

STUDENT'S PERCEPTION OF A SEMANTIC SEARCH ENGINE

Monique Reichert

*Faculty of Language and Literature, Humanities, Arts and Education, University of Luxembourg
Campus Walferdange, BP 2, L-7201 Walferdange
monique.reichert@emac.s.lu*

Serge Linckels, Christoph Meinel

*Hasso-Plattner-Institut for Software Systems Engineering (HPI), University of Potsdam
D-14440 Potsdam, Germany
{linckels, Meinel}@hpi.uni-potsdam.de*

Thomas Engel

*Luxembourg International Advanced Studies in Information Technologies (LIASIT), University of Luxembourg
6, r. Richard Coudenhove-Kalergi, L-1359 Luxembourg
thomas.engel@uni.lu*

ABSTRACT

This paper focuses on the question how students appreciate and value the possibility to query a multimedia knowledge base by entering complete questions instead of keywords. We aim at examining how far the result(s) of semantic queries are more appealing to students than those from a keyword-based search engine. In order to investigate this issue, two versions of the e-learning tool CHESt – an ontology-driven interactive expert system focusing on computer history – were tested in a secondary school. While the first version of CHESt implements a simple keyword search, the second version of CHESt carries out a semantic search. The aim was to assess whether students are satisfied with the number and the pertinence of the search results, and whether they generally appreciate the option to ‘communicate’ with the system by asking complete questions in natural language.

The outcome of our investigation shows that students generally preferred to use the keyword instead of the semantic search function, independently from the judgment on the accuracy of the results yielded by the respective search engine. The results suggest that the pertinence of the results as judged by the students strongly depends on the familiarity of the users with both the formulation of questions and the domain of interest. Also the semantic search engine needs to be improved in order to extract more semantic information.

KEYWORDS

Semantic search engine, information retrieval, multimedia knowledge base, educational tool.

1. INTRODUCTION

CHESt (*Computer History Expert System*) is an ontology-driven interactive e-learning tool that understands students' questions. It automatically retrieves only pertinent documents from an archive of educational material. CHESt disposes of a multimedia knowledge base, which is composed of 300 short multimedia sequences (*clips*). Each clip is between one and four minutes long and was recorded with *tele-TASK* (<http://www.tele-task.de>). The clips can be accessed from home or from a classroom either by streaming via the Internet or locally (for example from a CD-ROM or a LAN).

The architecture of CHESt is composed of four layers: the knowledge layer that includes the knowledge base and the semantic repositories, the inference layer that includes the search engine, the communication layer that guarantees a transparent communication between the user and the search engine, and the presentation layer that is responsible for the interaction with the user. Except for the presentation layer, all other layers are platform-independent. Furthermore, each layer can be located somewhere on the Internet or

on the local machine. The user question in natural language is transmitted to the inference layer, which tries to find the best matching clip(s). The URI (*Uniform Resource Identifier*) of the resulting clips are returned as an XML file to the presentation layer. Then, the user can select the clip(s) (s)he wants to watch.



Figure 1. CHESt with a semantic search and the question: "Who invented the transistor?"

We conceived two versions of CHESt, each using a different retrieval strategy; a keyword search for the first version, and a semantic search for the second version. Both versions are identical except for the inference layer. The aim of the retrieval mechanism of version two (semantic search) is to find fewer but more pertinent results by performing a semantic search rather than a keyword search. For example the question "Who invented the transistor" would return different results, whether a keyword search or a semantic search was applied. As for the first search engine, all clips that contain one of the keywords are potential results. But, as for the second search engine, only clips that are about the inventors of the transistor are potential results (figure 1). We developed an application and a Web user interface for CHESt. Details of our semantic search engine can be found in Linckels, S. & Meinel, Ch. (2005).

2. TESTING CHEST IN SCHOOLS

Given our premise that the development of an e-learning tool such as CHESt is not an end in itself but has to be proven useful in a target setting, we asked students to test and to judge the two versions of the e-learning tool. In the following we are going to report on these evaluations of CHESt that we carried out at the *Lycée Technique d'Esch/Alzette*, a technical school in Luxembourg/Europe, at the beginning of the year 2005.

