured text when, without it, the topic is too easily open to multiple interpretations. Considering the different capabilities of each search tool, such as Boolean search, proximity search, field search, case-sensitivity search and concept search, as well as truncation, the query suite design included these factors and defined the complex query by using advanced searching skill such as Boolean or truncation search (see Appendix A).

**Consideration for Relevance of Query Questions:** In their study, Chu and Rosenthal (1996), they had two different persons judging some of the same results, so that one person’s biases counterbalanced by those of the other person. Leighton (1997) used criteria to establish categories for relevancy, such as irrelevant links (0), technically relevant links (1), potentially useful links (2) and the most probably useful links (3). The former method is inherently subjective and the latter seems to be too complicated and boring for subjects to perform the test. Fortunately, the experimenter himself for this study is also a general user as most others. Obviously, there should be a relevant criterion for subjects' judging relevance of the results concerning the different interpretation or interest for a query, and the search efficiency. It is also necessary to set a criterion for each query for both subjects' testing and precision calculation. From the practical point of view and the experimental limitation (time, subjects effort, different computer and browser such as Netscape or Explorer, etc.), there should be a general relevance criterion for each query, and this criterion should include the basic aspects of search query. Since the experimenter who is also a common user develops this general criterion, there should be no bias though this criterion is still subjective. And then the precision can be defined according to the number of the hits.

### 2. Concerns about the Testing of Dependent Variables

**Precision on Relevance:** considering the general case for people searching on the Internet, the first ten retrieval references are more important (or the most important). Every user hopes that the first reference matches what exactly he/she is looking for, but usually this expectation cannot be satisfied. However, if people could not get the targeted information in the first ten references, the efficiency would not be ranked high, even some users would give up and try another engine. Actually, the first ten precisions were designed to reflect the quality of service delivered to the user: how good is the relevance within the first ten pages of results. Therefore, the precision can be defined as the hits in the first ten references related to the relevance, coupling with another benefit of relatively high efficiency.

Here, there were still two aspects regarding to the relevance criteria should be noticed as Leighton (1997) pointed out. First, the duplicate link, a link in question has the same URL as a link earlier in the same return list or the URL is capitalised in one instance but not in another, should not be counted as a hit. Second, the inactive link, in which the access to this page is forbidden although the server is contacted, or the page has been moved to a new address but the server is not responding. In this case, the link should not be counted as a hit too. Actually, once a duplicate link or inactive link found, eliminating these kinds of links from the numerator should reduce the hit rate and the search engine is penalised due to the waste of time.

Therefore, each trail included a query and associated the first 10 numbers for three search tools, and the order of trails are individually randomised for each subject.

**Analysis of Sensitivity and Response Bias**: considering the efficiency for searching on the Internet, it is impossible for users to check every link retrieved by search engines. Especially for the huge search engines today, like AltaVista, containing up more than 30 million homepages, it can be suspected that testing the number of hit is still an effective measure. In addition, the duplicate links, inactive links, and irrelevant links not only reduce the searching efficiency, but also distort the measurement. Therefore the quality of the hits rather than their quantity is becoming more important concern. In order to objectively evaluate the quality of the results created by the search engine, the Signal Detection Analysis (SDA) has also been conducted in this study.

First, assign retrieved links by the three search engines into one of four categories as Table 2 shown. This categorization is based on the subject's binary

Yes/No judgement of each link's relevance, and this procedure was taken separately for each search engine associated with eight queries (as Schlichting mentioned in 1997).

Table 2 **SDA Categories for Returned Links**

| | Reported Links | Unreported Links |
|---|---|---|
| Good Links | **Hit**<br>Relevant Link found by target search engine | **Miss**<br>Relevant Link not found by target search engine but found by others |
| Bad Links | **False Alarm**<br>Irrelevant Link found by target search engine | **Correct Rejection**<br>Irrelevant Link not found by target search engine but found by others |

Second, calculate the hit rate and false alarm rate for each search engine. The hit rate is the proportion of "good links" (Hits) found by a search engine relative to the total number of "good links" found by all three search engines (Hits and Misses). Similarly, the false alarm rate is defined as the proportion of "bad links" (False Alarms) found by a search engine relative to the total number of "bad links" found by all three search engines (False Alarms and Correct Rejections).

Then, find out the relevant z scores by using standard formulas and tables based on the hit and false alarm rates (Wickens, 1992). The SDA uses the conceptual model shown in Figure 1 below, where the two distributions are, in the basic model, assumed to be normal (Gaussian). The SDA will also yield two scores for each search engine: d' = z(HIT) +z(CR), which measures the sensitivity of the search engine in finding useful information in its index, the larger the d', the better; beta = ordinate(z(HIT))/ordinate(z(CR)), which measures decision criterion bias, how conservative or risky the search engine is in reporting links, the larger the beta, the more conservative. Here, conservative behaviour means missing some hits in an effort to keep a lower false alarm rate, while the risky is accepting a bigger number of false alarm in exchange for reporting the higher hit rate.

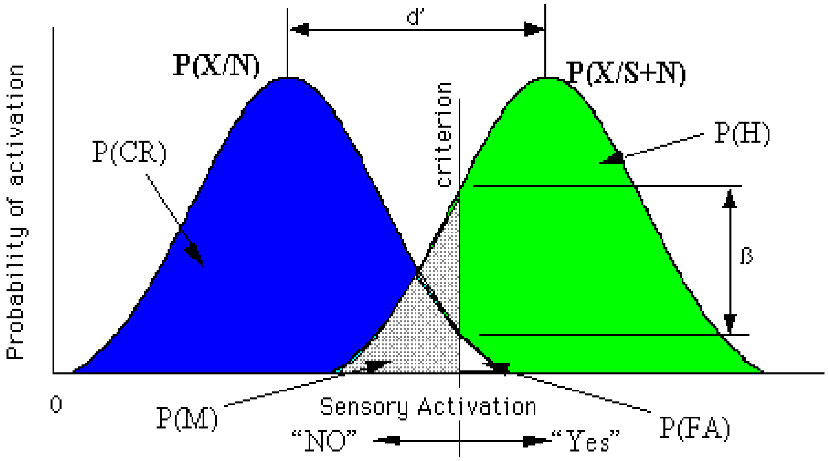

Figure 1. Signal Detection Analysis Model

Generally, d' ranges from 0 to 2. A score of 0 means that the search engine is unable to discriminate between good and bad links (signal and noise). If d' got a negative value, that means that the search engine would even report the bad link as a good one.

**Significance Test**: in order to verify the significant effect of search engine on the hit rate, sensitivity and decision criterion bias, the ANOVA has been conducted for analyzing the experimental results.

