The Overall Study of the Business Intelligence Explorer

Dr. Zhaohui Yu
Post-doctor of Fudan University
Associate Professor of College of International Business
Shanghai International Studies University
Shanghai 200083, China
E-mail: yuzhaohui2010@163.com

Xiang Ji
College of International Business
Shanghai International Studies University
Shanghai 200083, China
E-mail: paralellmiles@gmail.com

Abstract

Business Intelligence Explorer uses a new browsing method and the framework incorporates visualization, web mining and clustering techniques to support effective exploration of knowledge. To examine whether the business intelligence explorer did optimize the search result or not, this paper chose three research objects, Google, Quintura, Clusty, and conducted an analysis of variance in terms of efficiency, effectiveness and usability. The result shows that visualization and clustering techniques offers practical implications for search engine users.

Keywords: Business intelligence, Information overload, Business intelligence explorer

1. INTRODUCTION

Internet is one the five important tools company uses in finding business information. Traditional explorer displays the results in a long textural list which causes much time to find the related ones. Too much information made people not to be able to summarize all the useful one and waste lot of time in exploring. This is the problem of information overload. Many researchers developed explorers with business intelligence techniques to solve this problem.

The objectives of this research were to compare the effectiveness, efficiency and usability of the business intelligence explorer with that of the traditional one. We chose Google represents the traditional explorer, Quintura represents the Knowledge Map explorer and Clusty represents the Web Community. And we used questionnaires to get first hand data. Then, all the data was put into Statistics Package for Social Science (SPSS) software to do the analysis of variance (ANOVA). The analysis report can clearly shows the advantages and disadvantages of business intelligence explorer.

This research is very meaningful, for the business intelligence explorer can be a new method in solving the information overload problem if it turned out to be effective and efficient in searching and displaying information. Moreover, the research will also reveal the disadvantages of business intelligence explorer which points out the direction of future research.

2. LITERATURE REVIEW

2.1 Business Intelligence Tools

Business Intelligence is the whole process of accumulation, transformation, adjusting, assessment and expanding of the information. It can help company to make decisions. Recent years, many information analyze and processing tools have been developed. And they are the important method in internet information analysis. Many of those tools are open to the internet. They all try to solve the information overload problem. Compared with the traditional explorer, business intelligence explorer uses visual frameworks in displaying the results. So, it is better than traditional explorer in information analyzing. For example, business intelligence tools can analyze the customers' behavior through data mining and knowledge discovery, which helps in sales increasing.

2.2 Information Overload

Research papers mainly talked about two ways in solving the information overload problem. They are Search Engine Optimization and Visual Framework.

2.2.1 Search Engine Optimization

Among all the search engines, Meta-fact Search Engine is the most mature one. This search engine conduct search through three stages, they are problem transformation, result finding and result display. All of them are based on natural language processing, information retrieval and computational linguistic. And Meta-fact search engine combined many search engines together which makes it more punctuate.

2.2.2 Visual Framework

Shneiderman separated the display method into several ways based on statistic category. Traditional display shows the result in a long texture list. This is widely used as it is easy to realize, but it does not work efficiently. Comparatively speaking, two-dimensional display, tree structures and net structures show the result clearly, but ask for a complex process. Lin divided display method into four ways: hierarchical displays, network displays, scatter displays, and map displays. Among them, hierarchy (tree) display was found to be an effective way for browsing.

2.2.3 Business Intelligence Explorer

The researches about business intelligence explorer are mainly focus on algorithms and result display. Some paper also studied the usage of business intelligence explorer in certain industries. For example, Alexandra Robbins introduced intelligence search into FBI work. Koen Pauwels and other people found that business intelligence can improve the sales quantity of an online business site.

Few of the researches are about the comparative study of the business intelligence explorer and the traditional one. So, this kind of study is urgently needed.

3. BUSINESS INTELLIGENCE EXPLORER

Business intelligence explorer works in three main phases: data collection; parsing, indexing, and analysis; and visualization.