2.1 Test Preliminaries

Students from the upper secondary school level were asked to try out both versions of CHESt and to provide feedback on the three main characteristics of the tool: the number of results, the pertinence of the results and the satisfaction with the possibility to enter complete question(s) in natural language instead of keywords. Three consecutive assessment sessions took place, which differed from each other only in concern of two variables, with one variable being revised from session one to session two, and another variable from session two to session three (table 1; additional details about the variables will be provided further below).

Table 1. Settings of both variables for the three assessment sessions

	Type of frame questions	Instructions given about how to use the search engines
Session 1	Precise questions, e. g. "Who invented the Z3-machine?"	Students were instructed to enter single or multiple words, the way they thought they would obtain the most pertinent results.
Session 2	General questions, e. g. "Describe the early years of the Internet."	Idem session 1
Session 3	Idem session 2	Students were told to enter questions while using the semantic search, and keywords while using the keyword search engine.

2.2 General Characteristics of the Three Sessions, Instructions and Procedure

For each of the three assessments, a different group of students was to try out both versions of CHESt. None of the subjects had further domain knowledge about computer history. One half of each group started with the keyword search, the other half with the semantic search. After 20 minutes, the students were asked their opinion about the just tested CHESt version on a number of questions, and then continued within a second trial – again lasting 20 minutes – with the respective other version of the search engine. In order to provide the subjects some general context within which they could search for information, three questions (in the following named "frame questions") were presented at the beginning of each trial (i.e., six questions per task for each student; see further below for some examples of frame questions).

At the beginning of each session, the students were informed that two search engines allowing searching for information from the domain of computer history would be presented to them. They were told that not their successful answering to the frame questions would be the aim of the session, but rather their personal judging of the efficiency and their general liking of the respective search engine. They were also briefed that the graphical user interface (GUI) would be the same for both versions, and that no questions would have the GUI as target. The students were informed that their main job would consist in judging whether the search results yielded by the respective search engine would match their queries, and whether they really found the information they had been looking for. After the respective version of CHESt had been tested, the students answered questions focusing on the following issues:

- *Did the just tested search engine yield too few, too many, or an adequate number of results?* This question aimed at clarifying the personal judgment concerning the quantity of the results. Students might actually find what they searched for, but they might have expected more results.
- *Did the search results exactly fit the queries?* This question aimed at knowing whether, in general, the subjects had the impression that the result(s) listed was/were pertinent in regard to the keywords or questions they entered within the search field
- *Did the subjects find the information they have been searching for?* This question is considered separately from others about the general fitting of the results, as the user might have found results that fitted the queries well, but still might have been unable to find what (s)he actually had been looking for.

After both versions of CHESt had been tested by the subjects, they were questioned on the following issues:

- Some of the questions asked for a comparison between both versions, aiming at finding out whether the participants had the impression that the one or other version would provide the more fitting results.
- The users were also asked which version they would choose if they had an exam within the domain of computer history during which they were allowed to use one of the CHESt versions. This was to find out about the general preference for one of the two versions within a concrete context.
- Finally, one question raised the issue about the students' opinion about having the possibility of asking complete questions instead of keywords.

2.3 First Session

18 male students from the 13th (terminal) grade of secondary school (technical formation; mean age 21.25) participated within this first evaluative assessment. No information was provided about the difference between the two search engines; students were instructed to enter single or multiple words, even complete questions, just the way they thought they would obtain the most pertinent results. Further, the participants received precise frame questions such as the following ones:

- Did Howard Aiken and Charles Babbage know each other?
- Find three interesting inventions from the software domain (e.g. operating system, programming language, application). Which company or person has invented this and when was it published?

