CONTROLLED EXPERIMENT ON SEARCH ENGINE KNOWLEDGE EXTRACTION CAPABILITIES

Pasquale Ardimento, Danilo Caivano, Teresa Baldassarre, Marta Cimitile and Giuseppe Visaggio

Department of Informatics, University of Bari, Via Orabona 4, Bari, Italy

Keywords: Search Engine, Knowledge Transferring, Knowledge Acquisition, Knowledge Acquisition Process.

Abstract: Continuous pressure on behalf of enterprises leads to a constant need for innovation. This involves exchanging results of knowledge and innovation among research groups and enterprises in accordance to the Open Innovation paradigm. The technologies that seem to be apparently attractive for exchanging knowledge are Internet and its search engines. Literature provides many discordant opinions on their efficacy, and to our best knowledge, no empirical evidence on the topic. This work starts from the definition of a Knowledge Acquisition Process, and presents a rigorous empirical investigation that evaluates the efficacy of the previous technologies within the Exploratory Search of Knowledge and of Relevant Knowledge according to specific requirements. The investigation has pointed out that these technologies are not effective for Explorative Search. The paper concludes with a brief analysis of other technologies to develop and analyze in order to overcome the weaknesses that this investigation has pointed out within the Knowledge Acquisition Process.

1 INTRODUCTION

In the last century Internet has represented the largest communication and knowledge trasferring media. Moreover data, information, knowledge, experiences contained in the Web increase every day. This phenomena encourages researchers and developers to study and to use all the Internet related aspects (Tonchia, 2003) (Hee-Dong Yang, 1998). In particular, research of knowledge resources through a search engine is an issue of great interest for both research and practitioner communities (Marchionini, 2006) (Gersh, 2006) (Ryen, 2006). Indeed, even if Search Engines have contributed to knowledge research and diffusion, we are aware that search engines have many limitations (Andrews, 2003) (Aswath, 2005). In this sense, (Andrews, 2003) states that 40 percent of companies rate the available search tools as “not very useful” or “only somewhat useful”; other studies emphasize that much time is needed for extracting the searched knowledge (Grandal, 2001). An explanation and description of these limits from the technological point of view are reported in (Papagelis, 2007). In this work we analyse search engine data collection, quality search, and updating of data characteristics. Another accredited analysis of this limit confirms that general queries produce a large amount of documents and that there is not a natural language interface of the search engine (Aswath, 2005). The analysis of Search Engine issues is usually characterized by alternative Search Engine solutions that overcome these issues (Papagelis, 2007), (Moldovan, 2000) (Joachims, 2007) (Al-Nazer, 2007). In (Papagelis, 2007), collaborative search engines that can be adopted between traditional search engines and web catalogues is proposed; while in (Joachims, 2007) a search engine that provides accurate training data towards learning techniques is proposed.

Moreover several new approaches in search engines are beginning to adopt intelligent techniques for improving search precision (Choi, 1998), (Zhang, 2004), (Mingxia, 2005).

Finally, we can’t avoid considering the Experience Base and Experience Factory approach that allow to store, select and search specialized Knowledge and Experience (Basili, 1994).

In this work we do not introduce our own approach to knowledge searching and transferring, which is described in previous papers of the same authors (Ardimento, 2007A), (Ardimento, 2007B). Aim of this work is to investigate the available Search Engine limitations from the user point of view in
order to extract some lessons learned and some useful suggestions for searchers and developers that are working in these areas of interest. In this sense we consider our observations of interest for knowledge searching independently from the proposed approach. In fact the new search engine approaches and tools need to start from an accurate analysis of the limitations related to existing approaches and tools. We observed that sometimes the search engine limits are mentioned but not rigorously investigated to overcome them.

Moreover, in spite of the large amount of works (Scoville, 1996), (Leighton, 1997), (Ding, 1996), (Leighton, 1996), (Chu, 1996), (Clarke, 1997) that have evaluated the efficacy of different Search Engines, to the authors’ knowledge, no replicable empirical investigations have been carried out concerning the capability of these instruments in an Exploratory Search concerning Knowledge Acquisition. For clearness, Exploratory Search is the set of activities for extracting existing knowledge and analyzing it in order to verify that its relevance allows to learn new results or technologies within a specific knowledge domain (Marchionini, 2006).