Data were collected in two steps: identifying key terms and meta-searching. Key topics identified in the first step were used as input queries in meta-searching.

Then, the term was get from the web pages and the term's level of importance is measured by term frequency and inverse Web page frequency.

Co-occurrence analysis converts the terms into a matrix that shows the similarity between every pair of Web sites. The similarity between every pair of Web sites contained its content and structural information.

The last phase is to define a web community or a knowledge map. To identify Web communities for each business intelligence topic, a combination of hierarchical and partition clustering was chosen. A normalized cut criterion was used as the genetic algorithms fitness function, which measures both the total dissimilarity between different partitions as well as the total similarity within the partitions. When partitioning the Web graph into Web communities, the genetic algorithms tried to make the web pages within one community closer to each other.

In Knowledge map creation, MDS is used to transform a high-dimension similarity matrix into a two-dimensional representation of points and displayed them on a map. Then, use formulas to find out the coordinates of points on the map.

4. STATISTICAL ANALYSIS

4.1 Introduction of the research objectives and object

The objective of the experiment is comparing the effectiveness, efficiency and usability of business intelligence explorer and the traditional one.

Business intelligence explorer can be classified into two groups, one is the Knowledge Map, and another is the Web Community. Knowledge Map search engine presents search results as interconnected objects on a map. The line between each object shows their relationship. Web Community search engine generate several communities according to the result. Pages in each community have a same or similar content.

This research chose three search engines as representatives. As Google (www.google.com) is widely used, it has been chosen represent the Traditional explorer. Quintura (www.quintura.com) shows its result in a knowledge map, so it has been chosen represent the Knowledge Map. And Clusty (clusty.com) generate different communities, so it has been chosen represent the Web Community.

4.2 Sample Collecting

To get first hand statistics, a questionnaire was designed. In this research, the respondents are asked to use three explorers to answer some questions. Each respondents need to use each of the three search engines to finish several questions. The questions are divided into three parts: close-ended questions, open-ended questions and user comment. In the close-ended questions, it asks for a company's URL. The key words and correct answer are given to the respondents, and their job is to do the search, note the time and check whether the answer is the same as the correct one. In the open-ended questions, it asks for the companies name in a certain field. The key words are also given, and the respondents should try to figure out the related answers within first 100 results. Finally, users give a comment of each of the search engines according to the general performance. The scores they give on a scale of 1 to 5.

The respondents were seldom chosen from different colleges and different ages in order to maintain variance. The total number is more than thirty. Half of them are female and they are younger than thirty.

From the questionnaires, we collected six types of statistics.

- Number of correctly answered questions
- Total number of questions
- Number of relevant results
- Total number of results
- Searching time
- User comments

Some of the statistics can be used directly for analysis, but some of them need to be assessed.

The explorer performance factors studies in the experiment were effectiveness, efficiency, and usability. Effectiveness is measured by three components: accuracy, precision, and recall. Accuracy refers to how well the search engines help users find the right answers. Precision refers to how well the search engines help users find the relevant result and avoid the irrelevant ones. Recall measures how well the search engines find all the relevant results. A P-Value was used to combine precision and recall together to ease the analysis procedure. An expert of the University spent more than 1 hour on each of the open-ended questions to find as much relevant results as possible. The formulas to calculate the above measurements are showed below:

$$\text{Accuracy} = \frac{\text{Number of correctly answered questions}}{\text{Total number of questions}} \qquad (1)$$

$$\text{Precision} = \frac{\text{Number of relevant results}}{\text{Number of all results}} \qquad (2)$$

$$\text{Recall} = \frac{\text{Numer of relevant results}}{\text{Number of relevant results identified by expert}} \qquad (3)$$

$$\text{P-Value} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (4)$$

In equation 1, Number of correctly answered is the correct answered number collected from the questionnaires. Total number is all the close-ended questions in the questionnaire.

In equation 2, the number of relevant results is the relevant results in the open-ended questions. The number of all results is all the open-ended questions in the questionnaire.