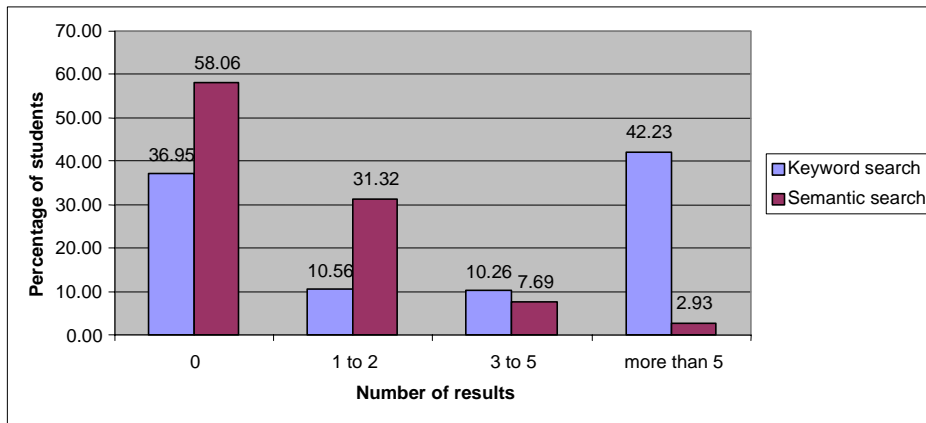


Figure 2. Number of results per CHES_t version: percentage of total number of results

Asked about the number of the yielded results, the majority of the students think there is either an adequate number of results (seven students) or even too many results (seven students) generated by the keyword search. Meanwhile, considerable 14 out of 18 students asserted that the semantic search function yielded too few results. The real number of results generated by the respective search engines (figure 2) confirms that a higher percentage of queries within the semantic search than within the keyword search yielded no results. In the meantime, the keyword search led to more than five results in 42% of the search initials.

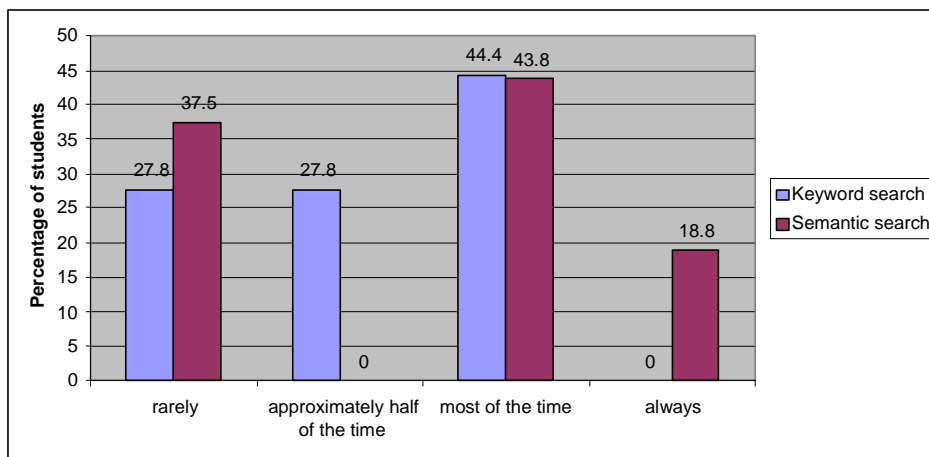


Figure 3. Within the listed results I found what I have been looking for

The question about the pertinence of the yielded results revealed an obvious superiority of the semantic search function. While 78% of the subjects said that in only a few or in approximately half of the cases the

keyword search would have provided fitting results, 78% considered that “most of the results” (61.1%) or “all of the results” (17%) fitted the search subjects within the semantic search engine.

There was not a similar obvious difference between the two search engines concerning the question whether students found what they were looking for. The subjects judged the efficiency of both search engines quite similarly (figure 3). Why is there this small discrepancy between the fitting of the results and the success in finding what was looked for? If taking a look at those last two questions in combination, one can observe that the incongruity is especially due to only a few of the students. Actually, for the keyword search, only one student meant that, although all of the results fitted the keywords, he only found what he looked for in approximately half of the search initials. Concerning the semantic search, three students (out of 16) said that most or all of the results fitted their search subjects, but that they still only rarely found what they had been looking for.

Asking students which one of the search engines they would prefer if they had an exam on the subject of computer history, the answers given were not clearly pointing in one or another direction (table 2). Although most of the students were fairly satisfied with the fitting of the results from the semantic search engine, and although there was no greater difference between the keyword search and the semantic search in terms of finding what has been searched for, only 39% would choose the semantic search engine, compared to 28% announcing their preference for the keyword search, and 22% not deciding on any of both versions.

