WHICH CONTRIBUTION OF THE WEB SERVICES IN THE IMPROVEMENT OF WEB SEARCHING?
A BEHAVIOURAL STUDY OF THE NET SURFERS

Christian Belbèze and Chantal Soulé-Dupuy

IRIT, Université Toulouse 1, 2 Rue du Doyen-Gabriel-Marty, 31042 Toulouse Cedex 9, France

Keywords: Information access, Web searching, Internet surfer behaviour, behavioural study, Web Services.

Abstract: The difficulty of finding information on the Web grows and this even for the most experts of us. In order to better understand how the Net surfers search, we observed five adults and four children. Two different protocols of observation, both presenting imposed and free searches, were defined for the children and the adults. Thus, we have defined a certain number of behaviours, attitudes and difficulties, and then we classified them. The result of these observations as well as the analysis of the behaviours are presented in this paper. The main goal of this paper is to introduce the contribution of Web services in the improvement of information access and Web searching.

1 INTRODUCTION

While the World Wide Web contents offer more and more each day, finding an answer or a piece of information becomes increasingly difficult. Before thinking about modifying the solutions of research suggested to the Net surfers, it appeared significant to observe them. The search engines are classically tools not guiding the Net surfer. According to his knowledge of search engines mechanisms and of his own capabilities (at the same time on the required subject and on the information format), Net surfers provide different reference frames of keywords. In a query without result, for example, an experienced Net surfer who is speaking English will quickly widen his research by the use of keywords in this language. But this extreme case is rare in a general public context. The search engines accept any reference frame of keywords and generally turn over thousands or million sites and it is of course there that the difficulties for the Net surfers start. Which adequacies and which relevance represent these proposals? How do Net surfers decode the few lines proposed to determine the contents of the page? How do they read then these pages once they selected them? To answer these questions, to imagine the Web of tomorrow and its information retrieval systems, we selected nine people from eight to sixty twelve years old, including five female,

from different sociocultural backgrounds and we observed them.

According to these observations, our goal is to propose some solutions to improve information access. Thus, our work aims at considering the contribution of Web services to help Net surfers in Web searching.

In this paper, we provide some related work. Then, we report the experimentation and the results of the observations. And in the last section we present the architecture that we propose to help Net surfers in Web searching by using selected services.

2 RELATED WORK

Since the beginnings of the documentary systems, and in particular since the work undertaken by Gerald Salton (Salton, 1970) (Salton & McGill, 1983), information retrieval enormously evolved. The evolutions were not carried on the models of information retrieval themselves, but rather on the contexts of use, the Net having upset the practices and the needs as regards access to information.

Many studies have lead to the same conclusion that information retrieval systems (more particularly on the Web) provide too disappointing results (Jansen & Al., 2000) (Savoy & Picard, 2001), (NachMias & Gilad, 2002), (Laurie, 2005).

Traditional information retrieval models prove to be ineffective because of the growing mass of information. Even the valuation campaigns, such as TREC (http://trec.nist.gov/), CLEF (www.clef-campaign.org), show the limits of experimentations feasibilities. Improvements might then come from new architectures including new technologies such as Web services and from researches led in the context of the Semantic Web: http://www.w3.org.

3 THE EXPERIMENTATION

This section describes an observation of nine users searching the web: All measurements, details and videos are present on the site http://sissiprojet.free.fr.

3.1 Protocols

Two protocols have been defined, one for the five adults and one for the four children. Both protocols mix imposed research subjects and free research subjects. A research is a set of reading tasks and a set of search requests using search engines. A new request will be considered to each search engine search button click.

3.1.1 Adults Protocols

The adult protocol is constituted of five researches:
- Song lyrics "All blues";
- The distance between two French town;
- A music score (Romania hymn);
- Two free research subjects.

3.1.2 Children Protocol

This protocol is constituted of two researches:
- The dolphin (life expectancy);
- One free research subject.

3.1.3 Notes and Measurements

The nine users give four notes (from 0 to 10), at the end of each research.
- Capacity to understand the met information;
- Interest of the visited site;
- Subjective Time length;
- General feeling.

For each research we measure a list of indicators
- Subjective Time length;
- General feeling.
- Research duration;
- Number of visited sites;

- Number of supplementary read pages;
- Success or failure of the research;
- Number of requests (number of times the search engine has been requested).