**Response Time**: clearly, the closer in time that a query is sought on the different search engines, the better for comparing. Ideally, a query should be sought on all three engines at the same time, but this is impossible for most subjects, especially when the search results come out simultaneously or mostly a subject has only one computer for searching. Hence, subjects should be asked to execute each search engine at a time for searching each query and record the relevant results or comments.

Certainly, once a result of a given query was popped up, checking the pages cited in this result should be as quickly as possible. Otherwise, the objective searching error will be larger. The whole questionnaire should be finished on the same day. However, considering the practical situation of individual subject, they are suggested to conduct the test at the same time of the other consecutive days. Obviously, subjects can distinguish the response difference among search engines when they retrieve every query on these search tools, and the response time can have only a relative definition with five ranks such as very slow, slow, fair, fast, and very fast (see Appendix A).

**Interface Format**: the interface output format was also tested in this experiment. Clearly, objective data cannot be collected, instead, the subjective ranking has been obtained from individual subject according to the five ranks (see Appendix A).

## RESULTS:

First, the raw data was analyzed in a three-way ANOVA, in which the group (experts vs. novices) was a between-subjects variable, and query questions and search tools were two within-subjects variables. Then further calculation provided the hit rate P(H), the false alarm rate P(FA), d' and beta, by conducting SDA. Another two-way ANOVA was conducted twice for sensitivity d' and decision criterion bias beta respectively. The relevant results can be acquired from below figures and tables.

Figure 2 indicated that there was no significant effect of expertise on the number of hits for the first ten retrieved links.

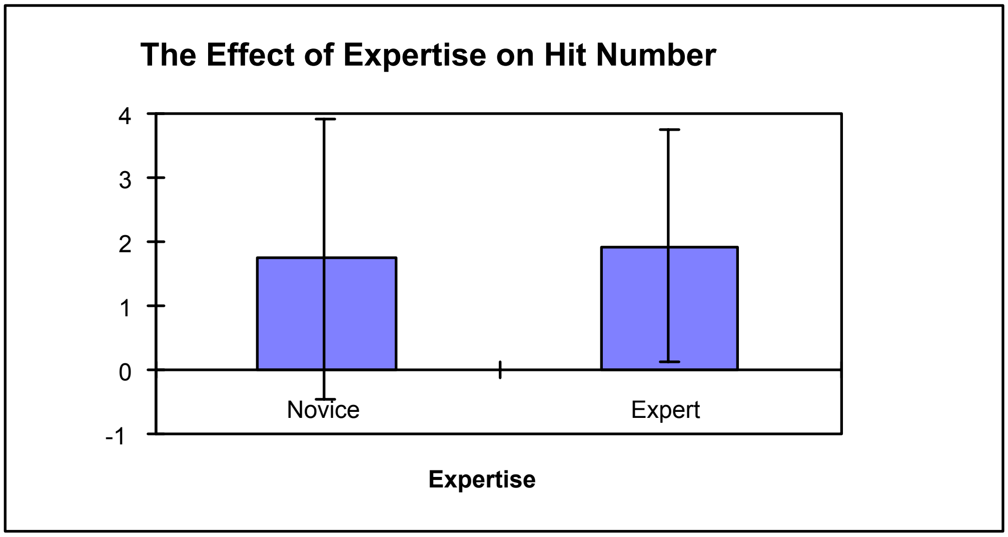

Figure 2. The Effect of Expertise on Hit Number

Figure 3 and Figure 4 showed that there were no interaction effects between expertise and query question, and between expertise and search engines.

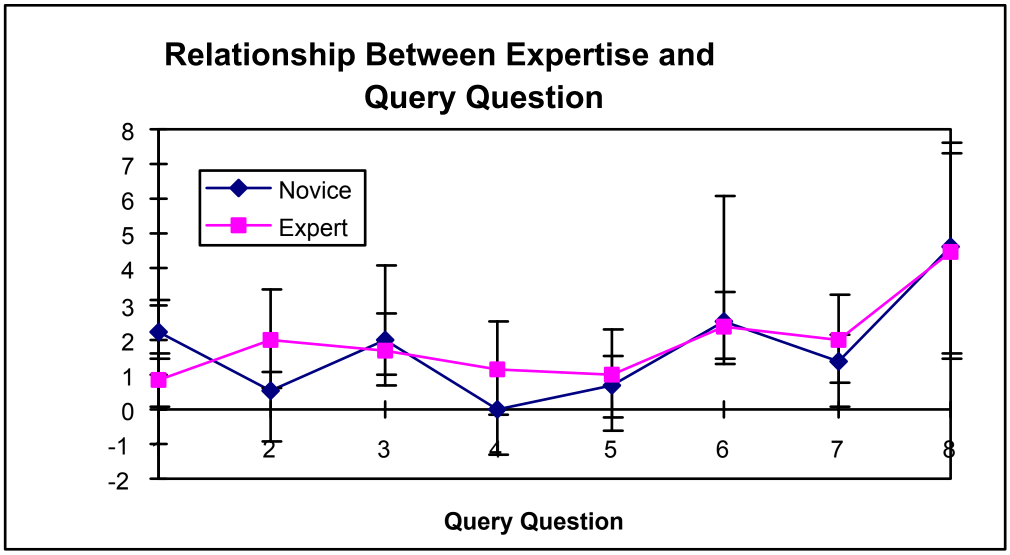

Figure 3. The relationship between Expertise and Query

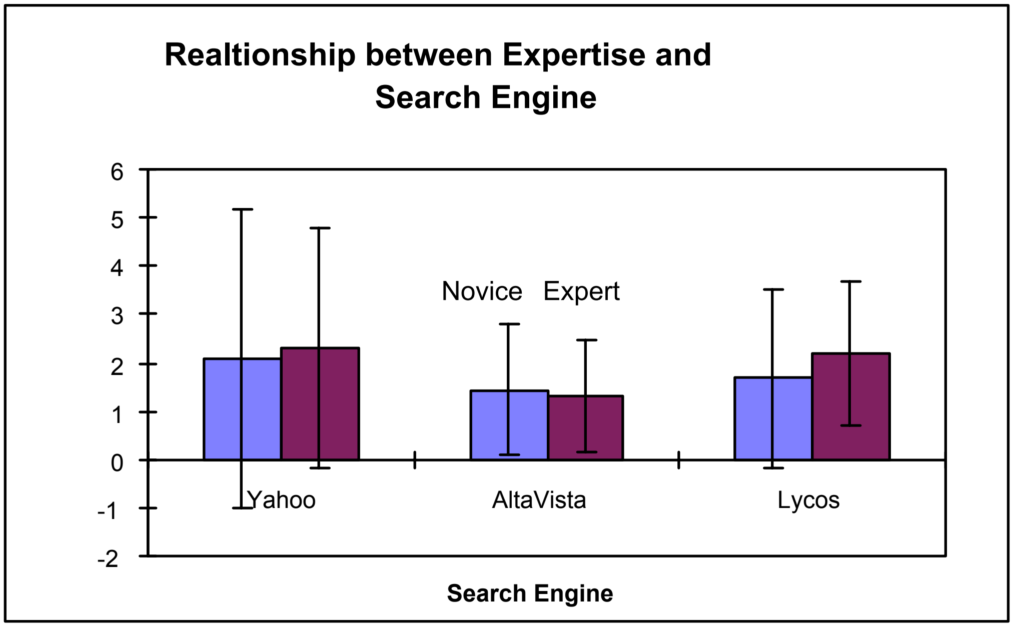

Figure 4. The relationship between Expertise and Search Engine

The effect of query question on hit number was significant (F(7, 48) = 4.944, P = 0.005). The more complex the question, the smaller the hit number. The main effect of query question can be seen in Figure 5.