In equation 3, the number of relevant results is the same as equation 2 and the number of relevant results identifies by the expert is the results the expert found out.

In equation 4, the precision is the result of equation 2 and the recall is the result of equation 3.

Efficiency refers to the amount of time users required to use a browsing method to finish the tasks. It is measured by searching time. Usability refers to how satisfied users are with using a browsing method. It is measured by user comments.

4.3 Analysis tool

This experiment uses Statistics Package for Social Science (SPSS) software to do the analysis of variances. SPSS is one of the world most prominent analysis software. It offers many analysis methods, statistic define and flexible display charts. As the research objective is to compare the difference between three explorers, the experiment chooses ANOVA test.

4.4 Statistical Analysis method

ANOVA compares group means by analyzing comparisons of variance estimates. In order to testify whether the variable have an influence on the dependent variables, we should hypothesis that the influence does not exist. Then we use test statistic to prove whether the hypothesis is true or not. The detailed steps are make hypothesis, build test statistic and analysis result.

4.4.1 Make hypothesis

In ANOVA, we should first hypothesis that there are no differences between each group that is H0. And H1 is that the differences exist. If the H0 is denied, then we can say H1 is proved to be true and there exist differences between each group.

4.4.2 Build test statistic

The sample mean value is calculated with the formula below:

$$x = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} \quad (5)$$

In equation 5, n_i is the number of sample i, x_{ij} is number j value of sample i.

The total statistic mean value is calculated with the formula below:

$$y = \frac{\sum_{i=1}^{k} n_i x_i}{n} \quad (6)$$

In equation 6, n is the total number of statistics and x_i is the value of sample i.

The variance is the mean of the squared deviations about the mean (MS) or the sum of the squared deviations about the mean (SS) divided by the degrees of freedom. Two independent estimates of the population variance can be obtained.

Sum of the squared deviations (SST), is the sum of the squared deviations of the value x_{ij} and the total means y. The formula is:

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - y)^2 \quad (7)$$

Sum of the squared deviations between groups (SSA), is the sum of the square deviations of the group means x and the total means y. The formula is:

$$SSA = \sum_{i=1}^{k} n_i (x - y)^2 \quad (8)$$

Sum of the squared deviations within groups (SSE), is the sum of the square deviations of the value x_{ij} and the group means x. The formula is:

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - x)^2 \quad (9)$$

The relationship between the three deviations is: SST=SSA+SSE.

The formula of the mean of the squared deviations between groups (MSA) is:

$$\mathrm{MSA} = \frac{\mathrm{SSA}}{k-1} \tag{10}$$

And the formula of the mean of the squared deviations within groups (MSE) is:

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{n-k} \tag{11}$$

Test statistic (F-Value) formula is:

$$F = \frac{\mathrm{MSA}}{\mathrm{MSE}} \sim F\ (k-1,\ n-k) \tag{12}$$

4.4.3 Analysis decisions

To make a decision, we should compare the F-value with the significance level (α). If F>F_α, then deny H0. If F<F_α, then H0 is proved to be true. According to the analysis procedures, this experiment's hypothesis is stated below:

H0: Web community is same with Traditional explorer in effectiveness, efficiency or usability.

H1: Web community is more effective, efficient and usable than traditional explorer.

5. RESEARCH RESULT

In this experiment, accuracy, P-Value, user comments and searching time were putted into SPSS. The significance level is set as 0.05. The analysis is ANOVA. Dependent variables are accuracy, F-Value, user comment and search time. The ANOVA result is showed in the note 1.

From the table 1 we can see that the accuracy and research time of each explorer is about the same. But the F-Value and user comments show some differences. Web community's F-Value is the highest one. The number is 0.490790. And Knowledge map has the highest user comment. The number is 2.91.

In the table 2, the F-Value and user comments significance level are 0.00 and 0.00. Both lower than 0.05. It means that the three explorers differ in F-Value and user comments.