Finally, when asked about their liking of the possibility of entering whole questions instead of single keywords, half of the students (N=9; 50%) indicated that this possibility is only considered to be advantageous if it also yields better results than a keyword search.

Table 2. What version did the users prefer? Choice of the version 1=keyword search, 2=semantic search, 3=both versions equivalent, 4=none of the versions

		Frequency	Percent	Cumulative Percent
Valid	1	5	27.8	27.8
	1 and 4	1	5.6	5.6
	2	7	38.9	38.9
	2 and 4	1	5.6	77.8
	3	2	11.1	88.9
	4	2	11.1	100.0
Total		18	100.0	

To summarize, the first evaluation session revealed that, although most of the subjects were rather satisfied with the fitting of the results provided by the semantic search engine, they were not completely convinced of the (possible) advantages of the semantic version of CHESt.

Before discussing these results in greater detail, the realization and results of the two other evaluative sessions shall be described. The principal aim of the subsequent session was to replicate the results of the first session with more general frame questions. Indeed, the analysis of the keywords and sentences entered showed that most of the students were sticking all to strictly to the respective frame questions in their formulation of the questions and keywords in the question bar. In order to investigate whether similar results would be obtained when the students are given greater liberty in their searching for information on the subject of computer history, more general tasks were formulated for the second and the third evaluation session, described in the following sections.

2.4 Second Session

18 students (17 male) from the 12th grade of secondary school (technical formation; mean age 19.76 years) participated within this second evaluative session. This time, the frame questions were more general than in the previous session; examples of frame questions are as follows:

- Give an overview of the last 60 years of computer evolution.
- Explain why, especially around World War II, computers had been developed. Name three examples of such computers and their respective inventor(s).

In correspondence to the results from the first assessment, the students in this second session showed to be more satisfied with the number of the results listed by the keyword search engine. 56% asserted that an adequate number of results were listed by the keyword search engine, only 17% said it had been too few.

This can be contrasted with respectable 78% of the students thinking the semantic search had generated too few results. The mean number of results yielded by the semantic search engine is way below the one in the first session, with no results yielded at all in considerable 80.5% of the search initials (figure 4). The keyword search in the meantime yielded no results in only 28% of the searches, while half of the search initials (50%) led to more than five results.

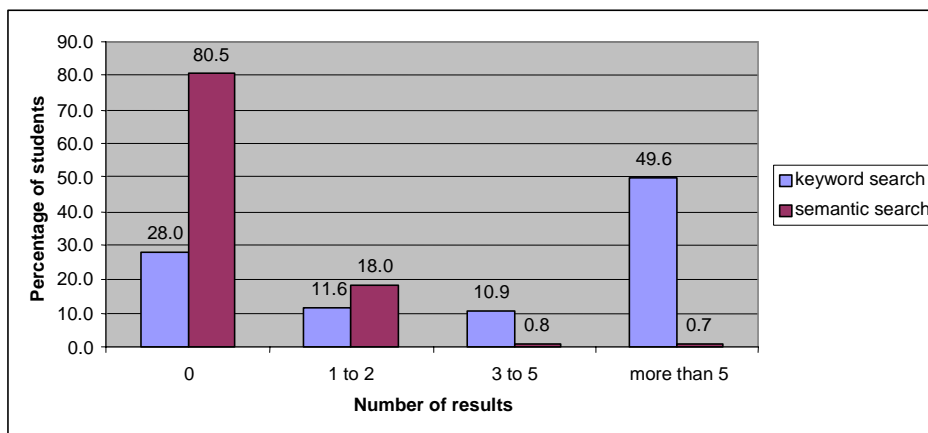


Figure 4. Number of results per CHESt version: percentage of total number of results

This time the question about the pertinence of the yielded results revealed a different answer pattern than within the first testing session. Actually, the keyword search was judged a little bit more positive than in the previous testing, while the semantic search engine was given a worse evaluation. This led to the results pattern of a nearly equivalent number of students (nine and 10, for the keyword and semantic search, respectively) asserting that “most of the results” or “all of the results” fitted the search keywords. Two students asserted not having had any fitting results at all during the semantic search (while none of the users did so concerning the keyword search).