As so, this work intends carrying out an empirical investigation that answers the following research question: Are the Search Engines available on Internet effective for Exploratory Search? The investigation analyzes the cause-effect relation among use of the technologies and their effectiveness in an Exploratory Search. It is rigorously described so that other researchers can replicate it to confirm or deny the results. Replication of an empirical investigation allows overcoming contrasting opinions in literature and, at the same time, collecting a set of lessons learned on the current Search Engines.

The rest of the paper is organized as follows: the controlled investigation is described in section 2; section 3 illustrates the measurement model used; results of the study including statistical analysis are presented and discussed in section 4; finally conclusions are made in section 5.

2 CONTROLLED EXPERIMENT

In order to assure experiment replications (Baldassare, 2008), we have introduced an Exploratory Search Process (ESP) representing the sequence of actions that experimental subjects (Searcher) have to carry out. The process is shown in figure 1. The Searcher specifies a query to a search engine which expresses its need of knowledge (Step 1); the engine provides a set of Extracted Resources (ER). The Searcher selects the resources, among the Extracted Resources, that can be classified as knowledge (Step 2). This set represents the Selected Resources (SR). The Searcher further reviews the SR and selects the ones that contain the requested knowledge (RK). The RK is a more specific search question than the topic specified in the Query. The selected resources make up the Relevant Resources (RR) (step 3).

Figure 1: Exploratory Search Process.

2.1 Research Goal

According to the process described above, the research goal is formalized as follows:

Analyze the Search Engine Tools in order to evaluate them with respect to Effectiveness from the Knowledge Searcher point of view, in the context of a controlled experiment.

The following research hypotheses have been made:

H₀: The available Search Engines are effective for extracting relevant knowledge.

H₁: The available Search Engines are ineffective for extracting relevant knowledge.

The effectiveness of a Search Engine for extracting relevant knowledge in this work is defined as the ability of a Search Engine to extract useful knowledge and experience according to a defined search request.
2.2 Experiment Variables

The efficacy of a search engine is measured through a factor (Leighton, 1997) that we will call Relevance. Relevance represents the dependent variable of our study, and is defined as the evaluation of in what terms the knowledge contained in the selected resources are relevant for the searcher. Evaluation of Relevance depends from variables that represent the independent variables of our experiment. The independent variables are described as follows:

**Searcher (S):** the experimental subjects are 4 software engineering researchers with a similar experience on research projects carried out in the SERLAB laboratory of the Department of Informatics at the University of Bari, Italy. They are able to evaluate whether the knowledge contained in a selected resource following a search, is relevant from a software engineer perspective.

**Search engines (SE):** the search engines have been selected according to the data on network traffic concerning searches on Internet (http://www.onestat.com):
- SE1: Google (http://www.google.com);
- SE2: Msn (http://www.msn.com);
- SE3: Yahoo (http://www.yahoo.com);
- SE4: Altavista (http://www.altavista.com);

**Query and Query Detail Levels (Qij):** 4 different queries, each with 3 levels of detail have been selected. Qi is the i-th Query with a j-th level of detail:

**Query 1**
- Q11: <"Software Engineering Quality">
- Q12: <"Software Engineering Quality” “Process Quality”>

**Query 2**
- Q21: <"Software Engineering Process”>

**Query 3**
- Q31: <"Software Engineering Best Practices”>

**Query 4**
- Q41: <"Software Engineering Development”>
- Q42: <"Software Engineering Development” “Product Development”>

2.3 Experiment Description

The experiment was organized in 4 experimental runs, one for each Search Engine. In each run 4 Queries, with the three levels of detail, were assigned to each Searcher. Searchers used the same Search Engine.

Each run was divided into two phases: first, each searcher was assigned to the lowest level Query. Step1 of the ESP was then executed; the Search Engine produced the Extracted Resources. According to the results, the searchers carried out the selections at Step2 of the ESP, within 30 minutes. In Step2 the i-th searcher produced a set of Selected Resources SRi1. After 30 minutes the intermediate level Query was given to each searcher, the process was iterated and led to SRi2. Finally, after 30 minutes the highest level Query was given to the searchers, and the process iterated, producing SRi3. At the end of this first phase the Searcher was informed of the Required Knowledge (RK). At that point, each Searcher extracted, among the SRij, the resources containing knowledge corresponding to the RK. A set of Relevant Resources, RRij, corresponding to the SRij, were produced.