Table 3 shows the detailed comparisons between every two explorers. We can see that the three explorers are common in accuracy and search time, but different in F-Value and user comments. According to the table 1, web community is the most effective one and knowledge map has the best user comments.

6. DISCUSSION

From the results, the F-value is bigger than test statistic. So the H1 is proved to be true. The experiment proved that the business intelligence is better than traditional explorer in effectiveness and usability. The means of F-Value of web community is 0.491, higher than 0.256 or 0.291. F-Value represents the precision and recall of an explorer. It means that the result web community displays is more relevant to the key words than the traditional one. It is very important in solving the information overload problem. Users may waste less time in finding what they want. But still, the statistic 0.491 is not good. The only half of the result it displays is relevant. It is the same in user comments. Although the business intelligence explorer's statistic is higher than that of the traditional one, 2.91 are not very high. Generally speaking, three points can be summarized from the experiment.

6.1 Business intelligence explorer has its advantages

The analysis proves that the F-Value of business intelligence explorer is much higher than the traditional one and the business intelligence explorer is convenient for the users. It is because it uses Torgerson's classical MDS procedures in forming the knowledge map. And the algorithms can classify the result into groups. In this way, the results are more relevant to the key words. Those irrelevant entries had been eliminated. With this character, people in the future can try to use this explorer or the same visual framework to deal with information overload.

6.2 Business intelligence explorer has its disadvantages

From the analysis, we can see that there is not much difference in accuracy and research time. The reason is not because the business intelligence explorer is not good enough. For the accuracy of the three explorers are all higher than 0.9, the web community is even 1. This explains that why there are no differences, because all of them are very mature in accuracy and research time.

6.3 The advantages need to be further developed

As we had mentioned before, business intelligence explorer is good in the field of efficiency. But the absolute score of F-Value and user comments are not high. A further development of grouping algorithms and visual framework is needed. In the future, researchers can pay attention to these two fields.

7. CONCLUSION

In this paper, we explained the researches about business intelligence tools and we did an experiment to prove that the business intelligent explorer can, to some extent, help solve the information overload problem and it does have some advantages compared to the traditional one. The experiment uses SPSS to do the ANOVA test. The statistics is getting from questionnaires. There are three search engines in this experiment. The result is that there exist difference between business explorer and the traditional one in efficiency and usability because of its accurate clustering and appealing visualization and there is no difference of searching time between the two kinds of explorer. Contrary to the expectations, traditional explorer also has a high accuracy and short searching time because the technology in these two fields had already fully developed.

The contributions of this research are threefold. First, this article uses a quantitative method to prove the superiority of business intelligence search engine. The ANOVA test shows that the business intelligence explorer is superior in effectiveness and usability. Second, business intelligence explore must consider the user experience. Some of the user comments mentioned that the inadequate zooming function is the weakness of knowledge map and people need some time to get familiar with the usage. Third, there is need for further development of business intelligence search engines in effectiveness and usability. The score of effectiveness and usability is not very high, so a further development is needed.

With the limit of time and competence, the experiment still has some limitations. The number of sample is not very large and this may influence the result of analysis. And the number of research objects is also very few.

References

Bowman, CM., Danzig, P.B., Manber, U. and Schwartz, F. (1994). Scalable Internet resource discovery: Research problems and approaches. *Communications of the ACM*, 37, 8, 98-10.

Cutting, D.R., Karger, D.R., Pederson, J.O. and Tukey, J.W. (1992). Scatter/gather: A clusterbased approach to browsing large document collections. Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, pp. 318-329.

Davies, P.H.J. (2002). Intelligence, information technology, and information warfare. *Annual Review of Information Science and Technology*, 36, 313-352.

Dmitri Roussinov and Michael Chau. (2008). *Journal of the Association for Information Systems*, Volume 9, Issue 3/4, pp. 175-199, Special Issue.

Fuld, L.M., Singh, A., Rothwell, K. and Kim, J. (2003). Intelligence Software Report™ 2003: Leveraging the Web. Cambridge, MA: Fuld & Company.