And for each request we recorded
- Used keywords;
- Number of keywords;
- Number of found sites;
- Whether the request multi-lingual or not.

3.2 The Users

Each user introduced himself/herself completely on the site http://sissiprojet.free.fr. We try to choose users with different ages, genders and social levels. Users have accepted the use of their images in a non commercial use and in the limit of the experimentation and analysis of this research.

3.2.1 The Five Adult Users

Jean
Age: 16
Use: Daily
Experiment: > 5 years
Maximum acceptable search time: 15 minutes
Trade: student or school
Comment: He is the referent. He uses internet more than often to search, to mail or to use instant messaging programs, since 1995! He is part of the Internet generation!

Annie
Age: 72
Use: Monthly
Experiment: > 5 years
Maximum acceptable search time: 25 minutes
Trade: Director (retired)
Comment: Annie only used Internet with a coach. Here are her first steps in solo.

George
Age: 49
Use: Never
Experiment: None
Maximum acceptable search time: 15 minutes
Trade: Craftsman
Comment: George is really a complete beginner but his interest for motorbikes and the brand Moto-Guzzi is a good motivation.

Marie
Age: 37
Use: Rare
Experiment: lower than one year

Maximum acceptable search time: 15 minutes
Trade: Liberal profession
Comment: "When I do not find I ask my children. I sometimes have big trouble to find". On the other hand for my emails, it is impeccable! ". Marie is an artist by passion and scientist by profession.

Yasmina
Age: 20
Use: Monthly
Experiment: superior to five years
Maximum acceptable search time: 15 minutes
Trade: Employee
Comment: Is Yasmina a girl who only uses Internet to make research on lyrics? Yasmina in spite of her engagements will not continue the protocol; thus we will have only the imposed requests.

3.2.2 The Four Children

Guillaume
Age: 11
Use: Weekly
Experiment: between one and five years.
Maximum acceptable search time: 15 minutes
Trade: student or school
Comment: Guillaume loves music and computers. He is a data processing passionate. Patient, posed, he will find what he seeks.

Celine
Age: 11
Use: rare
Experiment: less than one year
Maximum acceptable search time: 15 minutes
Trade: student or school
Comment "I never went on Internet to seek all alone…" Celine is a dreamer but a child who however does not miss tenacity.
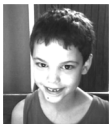
Lou
Age: 7
Use: Weekly
Experiment: between one and five years
Maximum acceptable search time: 15 minutes
Trade: student or school
Comment: Lou works well at school. She always wants to do her best.

Paul
Age: 7
Use: Rare, he is a real beginner.
Experiment: Never
Maximum acceptable search time: 15 minutes
Trade: student or school
Comment: Paul is often "on the moon".

3.2 Measurement

The 9 users completed 21 researches composed of 113 requests for a total of 829825484 returned sites.

3.3.1 Research Measurement

Table 1: The research: Time and number of requests by research

/	Time	Requests
Max	01:03:30	20
Average	00:18:45	4.5
Min	00:01:30	0

3.3.2 Request Measurement

Table 2: The search request – Number of keywords and number of sites returned by request.

/	Keywords	Found sites
Max	8	230000000
Average	3.44	7343588.
Min	1	0

Table 3: The navigation – Number of read sites, number of read pages in these sites, and number of search engine consulted pages.

/	Sites	Pages	Search engine pages
Max	44	17	9
Average	8.73	3.53	2.6
Min	1	1	0

Table 4: Search success – This tab is done by users in concordance with usage and experience.

Users	Usage	Experience in year	%found
Annie	rare	>5	40
Céline	monthly	<1	50
Georges	never	0	40
Guillaume	weekly	between 1 and 5	100
Jean	daily	between 1 and 5	60
Lou	never	between 1 and 5	50
Marie	rare	< 1 year	60
Paul	never	0	50
Yasmina	monthly	> 5	33
Average	*N/A*	*N/A*	*53,3*

3.3.3 Keyword Measurement

With 3.44 we can notice that the number of keywords increased from precedent observations. But even with an increase of the number of keywords, the average number of found sites is enormous. The average number of returned sites, using mainly Google was up to 7 millions. So we can affirm that keywords used in the observation are

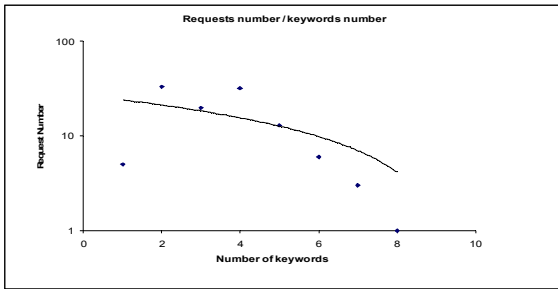not efficient enough and compel to filter the mass of information in the web.