As far as the question of whether the students were able to find what they were looking for is concerned, a rather negative picture is being presented for the semantic search. Three people never found what they were looking for, and four users were only rarely able to locate some interesting information. This is the case “only” two times within the keyword search. Clearly most of the students (76%) confirmed being successful in finding the information they had been looking for with the keyword search, compared to 47% concerning the semantic search engine.

If comparing and combining the answers to these last two questions (fitting of the results and finding what one has been looking for), we noticed that only for very few of the students there was an incongruity between the judgment on the general fitting of the search results to the entered queries on the one hand, and the judgment about whether the searched information was found on the other hand. For the keyword search one student (out of 17) believed that, although most of the results fitted the keywords, (s)he rarely found what (s)he had been looking for. Two students asserted that, although only few of the keywords fitted the results, they still found what they had looked for in most of the search initials. In concern of the semantic search, a similar pattern of the results is obtained, with two students saying that most or all of the results fitted their search words, although they only rarely found what they had looked for. One user indicated that although only few of the keywords fitted the results, (s)he still found what (s)he had looked for in most of the search initials.

Table 3. What version did the users prefer? Choice of the version 1=keyword search, 2=semantic search, 3=both versions equivalent, 4=none of the versions (more answers permitted)

		Frequency	Percent	Cumulative Percent
Valid	1	17	94.4	94.4
	2 and 4	1	5.6	100.0
Total		18	100.0	

Even though the previous questions already revealed that users did not have such an obvious judgment in the one or other direction about one of the CHESt versions any more, asking them which version they would

prefer to use during an exam in the domain of computer history revealed an obvious preference of nearly all of the students (17 out of 18) for the keyword search (table 3).

Approximately half of the students (10 out of 18) finally stated that the possibility to enter whole questions instead of single keywords would be a good option only if this would lead to better results than with keywords.

In summary, the second evaluative session – using more general frame questions and thus providing the participants a greater liberty in their information search – revealed a slightly more negative judgment of the semantic search engine than the first session. Firstly, the students indicated being more satisfied with the number of results provided by the keyword search than those by the semantic search. Secondly, although the judgment about the fitting of the listed results was comparable for both CHESt versions, more users pointed out finding what they were searching for within the keyword search rather than with the semantic search. The clear preference for the keyword search finally underlines that the characteristics of the keyword version of CHESt are the more appealing ones.

One final point still remains to be pointed out in this context: actually, the analysis of the log-file of this session reveals that students did not refer very often to the possibility of combining keywords and/or even entering whole questions, which would ensure optimal “communication” with the semantic version of CHESt. We hypothesized that the high number of semantic search initials yielding no results at all is due to this problem. In order to investigate whether the rather negative results concerning the semantic version of CHESt are the result of whether users do or do not fully exploit the potential of CHESt, a third evaluative session took place in which the explicit instruction was given to enter complete questions in the search field when using the semantic version of CHESt.

2.5 Third Session

14 students (11 male) from 12th grade of secondary school (general technical education; 17 to 20 years old) participated in this third assessment. While the frame questions remained rather general (as within the second session), the students were, this time, explicitly told to enter complete questions while using the semantic search, and to enter single or multiple words while using the keyword search. The judgments given by the students during this third session were still comparable to the ones provided in the second assessment, as outlined in the following section.

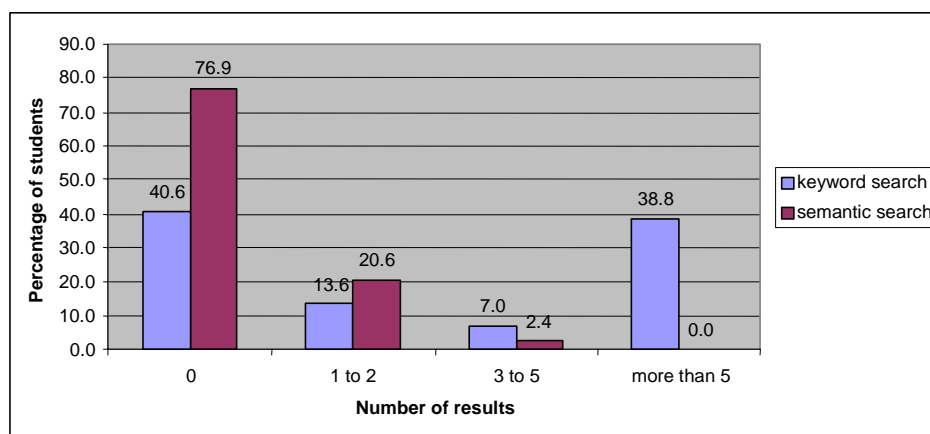


Figure 5. Number of results per CHESt version: percentage of total number of results