Futures-Group Ostriches & Eagles. (1997). The Futures Group Articles, Washington, DC.

Gessner, Guy & Scott, Richard A. (2009). Information Systems Management, Vol. 26, Issue 2, p199-208, Spring.

Lin, X. (1997). Map displays for information retrieval. *Journal of the American Society for Information Science*, 48, 1, 40-54.

Richard J. Harris. (1994). ANOVA: an analysis of variance primer. ISBN 0875813739, pp. 10–20.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of IEEE Symposium on Visual Languages. Los Alamitos*. CA: IEEE Computer Society Press, pp. 336-343.

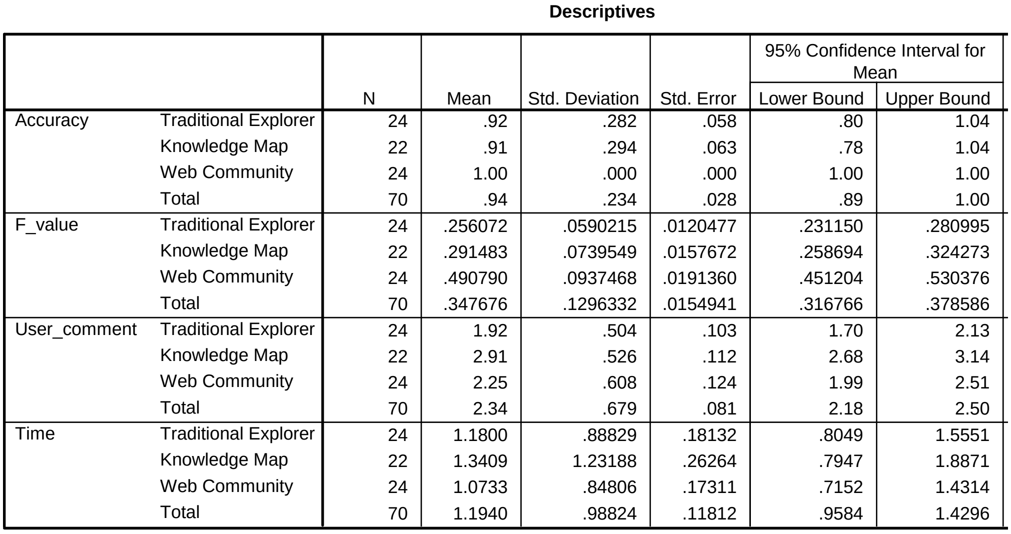
Torgerson, W.S. (1952). *Multidimensional scaling: I. Theory and method.* Psychometrika, 17, A, 401-419.

Wingyan Chung, Hsinchun Chen, Nunamaker Jr. & Jay F. (2005). *Journal of Management Information Systems,* Spring, Vol. 21, Issue 4, p57-84.

Notes

Note 1.