Figure 1: Groups of dots – number found sites/ keywords number.

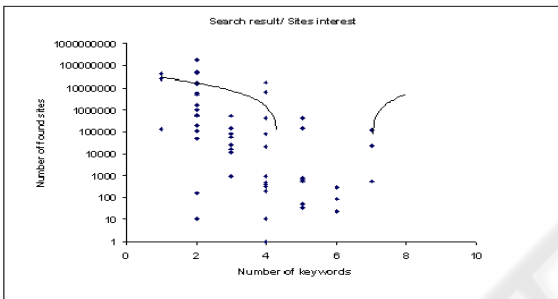Most requests are using between two and 6 keywords (Figure 1).

Figure 2: Groups of dots – Request number / Keywords number.

The found sites number decrease with the increase of keywords until the number of 6 keywords, we can notice an inversion of this tendency for highest value (Figure 2).

3.4 Users Feeling

With an average general satisfaction note at 6.25 the feeling on internet as information base and on its capability to find information could be improved.

Results are very different from a user to another (Table 5).

The exercise was based on the research of information. The users play the game so they were very impacted by the fact that they found the answer or not. The real time spent on the research does not match the general note.

Table 5: Average satisfaction note by user (from 1 to 10).

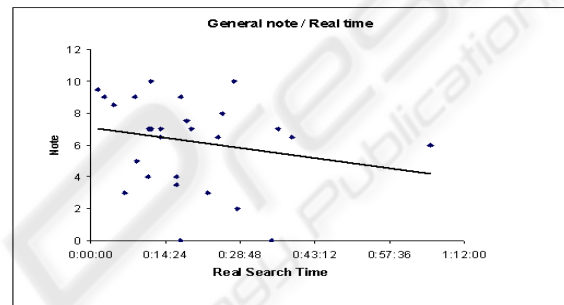| Users	Readable	Interest	Time	General
Annie	7.4	6	7	6.3
Céline	7.5	6	4	4
Georges	5.5	5.5	6.5	6
Guillaume	7.5	7	7.5	7
Jean	8.4	6.5	5.8	6.3
Lou	2	3.5	5	5.5
Marie	4.8	5.4	5.2	5.7
Paul	7	10	5	10
Yasmina	7	5.75	7.25	6.5
Max	*10*	*10*	*10*	*10*
Average	*6.41*	*6.05*	*6*	*6.25*
Min	*0*	*0*	*0*	*0*

Figure 3: Groups of dots – Real Search Time / Note.

Even the time perception does not match the real time. But the fact to find or not is in correlation with all notes.
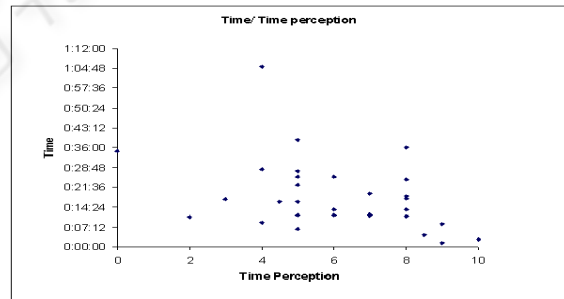

Figure 4: Groups of dots – Real Search Time/Time cost.
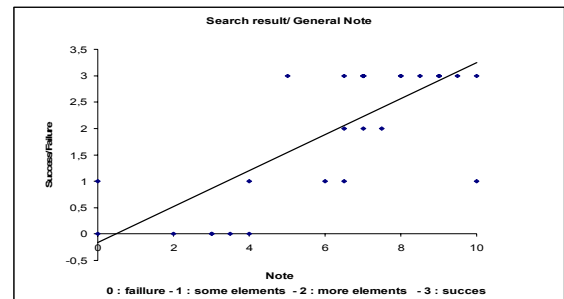

Figure 5: Groups of dots – Search Success or Failure /general Note.

Time perception and site interest seem to be more in correlation with the success or the failure of the research than anything else. We have to read the users notes very carefully. Users were obsessed with finding answers. If they found them, it was short and interesting if not it was too long and boring.