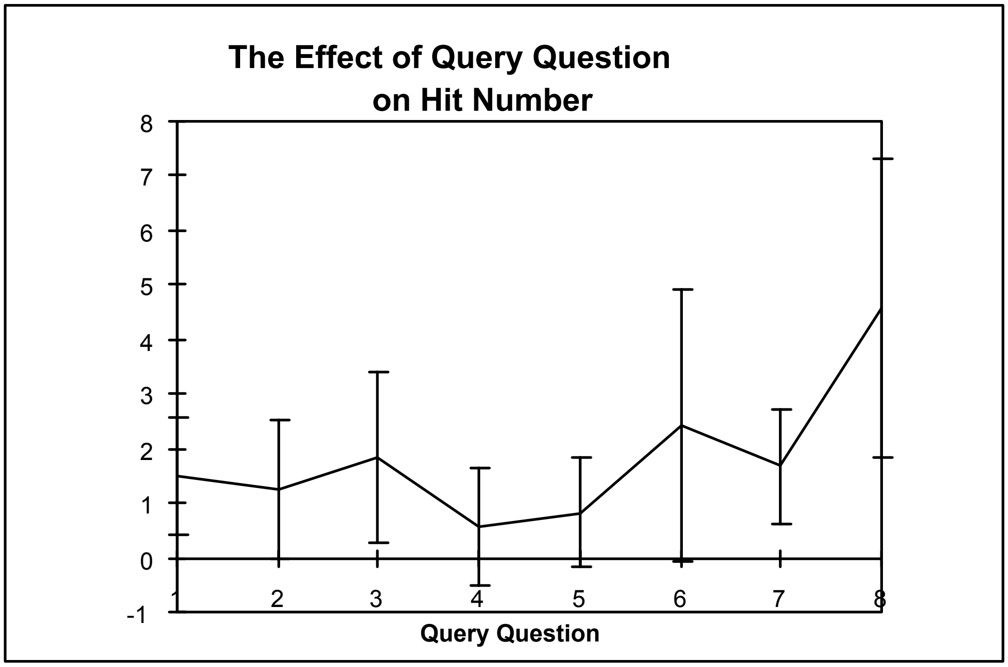

Figure 5. The Effect of Query Question on Hit Number

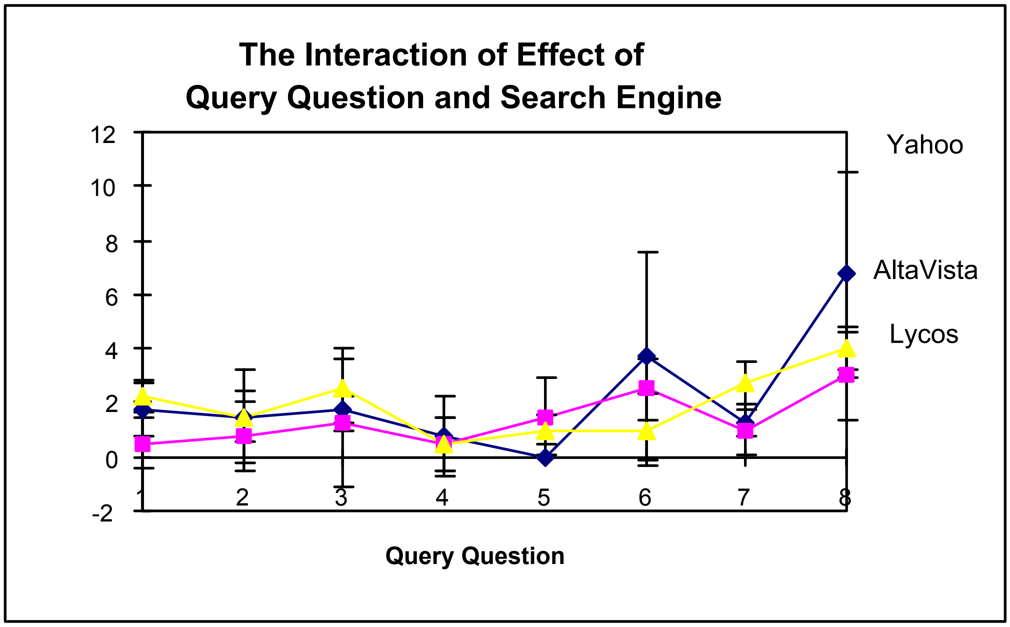

Figure 6. The Interaction of Effect of Query Question and Search Engine

An interaction effect between query question and search engine was significant (F(14, 48) = 2.127, P = 0.043), and is shown in Figure 6. There was no significant effect of search engine on the hit number, and the main effect of search engine is shown in Figure 7.