Table 1. DESCRIPTIVE

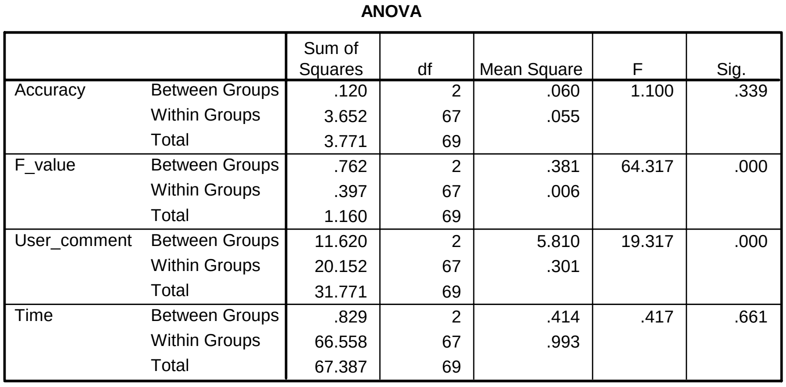


Descriptives

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
						Lower Bound	Upper Bound
Accuracy	Traditional Explorer	24	.92	.282	.058	.80	1.04
	Knowledge Map	22	.91	.294	.063	.78	1.04
	Web Community	24	1.00	.000	.000	1.00	1.00
	Total	70	.94	.234	.028	.89	1.00
F_value	Traditional Explorer	24	.256072	.0590215	.0120477	.231150	.280995
	Knowledge Map	22	.291483	.0739549	.0157672	.258694	.324273
	Web Community	24	.490790	.0937468	.0191360	.451204	.530376
	Total	70	.347676	.1296332	.0154941	.316766	.378586
User_comment	Traditional Explorer	24	1.92	.504	.103	1.70	2.13
	Knowledge Map	22	2.91	.526	.112	2.68	3.14
	Web Community	24	2.25	.608	.124	1.99	2.51
	Total	70	2.34	.679	.081	2.18	2.50
Time	Traditional Explorer	24	1.1800	.88829	.18132	.8049	1.5551
	Knowledge Map	22	1.3409	1.23188	.26264	.7947	1.8871
	Web Community	24	1.0733	.84806	.17311	.7152	1.4314
	Total	70	1.1940	.98824	.11812	.9584	1.4296

Note 2.