Figure 6: Groups of dots – Search Success or Failure / Subjective Time Cost.
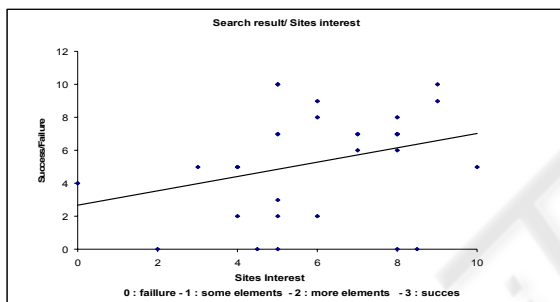

Figure 7: Groups of dots – Search Success or Failure /Subjective site interest.

3.5 The Users' Difficulties

The users' difficulties are located on each basic task that composed the internet information search task.

- Determine the efficient keywords in charge of filtering the web and use words in non mother tongue Languages ;
- read and choose in the list results including URL, acronyms and different languages mixed;
- extract the information from an object;
- manage research time and time perception

3.5.1 Difficulties to Find the Efficient Keywords

With an average of found sites number up to 7 millions, we can certify that the 3.44 keywords are not efficient enough to filter the Web. Users have difficulties to use more keywords and to choose these words. This is a problem of search nature user

competences: i.e. users do not know the subject well enough to enrich the request. It will also be a mix between indolence and confidence in search engine.

The user does not manipulate words in an unknown language. Looking for rumanian music information, nobody has used rumanian keywords (even if translation tools were known). Only Jean will use the song words to found the lyrics in English (Jean is bilingual French-English).

Jean has mixed French and English, using the word "resume" in French to mean abstract.

Keywords are selected in concordance with her or his knowledge of the subject, and the knowledge of the search engine. We also have to consider that keywords could be use with different spellings.

More and More young people use SMS. During the observation children will find information using SMS language (by mistake?) in their request.

3.5.2 Choose in the list of Proposed Sites

The list returned by search engines is constituted of some elements, Here Google is the sample: the site title; the snippet (a part of sentences including the keywords) and the URL.

Site title: In the site the title is very often non present. In this case Google chose other meta tags as "title" "author" and build a sort of title. If the title is too long Google chooses to cut and present only a part.

Snippet: This is just a few words around keywords extracted from the site. The extract can include any words from any part of the site. There is no global sense in the snippet when code is not mixed in it.

The URL: Uniform Resource Locator. See "RFC 1630"

We can imagine, a beginner looking for this information: The title has no sense, the Snippet is just a list of sentence extracts, and the URL Looks as binary code. For example, Annie who is 70 years has asked Google on *« natura 2000 marais de gabarret »* . Natura 2000 is a French ecologic organisation, "marais" means marsh and "gabarret" is a small village in France. Google sends back this!
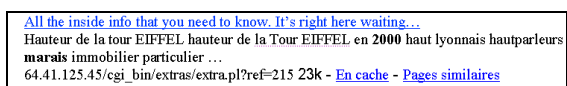

Figure 8: Google typical sites list.

133

In figure 8, the title is in English, the snippet in French speaking about the Eiffel tower and the URL is an IP address. Annie had chosen to burst out laughing.

3.5.3 Extract Information from the Object

From the page: It is difficult for the users to get information from the page. All the users have missed information on the page. It could be because:
- they stop reading the page before the information was displayed , the page is too long;
- they just read the posted part of the page;
- the page was too loaded, especially with advertising;
- they used a search function on a site instead of using a Web search engine (Quid, Amazon, …);
- they always used the same link instead of read the read the full page;
- they stopped reading if the page or part of it is in an unknown language.

From a file: Jean was looking for a music partition. He found a midi file. But he did not know that it is possible to extract the partition from it.

3.5.4 Manage Research Time and Time Perception

The majority of users have declared that the maximum research time is around 15 minutes. Annie has given 25 minutes. With an average search time of 18 minutes and 45 seconds and 50% of the research up to 15 minutes this is a reason of disappointment. If research time is less important than success or failure in the average general note on the research up to 15 minutes is only 5.5.

3.6 Observation Conclusion

Internet users are disappointed by the search process. They have difficulties to communicate efficient keywords and to extract information from the found sites list and from the site itself.