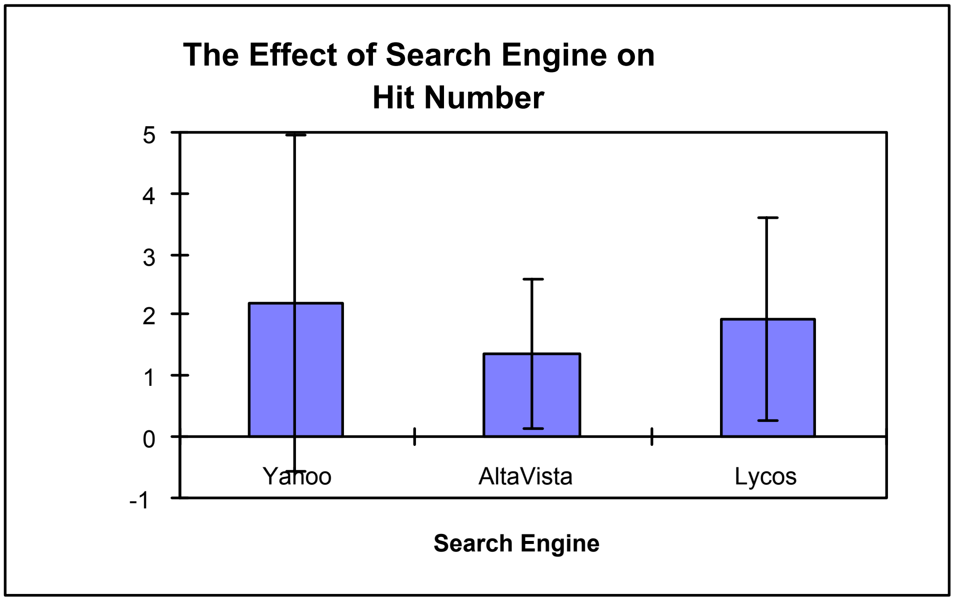

Figure 7. The Effect of Search Engine on Hit Number

Table 3. **Subjective Evaluation on Performance**

| **Search Engine** | **Response Time**<br>**(out of 5 ranking)** | **Output Format**<br>**(out of 5 ranking)** |
|---|---|---|
| Yahoo | 4.25 | 3.75 |
| AltaVista | 3.75 | 3.75 |
| Lycos | 3.75 | 2.75 |

For the performance of these three search engines, the subjective ranking scores were also collected from subjects according to the relevant response time and output format mentioned in above Method section (see Table 3).

Table 4. **Precision (P) Chart for Three Search Engines**

| | **P Value for Yahoo** | | **P Value for AltaVista** | | **P Value for Lycos** | | **Mean P Value** | |
|---|---|---|---|---|---|---|---|---|
| **Query** | First 5 | First 10 | First 5 | First 10 | First 5 | First 10 | First 5 | First 10 |
| **1** | 0.35 | **0.175** | 0.1 | **0.05** | 0.2 | **0.225** | 0.216 | **0.15** |
| **2** | 0.3 | **0.15** | 0.1 | **0.075** | 0.15 | **0.15** | 0.183 | **0.175** |
| **3** | 0.35 | **0.175** | 0.15 | **0.125** | 0.25 | **0.25** | 0.25 | **0.183** |
| **4** | 0.05 | **0.075** | 0.0 | **0.05** | 0.05 | **0.05** | 0.033 | **0.058** |
| **5** | 0.0 | **0.0** | 0.2 | **0.15** | 0.1 | **0.1** | 0.1 | **0.083** |
| **6** | 0.4 | **0.375** | 0.3 | **0.25** | 0.05 | **0.1** | 0.25 | **0.233** |
| **7** | 0.25 | **0.125** | 0.25 | **0.1** | 0.35 | **0.275** | 0.283 | **0.167** |
| **8** | 0.8 | **0.675** | 0.4 | **0.3** | 0.4 | **0.4** | 0.583 | **0.465** |
| **Mean P** | 0.31 | **0.218** | 0.187 | **0.137** | 0.22 | **0.193** | 0.275 | **0.189** |

Table 4 shows the precision of the first five and the first ten retrieved links (0.275 and 0.189), it is clear that the hit rate for first five retrieved links is a little bit higher than the hit rate for the first ten returned links.

By performing the Signal Detection Analysis (SDA), the sensitivity parameter d' and decision criterion bias value beta for each search engine were shown in Table 5. Yahoo got the highest d' and beta value (0.415 and 1.25), but AltaVista had a negative d' score (-0.425) and the lowest beta value (0.81) among three search engines. Table 5 also illustrates that the hit rate of Yahoo is the highest (41.05%) and AltaVista is the lowest (24.22%) among three search engines. On the contrary, the false alarm rate P(FA) of AltaVista is the highest (39.66%) and Yahoo is the lowest (23.42%).

Table 5 **SDA Scores for Each Search Engine**

| **SDA Features** | **Yahoo** | **AltaVista** | **Lycos** |
|---|---|---|---|
| Hit Rates P(H) | 41.05% | 24.22% | 36.82% |
| False Alarm Rate P(FA) | 23.42% | 39.66% | 35.70% |
| Measure of Sensitivity d' | 0.415 | -0.425 | 0.135 |
| Response Bias (Beta) | 1.25 | 0.81 | 1.05 |

Figure 8 depicted that search engines have significant effects on searching sensitivity d', and it can also be verified by the ANOVA result F(1,7) = 14.853, P = 0.014 (see Appendix C).

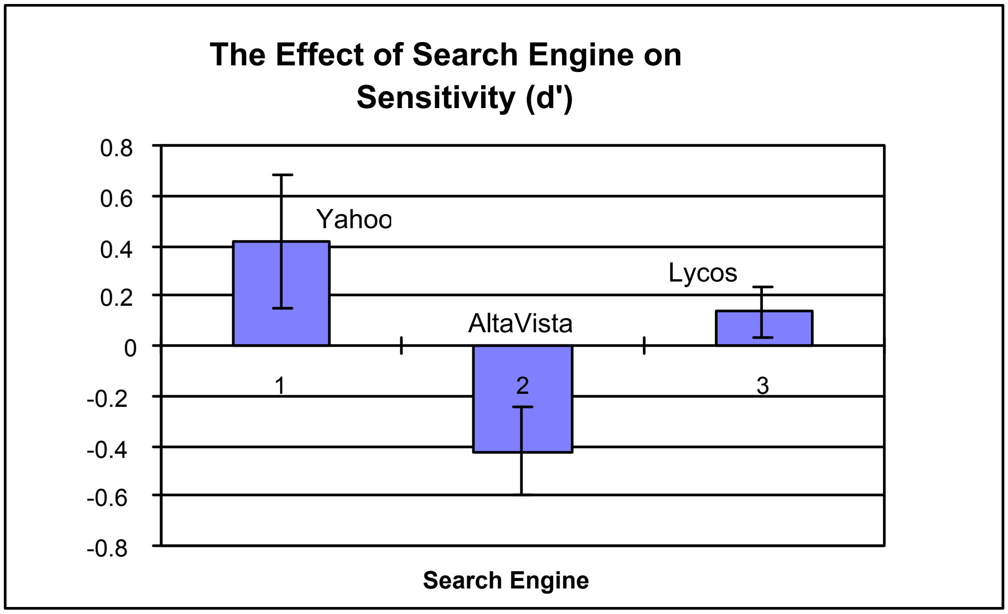

Figure 8. The Effect of Search Engine on Sensitivity d'

Figure 9 indicated that search engine has significant effect on decision criterion bias beta, and it can also be verified by the ANOVA result F(1,7) = 17.931, P = 0.010 (see Appendix D).