The RK, corresponding to the queries are:

**RK1:** Quality models to evaluate process performances in Software Engineering, described so that they can be transferred without help of their producers.

**RK2:** Metrics for evaluating the quality of the process models, described so that they can be transferred without help of their producers.

**RK3:** Best Practices on the Software Engineering development processes, described so that they can be transferred without help of their producers.

**RK4:** Quality metrics of Software Engineering for product development, described so that they can be transferred without help of their producers.

Each searcher used their own self-defined process for selecting the detailed resources, according to their own experience in the knowledge domain. This procedure remained tacit, in that it was out of the
scope of the investigation. Selection was carried out within a time limit of 45 minutes, in particular 20 min. for SR1, 15 min. for SR2, and 10 min. for SR3. The time available for the Searchers was less than the time estimated for evaluating the Extracted Resources and the Selected Resources. This restriction was necessary to be sure that the Searchers dedicated the same amount of time to their tasks and were not influenced by secondary effects that could have biased the results. Given the previous considerations, each RUN lasted 135 min, other than the time that each Search Engine implied for producing the Extracted Resources. This time was considered non relevant. The experimental design is reported in Table 1.

Table 1: Experimental Design.

<table>
<thead>
<tr>
<th>Experimental Subject</th>
<th>RUN1</th>
<th>RUN2</th>
<th>RUN3</th>
<th>RUN4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Searcher 1</td>
<td>SE1,Q1j</td>
<td>SE2,Q2j</td>
<td>SE3,Q3j</td>
<td>SE4,Q4j</td>
</tr>
<tr>
<td>Searcher 2</td>
<td>SE1,Q2j</td>
<td>SE2,Q4j</td>
<td>SE3,Q1j</td>
<td>SE4,Q3j</td>
</tr>
<tr>
<td>Searcher 3</td>
<td>SE1,Q3j</td>
<td>SE2,Q1j</td>
<td>SE3,Q4j</td>
<td>SE4,Q2j</td>
</tr>
<tr>
<td>Searcher 4</td>
<td>SE1,Q4j</td>
<td>SE2,Q3j</td>
<td>SE3,Q2j</td>
<td>SE4,Q1j</td>
</tr>
</tbody>
</table>

2.4 Metric Model

The research question, related to the goal of the study, that we have tried to answer is the following:

- What is the relevance of the search engines in Internet?

We will consider a search engine relevant if it allows the user to extract useful knowledge according to an assigned search scope. The search scope is assigned with refer to the queries and RK assignment. In order to answer to the proposed search question we have introduced the following metrics, which are named and described in Table 2:

3 EXPERIMENTAL RESULTS

The data collected during the investigation have been synthesized through descriptive statistics in order to represent them graphically, identify possible outliers and decide if they must be eliminated from the sample. Finally, data have been analyzed through hypothesis testing, where observations of statistical analysis were statistically validated with respect to a significance level.

3.1 Descriptive Statistics

Figure 2 reports the Line-Plot of the Mean Relevance for each Searcher, figure 3 reports the Line Plot of the Mean Relevance for each Search Engine and finally, figure 4 illustrates the Box-Plot of the Relevance distribution with respect to a level of detail of the Query. RDetail, is the distribution of Relevanceklij as k, l, and i vary.

According to the Relevance definition given in Table 2, we obtained that the Relevance value is very smaller than 100%. It means that the number of extracted resources that can be considered relevant according to a given RK is very small.

This Relevance value is conditioned by the introduced independent variables as reported in the shown graphs.

In particular, according to figure 2, it is possible to observe that there are no significant differences in Mean Relevance between the results of the Searchers, even if there are some minor differences caused by the ad hoc selection procedures carried out by each Searcher. These differences are reflected in the results.
According to figure 3 the Mean Relevance achieved by each Search Engine is small. Some non significant differences could be linked to their different navigation techniques and/or different selection algorithms.