Following the observations, the questions and the difficulties of the Net surfers, it appears that:
- the choice of the keywords in a semantic field does not seem sufficient to bring back a reasonable number of sites and to guarantee a scheduling related to this semantic field
- the Net surfers cannot carry out requests in foreign languages;
- the Net surfers do not understand the texts turned over by the search engine;

- if the turned over texts are in a foreign language; the situation can be even lived as disappointing;
- the Net surfers cannot extract information from a starting format containing it.

The implementation of an "intelligent" system of simplification of the services of Internet based on a better knowledge of the Net surfer, of his tastes, competences and resources could contribute to a better adaptation of the turned over sites. The taking into account of a machine translation of the nonknown languages of the Net surfer towards known languages, the indication of the contents and nature of the sites by a system of icons would reassure this one. The proposal for a change of format or nature of information by Web services is also a possible solution to help all one each one with better exploiting the layer of knowledge of the Web.

4 SEARCH AND INFORMATION AQUISITION INDIVIDUAL HELP BY USING SELECTED WEB SERVICES

4.1 Identify Individual Capabilities and Knowledge Including Handicap

Today users are not included in the search process in term of capabilities, knowledge and taste. Search engines work in the same way for very different people.

The general concept of this project is to propose web services to a user. The goal is to help him or her in the search process.

To include the user in the process, we define a user as a set of resources. These resources recorded in a profile, can be divided in two main parts:
- Communication resources;
- Knowledge resources.

Communication resources: these resources describe the capability as sight or hearing of the users.

Knowledge resources: these resources describe the knowledge of the user like languages, music reading or practice, math, Internet knowledge etc …

Handicap: These capacities and knowledge could be less than the standard or average ones.

Can we speak of handicap?

Yes, but the handicap is not where we can think it is. In searching the web a light sight or hearing handicap could be less penalizing than the fact that you don't speak English, or you do not know how to read URL.

In this way, we are all in trouble when facing the internet. Who can say "I speak all the languages, I know how to read music partition, I know enough math to read all demonstrations, I understand the SMS language and I can read PDF documents from binary. So I do not need any software to do it"?

In reality when facing massive and diverse information we are all handicapped.

We will add in this profile a list of proffered subjects or user tastes. This is an important element to help user to search filtering the web. The full user profile including his/her tastes, experiences, recorded researches; used keywords will be saved, in anonymous way, to have the have the best knowledge of the user.

4.2 Offering Services to Help User to Be More Efficient and Less Disappointed

We can propose solutions to help users in their difficulties. These solutions have to be individual. Each user needs adaptive solution. These solutions could be very complex to implement. The idea is to create or use for each solution an independent Web service(s). Trying to start from this observation we can imagine a list of services to help our users.

4.2.1 General Services to Help Users

In chapter 3.4 we identified ours user's difficulties. We can propose solutions to help then. These solutions have to be individual. Each user needs adaptive solution.

Which services to help users to determine the efficient keywords in charge of filtering the web?
- a service to build smart search requests based on user keywords, on previous requests and on semantic space. The semantic space is determined by the knowledge of the user taste and previous researches. These requests could filter the web more (efficiency) efficiently.
- a service translating Keywords from SMS to official orthographic and from official orthographic to SMS ;
- a service to translate keywords to a specific language and an automatic translation of the results;
- a service to collaborate with others users by sharing request and resources;
- a service to correct and create spelling mistakes in order to widen researches.

Which services to help users to read and choose in the list results?
- a service to give an explicit reading of URL;
- a service to give a more detailed snippet;
- a service to add icons to characterize the site contents (shop, forum, encyclopaedia,..)

Which services to extract the information from an object?
- a service to highlight key-word.
- a service to hide pictures and entertainments ;
- a service to display a picture (map) of the page to help to the localisation;
- a service to extract files from any archive type a web service to send by email the content.

Which services to help user to manage research time and time perception?
- A service to measure and control time with the possibility to save and reuse the research key-words and results.
- a service to analyse found sites to help users to modify the request;

4.2.2 Offering Services to Help "Handicapped" Users

These web services could be use to help people with accessibility troubles. For example we can imagine web services:
- for people with myopia trouble a service guaranteeing a minimal font size to display pages;
- For colour-blind people a service correcting colours to get the best readable pages;
- for people having understanding handicap a service placing a link to definitions for all detected words as "difficult words";
- for blind people a service reading out loud the site.