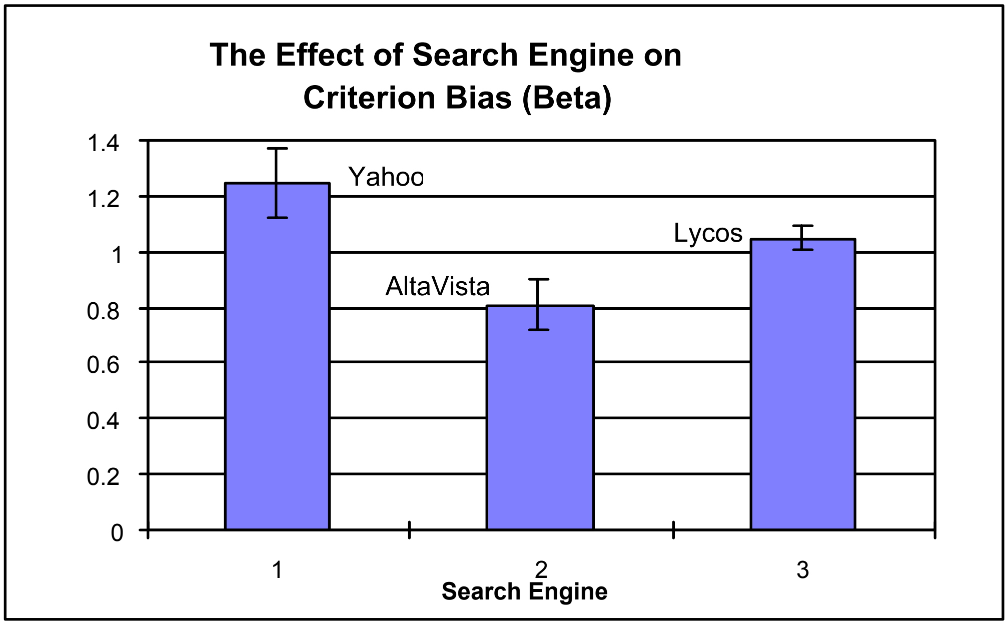

Figure 9. The Effect of Search Engine on Decision Criterion Beta

# DISCUSSION

## 1. Significant Factors and Their Interaction

The lack of significant effect of expertise shown in Figure 2 was not expected although there was some difference. The reason is that the criteria for each query is the same for experts and novices, regardless the experience in surfing on the internet. There should not be big difference for the same query and its relevance. This can also explain why there was no interaction between expertise and query question, and between expertise and search engine (see Figure 3 and Figure 4).

Figure 3 showed that experts did better on query 2 and 4, but worse on query 1. This is due to the nature of those query questions, since query 2 and 4 ("Austria AND classical music", "unemployment rate AND Toronto) are both complex query with more than two concepts and a stricter definition of relevance (see Appendix A). Experts with more knowledge and experience of searching skill (Boolean search for this case) are expected to get better scores. However, since query 1 ("WTO") is very simple and its criterion of relevance is also very specified, it is not surprised for novices to do better.

Contrary to the hypothesis, Figure 5, Figure 6, and the ANOVA results in Appendix B (F(7, 48) = 4.944, P = 0.005) indicated that the effect of query was significant. The interaction between query question and search engine was significant too. This unexpected result is due to the different features for different search engines. Yahoo and AltaVista support simple and advanced queries, and Lycos does not, even though the hit rate for Lycos is higher than that for AltaVista.

On the other hand, different search engine collects different information according to its own algorithm and constructs its own information resource. This suggests that the constructions (e.g. spiders and indexes) of each search engine are diversified enough that they consequently represent different portions of the entire magnificent web system. Hence, the experimental results cannot avoid the effect of search engine's own structure. Certainly, the experimental error may be the other reason.

Here, the interesting thing in Figure 6 is that Yahoo did exceptionally well on query 6 and 8 relative to other two search engines, but very bad on query 5. Why?. First, query 6 and 8 ("portable computer" and "human cloning") have a common feature----most current technologies, but query 5 ("Ming Dynasty") is an ancient story about one thousand years ago in the Orient China. On the other hand, the nature of Yahoo is a directory, which builds its construction and updates its index by people rather than software "Robot" or "Spider" for other two search engines. Generally, educated people are much more sensitive and interested in new technology around them, not a very old story happened very far away from them. Thus they possibly put much such information on Yahoo's index and update it frequently. This suggests that Yahoo is more suited for current technology searching rather than the ancient arts. Certainly, this argument can also be verified in the future research by comparing the results for searching old arts like "Shakespeare" etc.

Figure 7 and Table 4 indicated the failure of rejecting the hypothesis that there is no significant difference of usability among three search tools. From Table 4, although Yahoo obtained the highest precision score (0.218) among the three search engines while Lycos had 0.193 and AltaVista got the lowest (0.137), the standard deviation in Figure 7 (or even in previous figures) are higher than the mean. This is due to the stricter definition of "relevance", since the criterion is a page that is very likely to be useful for users. This criterion made the hit number very small, even for the precision of the first five returns (see Table 4). Therefore, we can say that there is no big difference among three search engines from this point of view.

### 2. Analysis of Sensitivity and Criterion Bias

From Figure 7 and the ANOVA results in Appendix A, it can be seen that there is no big difference of the precision among search engines, and the values for all three search engines are lower (Yahoo: 0.2178, AltaVista: 0.1375, Lycos: 0.1973). Considering the efficiency for searching, the experiment for checking the first 5 retrieved links has also been conducted and no big improvement was found for hit rate (0.189 and 0.275 in Table 4). This suggests that the number of hit does not permit any valuable conclusions about the usefulness of the various search engines because duplication of links and irrelevant links distort the measurement. Therefore, the quality of the hits rather than their quantity is becoming more important concern in this experiment. In order to objectively evaluate the quality of the results created by the search engine, the SDA and ANOVA mentioned in section of Method were conducted.

The relevant results shown in Figure 8, Figure 9 and Table 5 illustrated that search tools had significant effects on sensitivity and decision criterion bias in finding useful information. From Table 5, it can be seen that only Yahoo got a respectable d' score (0.415), this means that Yahoo can distinguish between good and bad retrieved links. Lycos displayed a small d' value (0.135), and made it difficult to be evaluated for its sensitivity. The negative value for AltaVista are especially confusing, that means it is very possible for this search engine to report a bad link as a good one (cannot discriminate between the signal and noise). It also suggested that the performance of this search engine in this experiment is too poor to fit the standard assumptions of SDA. One of the central assumptions is that the distribution of links containing relevant links (Hit and Misses) is more likely to be indexed by the search mechanism than those sites consisting irrelevant links (False Alarms and Correct Rejections) in search engine structure (Schlichting and Nilsen 1997). This violation of assumptions also makes the interpretation of the beta score more complex. Yahoo has a less risky decision criterion level compared to Lycos, but AltaVista had a beta value less than one (0.81) due to negative d' score.