Comparable to the results from the previous assessment, the students were more satisfied with the number of results listed by the keyword search than with the number generated by the semantic search. While 64% asserted that an adequate number of results were listed by the keyword search engine, only 21% said so with relation to the semantic search engine. 79% had the opinion that the semantic search listed too few. The analysis of the mean number of results yielded by the respective queries further consolidates the finding out of a very elevated percentage of semantic search initials not having generated any results at all (77%) during the previous sessions (figure 5).

The overall satisfaction with the fitting of the results listed by the semantic search engine was even lower than in the previous assessments. While 57.1% affirmed that most of the results fitted the words they entered when using the keyword search, only 33% (25% saying “most of them”, 8% saying “all”) said so as to the results yielded by the semantic search.

The same is true for the success the users experienced concerning the finding of what they were looking for: with the keyword search, eight people (out of 13) asserted having found what they were looking for “most of the time” or “always”, but only two people were equally confident concerning the semantic search engine. No greater discrepancies were found as to these two questions, with only one student asserting that, although only few of the results fitted the queries, (s)he still found what (s)he had looked for most of the time (both CHESt versions).

Asking students which version they would prefer to use during an exam in the domain of computer history revealed an obvious voting of the great majority of the students (11 out of 14) for the keyword search. Finally, 79% of the users emphasized that they would like the option of being able to ask complete questions instead of keywords only if this yielded the better results.

In summary, this last assessment session left us with a rather devastating image of the semantic search engine. Firstly, students were again not as satisfied with the number of results provided by the semantic search as those listed by the keyword search. Also, even when giving concrete instructions as to how to use the search engines in order to guarantee that all qualities of the semantic engine are exploited, a very high percentage of the queries still remain without any yield of results. In addition, even few students than in the previous sessions really seemed to be satisfied by the pertinence of the listed results. The same is true for the judgment of the success in finding what the users were looking for. This third assessment session thus confirms and consolidates the finding that several of the qualities of CHESt have to be revised in greater detail.

3. DISCUSSION AND CONCLUSION

The present investigation aimed at evaluating the qualities of the keyword and the semantic version of the e-learning tool CHESt in an educational environment. Students from the upper secondary school level were asked to test both versions and judge them on the number and the pertinence of the yielded results, as well as to give their opinion about the possibility to enter questions instead of keywords. The results from the three evaluative sessions especially revealed two things in particular:

The **first striking result** was that the subjects generally preferred the keyword search to the semantic search. This was found to be independent from the judgments on the appropriateness of the yielded results. Even in the first session, where the majority of the users claimed that the semantic version yielded pertinent results, compared to only a few concerning the keyword search, most of the users decided to use the keyword version of CHESt within an exam on computer history.

There are multiple reasons that could stand as explanation for this result. Actually, the students could just be more accustomed to use keyword search engines (nearly all search engines on the Web are keyword based). Further, entering keywords might have been experienced as an easier and more comfortable task than entering whole sentences. Blondel, F.-M. (2001) concludes in his survey that the students' preferred method to query search engines is without any doubt by keywords. In the light of the generally valid claim of the users that they like the option of being able to ask complete questions instead of keywords, but only if this would yield better results, we however realized that the supplementary intellectual task of thinking and formulating whole questions must not necessarily be considered as a real burden compared to the entering of single, maybe only loosely related keywords. Finally, the number of results generated by the respective search engine is considered to be a factor of central importance concerning the rating of the CHESt-versions. Actually it turned out that, more than half of the semantic search initials did not lead to any result at all. By contrast to this, the keyword search generally listed a high number of results (judged as being “too many”, or – more often – as being “neither too many, nor too few”). Seeing the differences within the number of results that were in the mean yielded by a keyword and a semantic search respectively, the users might have experienced to have many more opportunities to explore the content of the knowledge base with the keyword than with the semantic search. The number of the yielded results is seen as a characteristic that might be of central importance, especially if users of an e-learning tool have not yet an elaborate knowledge within the

domain of interest (such as computer history). The fact that during the last two sessions users more often found what they had searched for with the keyword than with the semantic search engine, further underlines the advantageous possibilities the students might have experienced with the higher number of results produced by the keyword search.