4.2.3 Offering Services to Modify Information Format

These Web services can extract, rebuild, complete information from a start-up document.
- For people with a specific editor a web service to transform files to the right editor format;
- for musicians extract partitions from midi file;

4.3 Defining an Open Structure Project to Record and Select Web Services: SISSI

4.3.1 Fundamentals

The goal is to define an open structure project to record and propose selected Web Services in charge of helping users to search and to adapt and complete
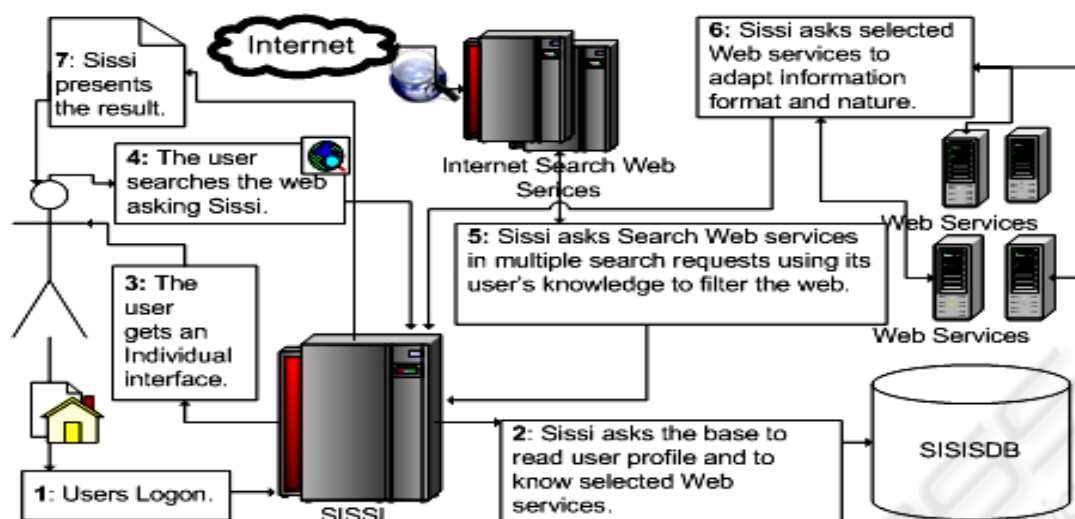
Figure 9: SISSI transactions architecture.

data nature and format to the users competency; knowledge and taste.

Articulated around a central service called SISSI: Individualized System of simplification of Internet Services (Figure 9). Web Services definitions and users definitions are stocked in a database called Sissidb.

Based on UDDI the Web services recorded in the base Sissidb present their services to the user who can select and parameter them in concordance with his/her needs and tastes.

4.3.2 SISSI and Web Services Communication

SISSI does not include any Web services. It is a connector between us. SISSI could be viewed as a generic web service client or as a generic web service IHM.

Based on UDDI web services are described more precisely in the base in the way that the user can understand the result they are expecting, and SISSI can communicate with then.

SISSI communication is based on SOAP and we know that SOAP is a heavy protocol. But the Web information manipulated by SISSI is mainly HTML pages or active code used on the client side. The complex objects as pictures stay on the Http server. If a Web service needs one of these objects it will be its responsibility to load it from the initial resource server. (In this way) big objects manipulated by SISSI are not to be loaded.

SISSI is here including search engine service but it could be invoked directly on a document by receiving an URL as argument.

5 CONCLUSION

The first feasibility works reveals a certain complexity. If the architecture model seems to be usable, we found some new challenges:

- The "real life" HTML language is not normalized. Working on it to modify a document structure or presentation as generic language is today a big work and due to this HTML management complexity writing efficient Web Services will be difficult.
- SISSI is a service based on services. That means that if a service falls down, SISSI falls down. We will have to think "fault tolerance".
- Even if they use a pseudo, users have the right to withdraw themselves from a base where their competencies, knowledge, tastes and handicaps are stored, especially if this base records their activity.

However, as regard to the related works, the observations and the experimentation carried on with SISSI, we come to the conclusion that the use of an open architecture based on Web services can be a solution to help Net surfers in Web searching. Indeed, it can make users aware from facing up to certain situations and prevent them from giving up the research, more particularly when they find some difficulties to access or to exploit some retrieved

136

information (handicapped users, surfers who do not speak any foreign languages …).