### 3. Subjective Performance Evaluation

As mentioned in section of Method, it is impossible to conduct a standard test for response time and output format due to the difference search time and utilities when subjects conducting the experiment. The subjective ranking scores shown in Table 3 indicated that Yahoo (4.25/5) could retrieve results relatively faster than AltaVista and Lycos (3.75/5). The output formats (interface) of Yahoo and AltaVista (3.75/5) are better than that of Lycos (2.75/5).

## CONCLUSIONS

Considering the different results in this experiment by comparing three search tools----Yahoo, AltaVista and Lycos, some conclusions can be obtained as below.

1. There is no big difference of the precision on relevance for the first 10 retrieved links by each search tools.

2. Expertise had no significant effects on the hit number, nor the interaction with query and search engine.

3. Query suite had significant influences on the hit number as well as the interaction with search engine due to the different construction and index (mechanism) for each search engine. Yahoo is most suitable for current technology searching not for ancient arts.

4. Subjectively, Yahoo did best among three search engines in performance evaluation for response time and interface output format.

5. Objectively, search tools had significant effects on the sensitivity and decision criterion in finding useful information. Yahoo is more sensitive to retrieve useful information than Lycos, and it is difficult to determine the sensitivity and decision criterion for AltaVista due to its negative d' value in this experiment.

# REFERENCES:

[1] Box, George E. P., Hunter, W. G., and Hunter, J. S. "Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building", John Wiley & Sons, 1978.

[2] Chu, Heting & Rosenthal, Marilyn, "Search Engines for the World Wide Web: A Comparative Study and Evaluation on Methodology", ASIA 1996 Annual Conference Proceedings, Baltimore, MD, October, 1996, pp127-125.

[3] DeMoss, Timothy, "Gentlemen, Start Your Engines", POWER ENGINEERING, August 1996, pp10-12.

[4] Gauch, Susan and Guijun Wang, "Information Fusion with ProFusion", Webnet 96 Conference, San Francisico, CA, October 1996.

[5] Kingoff, Andrew, "Comparing Internet Search Engines", Computer, April 1997.

[6] Leighton, H. Vernon, "Performance of Four World Wide Web Index Services: Infoseek, Lycos, Webcrawler and WWW Worm", URL: http://www.winona.msus.edu/is-f/library-f/webind.htm. July 1996.

[7] Leighton, H. Vernon and Jaideep Srivastava, "Precision among World Wide Web Search Services (Search Engines): Alta Vista, Excite, Hotbot, Infoseek, Lycos", URL: http://www.winona.msus.edu/is-f/library-f/webind2/webind2.htm, June 1997.

[8] Liu, Jian, "Understanding WWW Search Tools", Reference Department, IVB Libraries. URL: http://www.indiana.edu/~librcsd/search. 1996.

[9] Lynch, Clifford, "Searching the Internet", SCIENTIFIC AMERICAN, March 1997, pp52--56.

[10] McDonald, Jason, "Simple Route to More Efficient Web Searches", MACHINE DESIGN 24, October 1996 pp78-80.

[11] Richard, Peter and Robert, Sikorski, "Smarter Searching", SCIENCE Vol. 277 15, August 1997, pp 976--979.

[12] Schlichting, Carton & Nilsen, Erik, "Signal Detection Analysis of WWW Search Engines", URL: http://www.microsoft.com/usability/webconf/schlitchiting.htm.

[13] Tomaiuolo, Nicholas G. and Joan G. Packer, "An Analysis of Internet Search Engines: Assessment of over 200 Search Queries", Computers in Libraries. V16 No. 6, June 1996, pp58.

[14] Westera, Gillian, "Search Engine Comparison: Testing retrial and Accuracy", URL: http://www.curtin.edu.au/curtin/staffpages/gwpersonal/senginestudy/result.htm

[15] Wickens, C. D., "Engineering Psychology and Human Performance" (second edition), Harper Collions Publishers, 1992.

[16] Willis, Lynn, "Touring the Internet", CHEMTECH, July 1996, pp19--20.

# Appendix A

## Questionnaire

Thanks a lot for your kindly cooperation for this study, your effort on answering this questionnaire is highly appreciated.

1. How often do you use the search engine to look for some information? Please check on one of the choices below to describe your state.

a. daily b. weekly c. biweekly d. monthly e. less frequently

2. How long have you been using search tools below? Please circle your answer:

a. 0--6 months b. 6--12 months c over 12 months

| | | | |
|---|---|---|---|
| Yahoo | a | b | c |
| AltaVista | a | b | c |
| Lycos | a | b | c |

3. Are you familiar with "Boolean" search, where you combine concepts by using "AND", "OR", "NOT" to get the search result. (circle your answer below)

Yes No

4. Are you familiar with "truncation" search, where allows you to search for strings of words or for difference of spelling within words (i.e. "wom*n" will find woman and women)? (circle your answer below)

Yes No

Please use each above search tool (Yahoo, Alta Vista and Lycos) at a time for each query question below and record your results and comments in the table regarding to the relevant instructions.

Thanks.

Instruction on how to count the hit number:

1. Type in each keyword in the list of questionnaire on each search engine and indicate the relevance of hit for the first ten links returned by putting in check marks on the table provided for you.

2. There will be four cases about accessing the links and please do not count a link as a relevant link if:

a. a link has the same URL as a link earlier in the same return list or the URL is capitalised in one instance but not in another, do not count this as a hit.

b. a link is forbidden to access, do not count it.

c. the information in a link return does not include all criteria following each query question, do not count it.

Do count a link as a relevant hit only if:

d. a link return can provide all information included by the criteria in each query.