Table 2. ANOVA

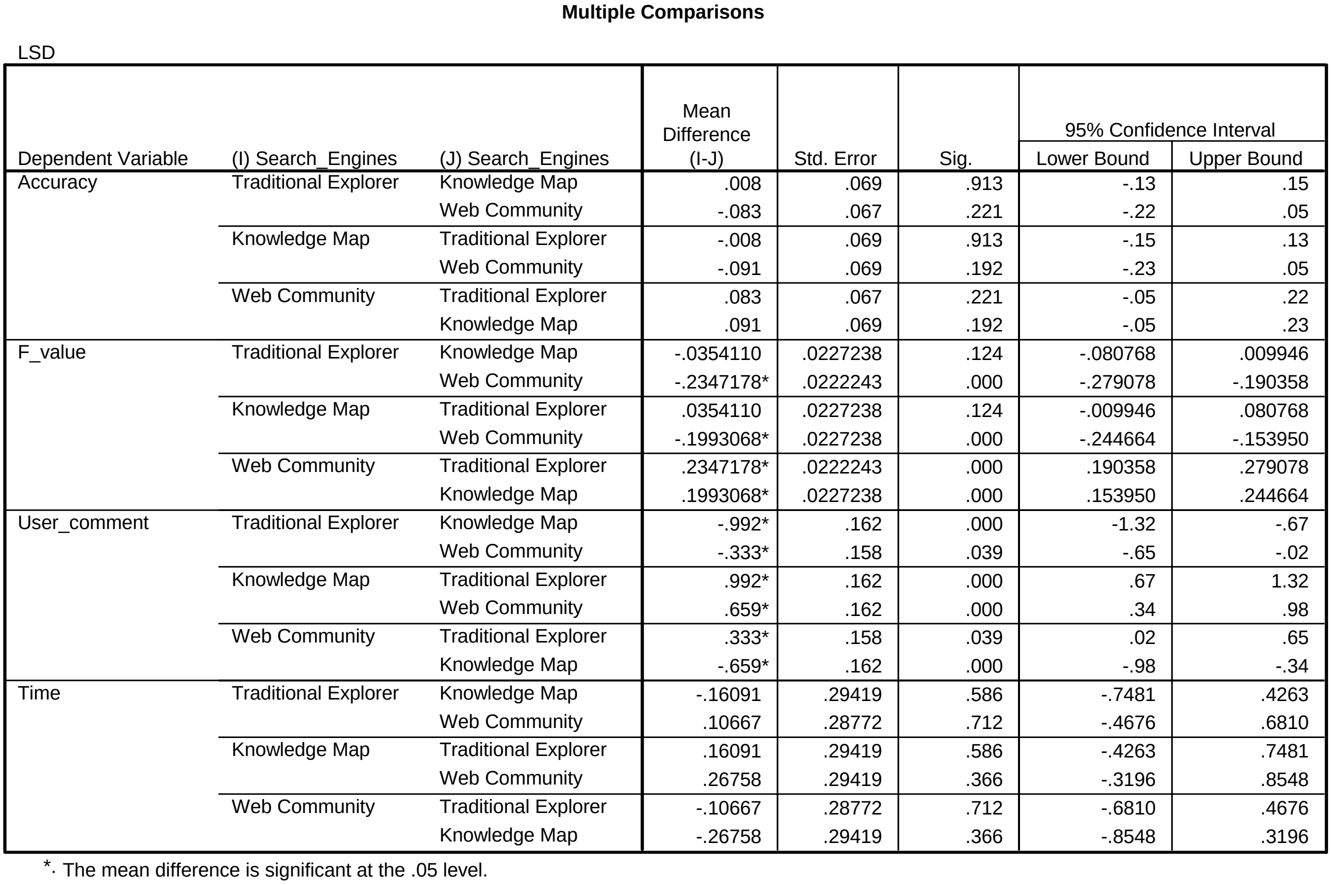


ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Accuracy	Between Groups	.120	2	.060	1.100	.339
	Within Groups	3.652	67	.055		
	Total	3.771	69			
F_value	Between Groups	.762	2	.381	64.317	.000
	Within Groups	.397	67	.006		
	Total	1.160	69			
User_comment	Between Groups	11.620	2	5.810	19.317	.000
	Within Groups	20.152	67	.301		
	Total	31.771	69			
Time	Between Groups	.829	2	.414	.417	.661
	Within Groups	66.558	67	.993		
	Total	67.387	69			

Note 3.

Table 3. Multiple Comparisons



Multiple Comparisons

LSD

Dependent Variable	(I) Search_Engines	(J) Search_Engines	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Accuracy	Traditional Explorer	Knowledge Map	.008	.069	.913	-.13	.15
		Web Community	-.083	.067	.221	-.22	.05
	Knowledge Map	Traditional Explorer	-.008	.069	.913	-.15	.13
		Web Community	-.091	.069	.192	-.23	.05
	Web Community	Traditional Explorer	.083	.067	.221	-.05	.22
		Knowledge Map	.091	.069	.192	-.05	.23
F_value	Traditional Explorer	Knowledge Map	-.0354110	.0227238	.124	-.080768	.009946
		Web Community	-.2347178*	.0222243	.000	-.279078	-.190358
	Knowledge Map	Traditional Explorer	.0354110	.0227238	.124	-.009946	.080768
		Web Community	-.1993068*	.0227238	.000	-.244664	-.153950
	Web Community	Traditional Explorer	.2347178*	.0222243	.000	.190358	.279078
		Knowledge Map	.1993068*	.0227238	.000	.153950	.244664
User_comment	Traditional Explorer	Knowledge Map	-.992*	.162	.000	-1.32	-.67
		Web Community	-.333*	.158	.039	-.65	-.02
	Knowledge Map	Traditional Explorer	.992*	.162	.000	.67	1.32
		Web Community	.659*	.162	.000	.34	.98
	Web Community	Traditional Explorer	.333*	.158	.039	.02	.65
		Knowledge Map	-.659*	.162	.000	-.98	-.34
Time	Traditional Explorer	Knowledge Map	-.16091	.29419	.586	-.7481	.4263
		Web Community	.10667	.28772	.712	-.4676	.6810
	Knowledge Map	Traditional Explorer	.16091	.29419	.586	-.4263	.7481
		Web Community	.26758	.29419	.366	-.3196	.8548
	Web Community	Traditional Explorer	-.10667	.28772	.712	-.6810	.4676
		Knowledge Map	-.26758	.29419	.366	-.8548	.3196

*. The mean difference is significant at the .05 level.