Finally, in figure 4, box plots describe the relevance results range for each query detail level. In the box plot, median values are indicated. We can observe that there is a consistent difference in results among the three query levels of detail. In the lowest level of detail the distribution of relevance values are concentrated around 0%. Increasing the detail level, relevance assumes values around a larger range. The Relevance values result being small also in the case of a higher detail level. Given the previous considerations, Figures 2, 3, and 4 confirm small relevance values.

### 3.2 Hypothesis Tests Analysis

Relevance has been investigated to confirm the considerations pointed out by the descriptive analysis and avoid threats. For this reason, two types of tests have been carried out:

- **Kruskal-Wallis**, a non-parametric test alternative to one-way (between-groups) ANOVA. It is used to compare three or more samples, and it tests the null hypothesis that the different samples in the comparison were drawn from the same distribution or from distributions with the same median. Thus, the interpretation of the Kruskal-Wallis test is basically similar to that of the parametric one-way ANOVA, except that it is based on ranks rather than means (Siegel and Castellan, 1988). Note that all the Kruskall Wallis tests have been carried out on mean measures of Relevance with respect to the different detail levels ($\frac{\sum\text{Relevance}_{klij}}{3}$), because the different detail levels have not been considered in this test, rather they have been analyzed through a separate one.

- **Friedman ANOVA**: it is a non parametric alternative to one-way repeated measures analysis of variance. In particular in the context of our analysis it is used to investigate presences of statistically significant differences in the values of Relevance collected with respect to the 3 different levels of Query details. This test assumes that the variables (levels) under consideration be measured on at least an ordinal (rank order) scale. The null hypothesis for the procedure is that the Relevance for the different levels of detail, contain samples drawn from the same population, or specifically, populations with identical medians.
3.3 Results

Tables 3, 4, 5, and 6 report test values for Relevance. According to the descriptive analysis results, Kruskal-Wallis points out that no statistically significant differences exist between Search Engine, Queries and Searcher. The significance values are the following: $p = 0.2695$; $p = 0.7924$; $p = 0.1064$.

It can be seen in table 3 that the number of measures considered for each Search Engine is 4, i.e. 1 for each Searcher that answered a Query using a SE only once, for a total of 16 measures.

For clearness, each of the 16 values has been associated to a Rank that corresponds to its position in an increasing order. In case of $n$ equal values in positions $p_1, p_2, ..., p_n$, the assigned rank is $(p_1+p_2+...+p_n)/n$. The Sum of Ranks corresponds to the sum of the ranks related to the 4 points of the Search Engine.

The Sum of Ranks is displayed in the rightmost column of the spreadsheet. The Kruskal-Wallis test isn’t significant ($p = 0.2695$). Thus, we can conclude that the Search Engines were not significantly different from each other with respect to Mean of Relevance.

So, we can conclude that Search Engines give different responses to the same queries although their differences in terms of knowledge relevance are not statistically significant.

In table 4, the number of measures considered for each Query is 4, corresponding to the 4 Searchers that have answered a query once and used a different Search Engine, for a total of 16 measures. The Kruskal-Wallis test isn’t significant ($p = 0.7924$). Thus, we can conclude that if we consider the Mean Relevance, there aren’t significant differences among different Searchers.

So, we can conclude that searchers have analogous experiences in the search knowledge domain, although different procedures are used for selecting relevant knowledge, the differences in results are not statistically significant.

Table 3: Dependence of Relevance from the Search Engine.

<table>
<thead>
<tr>
<th>Measures</th>
<th>Sum of Ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google 4</td>
<td>49,000000</td>
</tr>
<tr>
<td>MSN 4</td>
<td>35,000000</td>
</tr>
<tr>
<td>Yahoo 4</td>
<td>27,000000</td>
</tr>
<tr>
<td>Altavista 4</td>
<td>25,000000</td>
</tr>
</tbody>
</table>

Table 4: Dependence of Relevance from the type of Query.