Here is the URL of these three search tools:

Yahoo: http://www.yahoo.com/

AltaVista: http://www.altavista.com/

Lycos: http://www.lycos.com/

## Query Questions

1. WTO

Criteria: a. Find information which allows you to write a summary about this international organisation

b. Find the headquarters address of this organisation

2. Austria AND classical music

Criteria: a. Find information which allows you to write a summary about Austria classical music

b. Find two typical works and two famous composers

3. Darwinism

Criteria: a. Find information which allows you to write a summary about Darwin and his theory of evolution

4. unemployment rate + Toronto

Criteria: what was the unemployment rate in Toronto of last year

5. Ming Dynasty

Criteria: a. Find information which allows you to write a summary about this ancient China Dynasty

6. portable* computer

Criteria: a. Find information which allows you to write a summary about portable computer

b. Find information which allows you to create a list of portable computer (brand and model)

7. Sigmund Freud

Criteria: a. Find information which allows you to write a summary about Sigmund Freud

b. Find two books written by Sigmund Freud

8. human cloning

Criteria: a. Find information which allows you to write a summary about this new technology

Query No._____ Time:_______

| **Link Number:** | **Yahoo** | **Alta Vista** | **Lycos** |
|---|---|---|---|
| **1** | | | |
| **2** | | | |
| **3** | | | |
| **4** | | | |
| **5** | | | |
| **6** | | | |
| **7** | | | |
| **8** | | | |
| **9** | | | |
| **10** | | | |

Now please answer these two questions again:

1. Comparing the response time for three search engines, I think this search tools should be (please circle your answer)

1---very slow 2---slow 3---fair 4---fast 5---very fast

| | | | | | |
|---|---|---|---|---|---|
| Yahoo | 1 | 2 | 3 | 4 | 5 |
| AltaVista | 1 | 2 | 3 | 4 | 5 |
| Lycos | 1 | 2 | 3 | 4 | 5 |

2. For the comment on the interface output format of three search tools as the description: it is concise and easy to get the relevant information, my option is (please circle your answer):

1---strongly disagree, 2---disagree, 3---neutral, 4---agree, 5---strongly agree

| | | | | | |
|---|---|---|---|---|---|
| Yahoo | 1 | 2 | 3 | 4 | 5 |
| AltaVista | 1 | 2 | 3 | 4 | 5 |
| Lycos | 1 | 2 | 3 | 4 | 5 |

Thanks a lot!!!!!

# Appendix B

## ANOVA Result for Hit Rate

```
SOURCE: grand mean

Experti      Query  SearchE     N     MEAN        SD          SE

                                96    1.8333      1.9982      0.2039

SOURCE: Expertise

Experti Query SearchE N MEAN SD SE

Novice 48 1.7292 2.1903 0.3161

Expert 48 1.9375 1.8031 0.2603

SOURCE: Query

Experti Query SearchE N MEAN SD SE

1 12 1.5000 1.0871 0.3138

2 12 1.2500 1.2881 0.3718

3 12 1.8333 1.5859 0.4578

4 12 0.5833 1.0836 0.3128

5 12 0.8333 1.0299 0.2973

6 12 2.4167 2.5030 0.7226

7 12 1.6667 1.0731 0.3098

8 12 4.5833 2.7455 0.7926
```

SOURCE: Expertise Query

Experti Query SearchE N MEAN SD SE

Novice 1 6 2.1667 0.9832 0.4014

Novice 2 6 0.5000 0.5477 0.2236

Novice 3 6 2.0000 2.0976 0.8563

Novice 4 6 0.0000 0.0000 0.0000

Novice 5 6 0.6667 0.8165 0.3333

Novice 6 6 2.5000 3.5637 1.4549

Novice 7 6 1.3333 0.8165 0.3333

Novice 8 6 4.6667 2.6583 1.0853

Expert 1 6 0.8333 0.7528 0.3073

Expert 2 6 2.0000 1.4142 0.5774

Expert 3 6 1.6667 1.0328 0.4216

Expert 4 6 1.1667 1.3292 0.5426

Expert 5 6 1.0000 1.2649 0.5164

Expert 6 6 2.3333 1.0328 0.4216

Expert 7 6 2.0000 1.2649 0.5164

Expert 8 6 4.5000 3.0822 1.2583

SOURCE: SearchE

Experti Query SearchE N MEAN SD SE

Yahoo 32 2.1875 2.7526 0.4866

AltaV 32 1.3750 1.2378 0.2188

Lycos 32 1.9375 1.6644 0.2942

SOURCE: Expertise SearchE

Experti Query SearchE N MEAN SD SE

Novice Yahoo 16 2.0625 3.0869 0.7717

Novice AltaV 16 1.4375 1.3647 0.3412

Novice Lycos 16 1.6875 1.8518 0.4630

Expert Yahoo 16 2.3125 2.4690 0.6172

Expert AltaV 16 1.3125 1.1383 0.2846

Expert Lycos 16 2.1875 1.4705 0.3676

SOURCE: Query SearchE

Experti Query SearchE N MEAN SD SE

1 Yahoo 4 1.7500 0.9574 0.4787

1 AltaV 4 0.5000 0.5774 0.2887

1 Lycos 4 2.2500 0.9574 0.4787

2 Yahoo 4 1.5000 1.7321 0.8660

2 AltaV 4 0.7500 0.9574 0.4787

2 Lycos 4 1.5000 1.2910 0.6455

3 Yahoo 4 1.7500 0.5000 0.2500

3 AltaV 4 1.2500 1.5000 0.7500

3 Lycos 4 2.5000 2.3805 1.1902

4 Yahoo 4 0.7500 1.5000 0.7500

4 AltaV 4 0.5000 1.0000 0.5000

4 Lycos 4 0.5000 1.0000 0.5000

5 Yahoo 4 0.0000 0.0000 0.0000

5 AltaV 4 1.5000 0.5774 0.2887

5 Lycos 4 1.0000 1.4142 0.7071

6 Yahoo 4 3.7500 3.8622 1.9311

6 AltaV 4 2.5000 1.2910 0.6455

6 Lycos 4 1.0000 1.1547 0.5774

7 Yahoo 4 1.2500 0.5000 0.2500

7 AltaV 4 1.0000 0.8165 0.4082

7 Lycos 4 2.7500 0.9574 0.4787

8 Yahoo 4 6.7500 3.7749 1.8875

8 AltaV 4 3.0000 0.8165 0.4082

8 Lycos 4 4.0000 1.6330 0.8165

```
SOURCE: Expertise Query SearchE

Experti Query SearchE N MEAN SD SE

Novice 1 Yahoo 2 2.5000 0.7071 0.5000

Novice 1 AltaV 2 1.0000 0.0000 0.0000

Novice 1 Lycos 2 3.0000 0.0000 0.0000

Novice 2 Yahoo 2 0.5000 0.7071 0.5000

Novice 2 AltaV 2 0.5000 0.7071 0.5000

Novice 2 Lycos 2 0.5000 0.7071 0.5000

Novice 3 Yahoo 2 1.5000 0.7071 0.5000

Novice 3 AltaV 2 1.0000 1.4142 1.0000

Novice 3 Lycos 2 3.5000 3.5355 2.5000

Novice 4 Yahoo 2 0.0000 0.0000 0.0000

Novice 4 AltaV 2 0.0000 0.0000 0.0000

Novice 4 Lycos 2 0.0000 0.0000 0.0000

Novice 5 Yahoo 2 0.0000 0.0000 0.0000

Novice 5 AltaV 2 1.5000 0.7071 0.5000

Novice 5 Lycos 2 0.5000 0.7071 0.5000

Novice 6 Yahoo 2 4.5000 6.3640 4.5000

Novice 6 AltaV 2 3.0000 1.4142 1.0000
```