To conclude, services based architectures like the one we have implemented (SISSI) must be freely extensible in order to offer more and more. At present we have to experiment this extensibility by permitting to any user to add new services when he judges it necessary. New observations could then be carried on to analyse the capabilities of Net suffers to interpret the results provided by the use of hidden services.

Finally a better knowledge of the users' tastes, their competences, their contexts and their resources will allow the implementation of communities to share researches and results. This information could be used to refine research by the creation of semantic contexts of associates.

REFERENCES

Alexander, P.A., & Jetton, T.L. (2002). Learning from text: A multidimensional and developmental perspective. In M.L. Kamil, P. Mosenthal, P.D. Pearson, and R. Barr (Eds.) *Handbook of reading research*, volume III (pp. 285-310). Mahwah, NJ: Erlbaum.

Berkowitz, B. (2002). Moving every child ahead: The Big6 success strategy. *Multimedia Schools*, 9, 17-22.

Bransford, J.D., Brown, A.LO., & Cocking, R.R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Coiro, J. L. (2003). Rethinking comprehension strategies to better prepare students for critically evaluating content on the Internet. *The NERA Journal*, 39(2), 29-34.

Eagleton, M.B., Guinee, K., & Langlais, K. (2003). Teaching Internet literacy strategies: The hero inquiry project. *Voices from the Middle*, 10, 28-35.

Gay, L.R. & Airasian, P. (2000). *Educational research: Competencies for analysis and application*. Upper Saddle River, NJ: Prentice Hall.

Guinee, K., Eagleton, M.B., and Hall, T.E. (2003). Adolescents' Internet search strategies: Drawing upon familiar cognitive paradigms when accessing electronic information sources. *Journal of Educational Computing Research*, 29, 363-374.

Henry, L. A. (2005). "Information search strategies on the Internet: A critical component of new literacies". *Webology*, 2(1), Article 9. Available at: http://www.webology.ir/2005/v2n1/a9.html

International Reading Association. (2002). *Integrating literacy and technology in the curriculum: A position statement*. Retrieved April 22, 2005, from http://www.reading.org/downloads/positions/ps1048_t echnology.pdf

Janesick, V. J. (1994). The dance of qualitative research design: Metaphor, methodolarty, and meaning. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 209-219). Thousand Oaks, CA: Sage Publications, Inc.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management*, 36, 207-227.

Laurie, A. H., (2005). Information Search Strategies on the Internet: A Critical Component of New Literacies University of Connecticut, Connecticut, U.S.A.

Leu, D.J., Jr. (2000). Our children's future: Changing the focus of literacy and literacy instruction. *The Reading Teacher*, 53, 424-431.

Leu, D.J., Jr., Kinzer, C.K., Coiro, J., & Cammack, D.W. (2004). *Toward a theory of new literacies emerging from the Internet and other communication technologies*. In R. Ruddell and N. Unrau (Eds.) Theoretical models and processes of reading (5th ed.). Newark, DE: International Reading Association.

Nachmias, R. & Gilad, A. (2002). Needle in a hyperstack: Searching for information on the World Wide Web. *Journal of Research on Technology in Education*, 34, 475-486.

O'Sullivan, M., Scott, T., Ph.D March/April 2000 Teaching Internet Information Literacy: A Critical Evaluation http://www.infotoday.com/MMSchools/mar00/osulliv an&scott.htm

Pressley, M, & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsible reading*. Hillsdael, NJ: Erlbaum.

RAND Reading Study Group. (2002). *Reading for understanding: Towards and R&D program in reading comprehension*. Retrieved March 3, 2004. Available Online: http://www.rand.org/multi/achievementforall/reading/r eadreport.html

Salton, G. (1970). The SMART Retrieval System. Englewood Cliffs, NJ: Prentice-HiII

Salton, G., & McGill, M.J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Savoy, J., Picard, J. (2001). Retrieval effectiveness on the Web. *Information Processing and Management*, 37, 543-569.

Sutherland-Smith, W. (2002). Weaving the literacy Web: Changes in reading from page to screen. *The Reading Teacher*, 55, 662-669.

Williams, P, & Nicholas, D., (2006) Testing the usability of information technology applications with learners with special educational needs (SEN) http://www.blackwell-synergy.com /doi/full/10.1111/j.1471-3802.2006.00057.x

Yin, R.K. (1989). *Case study research: Design and methods* (2nd. ed). Sage: Newbury Park, CA.