<table>
<thead>
<tr>
<th>Measures</th>
<th>Sum of Ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1 4</td>
<td>32,000000</td>
</tr>
<tr>
<td>Q2 4</td>
<td>29,000000</td>
</tr>
<tr>
<td>Q3 4</td>
<td>33,000000</td>
</tr>
<tr>
<td>Q4 4</td>
<td>42,000000</td>
</tr>
</tbody>
</table>

In Table 5, the number of measures considered with respect to each Searcher is 4, corresponding to the results that each Searcher obtained in each of the 4 Runs, for a total of 16 measures. The Kruskal-Wallis test isn’t significant ($p = 0.1472$). Thus, we can conclude that if we consider the Mean of Relevance, there aren’t significant differences among different Searchers.

So, we can conclude that searchers have analogous experiences in the search knowledge domain, although different procedures are used for selecting relevant knowledge, the differences in results are not statistically significant.

Table 5: Dependence of Relevance from the Searcher.

<table>
<thead>
<tr>
<th>Measures</th>
<th>Sum of Ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Searcher1 4</td>
<td>49,000000</td>
</tr>
<tr>
<td>Searcher2 4</td>
<td>22,000000</td>
</tr>
<tr>
<td>Searcher3 4</td>
<td>42,000000</td>
</tr>
<tr>
<td>Searcher4 4</td>
<td>23,000000</td>
</tr>
</tbody>
</table>

Finally, Table 6 reports the value of Sum of Ranks, the average Relevance, the mean of rank order correlation between the cases and the standard deviation for each of the levels of detail of the Relevance Query. For clearness, the average is intended as the average value of the ranks calculated for each of the sample data points.

The first observation on these results is that the Average and the Sum of Ranks increases as the level of detail of the query increases. This confirms that the level of detail of the question allows for a greater relevance of the Search Engine. Also, the Friedman Anova test shows that there are highly significant differences ($p < 0.00176$) between the different Detail Levels. This difference is statistically significant.

This confirms that for relevance, a statistically significant difference exists among the results obtained with queries of different detail levels. So, a higher level of detail in the queries increases effectiveness in the relevance of the resources extracted from the Search Engine.
Table 6: Dependence of Relevance from the levels of detail of the Query.

<table>
<thead>
<tr>
<th></th>
<th>Average</th>
<th>Sum of Ranks</th>
<th>Mean</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>1,437500</td>
<td>23,00000</td>
<td>0,000390</td>
<td>0,001117</td>
</tr>
<tr>
<td>R2</td>
<td>1,906250</td>
<td>30,50000</td>
<td>0,030081</td>
<td>0,055850</td>
</tr>
<tr>
<td>R3</td>
<td>2,656250</td>
<td>42,50000</td>
<td>0,219395</td>
<td>0,215998</td>
</tr>
</tbody>
</table>

4 CONCLUSIONS AND FUTURE WORK

The experiment carried out has allowed us to give a preliminary answer to the research question. We can conclude that: the available Internet Search Engines are not relevant. Their capabilities in extracting relevant knowledge according to an assigned search goal are very low and they can’t be used to extract reusable innovative knowledge to transfer between research organizations or enterprises.

These results are independent both from the Search Engines, from the Searchers and the searching queries. Moreover, we identified a relationship between Search Engine Relevance and detail levels of the searching Queries. As the detail level increases, the Search Engine appears more relevant, although, in all these cases, results are not satisfactory.

Given these considerations, the only difference that can be used to improve the rate of resources containing knowledge and relevant knowledge is the Query Detail Level. Note that the level of detail is not managed in the same way by all search engines, and usually depend by their parsers. So, a greater level of detail in the Query, may not necessarily assure that the Search Engine is able to satisfy the knowledge content the Searcher is interested in. Also, consider that the results are not satisfactory what ever the level of detail of the Query.

The proposed work has empirically shown some shared opinions about the low quality of the knowledge available using the internet search engines. These considerations need to be validated through the replication of the experimentation and furthermore through a family of investigations. The family of experiments will allow us to obtain a rigorous list of search engine limits from the user point of view. These lists will be used to characterize and define a new and innovative approach to knowledge searching processes and tools.

The experiment also has suggested to investigate the proposed ESP. The ESP could be encapsulated and refined in a knowledge transferring and searching approach.

REFERENCES