One **second striking finding** concerns the pertinence of the results. During the first evaluative session, the subjects were more confident in this concern while using the semantic search function. As opposed to this, subjects judged the keyword search more positively in this regard during the other two sessions, where more general frame questions had been given. Thus, providing the users with additional freedom to explore the knowledge base has led to search queries that provided less convenient results than queries that were initiated within a more restrictive context. We refer to three explanations concerning this pattern of results:

First of all, this finding might be explained by the fact that users directly associated this question with the one about the number of the results: as a very high percentage of the search initials didn't yield any results at all during the semantic information queries, the judgment on the pertinence of the results might have been strongly influenced by this (with an interpretation such as "no results at all means no fitting").

Secondly, the difference in this matter between the first and the other two sessions might suggest that students have had general difficulties to formulate own questions in order to explore a domain such as computer history. This might have resulted in a general sticking to the rather specific frame questions during the first session. Indeed, it was possible to take exact questions/words that were given as frame questions (or to slightly change the formulation of those) and to enter these within the search field. Such a strategy was not possible any more to the same degree during the two subsequent sessions, where users were expected to think about and formulate questions more autonomously. Thirdly, this finding of a difference between the first and the other two sessions however might also suggest that we have to improve the semantic search engine concerning its main characteristic: the understanding of the users' questions. As already outlined above, although many of the yielded results were judged to fit the search subjects, just too many queries have yielded no results at all. Given the fact that both search engines access the same knowledge base, and that the judgment about the pertinence of the results did not differ significantly between sessions two and three (where emphasis was placed on the instruction to enter whole questions) this finding underlines that future efforts will need to focus on the improvement of the inference engine. It seems that the semantic search engine is weak in inferring over more general questions. Current research within our laboratory aims to improve the linguistic pre-processing of the user's question in order to extract more semantic information.

In conclusion, the results of our assessments suggest that the satisfaction users experience with a search engine like CHESt (in its two versions) seems to strongly depend on three factors:

- The practice users have with the respective search engine (with the formulation of questions). We agree with Fidel, R. et.al. (1999), Blondel, F.-M. (2001), and Navarro-Prieto, R. et al (1999) that users need guidance in how to formulate effective queries, especially if they are not expert in the focused domain
- The background knowledge users have concerning of the focused domain (here: computer history); little knowledge within a domain of interest seems to require good basic opportunities (such as a list of domains to be explored or a search tree) to explore a given knowledge base;
- The factual success of a search engine to find the requested content – which, again, might depend on the content of the knowledge base.

REFERENCES

- Blondel, F.-M., 2001. La Recherche d'Informations sur Internet par des Lycéens, Analyse et Assistance à l'Apprentissage. Proceedings *5e Colloque Hypermédias et Apprentissages*, Grenoble, France, pp. 119-133.
- Fidel, R. et.al., 1999. A Visit to the Information Mall: Web Searching Behavior of High School Students. In *Journal of the American Society for Information Science*, vol. 50-1, pp. 24-37.
- Hölscher, C. and Strube, G., 2000. Web Search Behavior of Internet Experts and Newbies. Proceedings *9th International World Wide Web Conference (WWW-9)*, Amsterdam, Netherlands, pp.337-346.
- Linckels, S. and Meinel, Ch., 2005. A Simple Solution for an Intelligent Librarian System. Proceedings *IADIS International Conference of Applied Computing (AC2005)*, Lisbon, Portugal, 2005; vol. I, pp. 495-503.
- Navarro-Prieto, R. et.al, 1999. Cognitive Strategies in Web Searching. Proceedings *5th Conference on Human Factors & the Web*, Gaithersburg, USA.