Novice 6 Lycos 2 0.0000 0.0000 0.0000

Novice 7 Yahoo 2 1.0000 0.0000 0.0000

Novice 7 AltaV 2 1.0000 1.4142 1.0000

Novice 7 Lycos 2 2.0000 0.0000 0.0000

Novice 8 Yahoo 2 6.5000 4.9497 3.5000

Novice 8 AltaV 2 3.5000 0.7071 0.5000

Novice 8 Lycos 2 4.0000 0.0000 0.0000

Expert 1 Yahoo 2 1.0000 0.0000 0.0000

Expert 1 AltaV 2 0.0000 0.0000 0.0000

Expert 1 Lycos 2 1.5000 0.7071 0.5000

Expert 2 Yahoo 2 2.5000 2.1213 1.5000

Expert 2 AltaV 2 1.0000 1.4142 1.0000

Expert 2 Lycos 2 2.5000 0.7071 0.5000

Expert 3 Yahoo 2 2.0000 0.0000 0.0000

Expert 3 AltaV 2 1.5000 2.1213 1.5000

Expert 3 Lycos 2 1.5000 0.7071 0.5000

Expert 4 Yahoo 2 1.5000 2.1213 1.5000

Expert 4 AltaV 2 1.0000 1.4142 1.0000

Expert 4 Lycos 2 1.0000 1.4142 1.0000

```
Expert 5 Yahoo 2 0.0000 0.0000 0.0000

Expert 5 AltaV 2 1.5000 0.7071 0.5000

Expert 5 Lycos 2 1.5000 2.1213 1.5000

Expert 6 Yahoo 2 3.0000 1.4142 1.0000

Expert 6 AltaV 2 2.0000 1.4142 1.0000

Expert 6 Lycos 2 2.0000 0.0000 0.0000

Expert 7 Yahoo 2 1.5000 0.7071 0.5000

Expert 7 AltaV 2 1.0000 0.0000 0.0000

Expert 7 Lycos 2 3.5000 0.7071 0.5000

Expert 8 Yahoo 2 7.0000 4.2426 3.0000

Expert 8 AltaV 2 2.5000 0.7071 0.5000

Expert 8 Lycos 2 4.0000 2.8284 2.0000

FACTOR : Subject Expertise Query SearchE Hitnumber

LEVELS : 4 2 8 3 96

TYPE : RANDOM BETWEEN WITHIN WITHIN DATA

SOURCE SS df MS F p

===============================================================

mean 322.6667 1 322.6667 22.610 0.041 *

S/E 28.5417 2 14.2708
```

Experti 1.0417 1 1.0417 0.073 0.812

S/E 28.5417 2 14.2708

Query 131.3333 7 18.7619 4.944 0.005 **

QS/E 53.1250 14 3.7946

EQ 17.2917 7 2.4702 0.651 0.709

QS/E 53.1250 14 3.7946

SearchE 11.0833 2 5.5417 2.375 0.209

SS/E 9.3333 4 2.3333

ES 1.5833 2 0.7917 0.339 0.731

SS/E 9.3333 4 2.3333

QS 57.4167 14 4.1012 2.127 0.043 *

QSS/E 54.0000 28 1.9286

EQS 14.5833 14 1.0417 0.540 0.887

QSS/E 54.0000 28 1.9286

# Appendix C

## ANOVA Result for d'

```
SOURCE: grand mean

Experti Tools N MEAN SD SE

12 0.0417 0.4047 0.1168

SOURCE: Expertise

Experti Tools N MEAN SD SE

Expert 6 -0.0217 0.4388 0.1792

Novice 6 0.1050 0.3975 0.1623

SOURCE: Tools

Experti Tools N MEAN SD SE

Yahoo 4 0.4150 0.2664 0.1332

AltaV 4 -0.4250 0.1756 0.0878

Lycos 4 0.1350 0.1038 0.0519

SOURCE: Expertise Tools

Experti Tools N MEAN SD SE

Expert Yahoo 2 0.3800 0.2263 0.1600

Expert AltaV 2 -0.5450 0.0919 0.0650

Expert Lycos 2 0.1000 0.0566 0.0400
```

```
Novice Yahoo 2 0.4500 0.3960 0.2800

Novice AltaV 2 -0.3050 0.1626 0.1150

Novice Lycos 2 0.1700 0.1556 0.1100

FACTOR : Subject Expertise Tools d

LEVELS : 4 2 3 12

TYPE : RANDOM BETWEEN WITHIN DATA

SOURCE SS df MS F p

==================================================================

mean 0.0208 1 0.0208 0.569 0.529

S/E 0.0732 2 0.0366

Experti 0.0481 1 0.0481 1.315 0.370

S/E 0.0732 2 0.0366

Tools 1.4635 2 0.7317 14.853 0.014 *

TS/E 0.1971 4 0.0493

ET 0.0193 2 0.0096 0.196 0.830

TS/E 0.1971 4 0.0493
```

# Appendix D

## ANOVA Result for Beta

```
SOURCE: grand mean
Experti Tools N MEAN SD SE
12 1.0358 0.2050 0.0592
SOURCE: Expertise
Experti Tools N MEAN SD SE
Expert 6 1.0100 0.2286 0.0933
Novice 6 1.0617 0.1964 0.0802
SOURCE: Tools
Experti Tools N MEAN SD SE
Yahoo 4 1.2475 0.1253 0.0626
AltaV 4 0.8100 0.0902 0.0451
Lycos 4 1.0500 0.0476 0.0238
SOURCE: Expertise Tools
Experti Tools N MEAN SD SE
Expert Yahoo 2 1.2500 0.0707 0.0500
Expert AltaV 2 0.7500 0.0707 0.0500
Expert Lycos 2 1.0300 0.0141 0.0100
```

```
Novice Yahoo 2 1.2450 0.2051 0.1450

Novice AltaV 2 0.8700 0.0707 0.0500

Novice Lycos 2 1.0700 0.0707 0.0500

FACTOR : Subject Expertise Tools Beta

LEVELS : 4 2 3 12

TYPE : RANDOM BETWEEN WITHIN DATA

SOURCE SS df MS F p

=============================================================

mean 12.8754 1 12.8754 1326.223 0.001 ***

S/E 0.0194 2 0.0097

Experti 0.0080 1 0.0080 0.825 0.460

S/E 0.0194 2 0.0097

Tools 0.3840 2 0.1920 17.931 0.010 *

TS/E 0.0428 4 0.0107

ET 0.0080 2 0.0040 0.374 0.710

TS/E 0.0428 4 0.010
```