

# SEEKING FOR CLUES ABOUT USERS' INFORMATION NEEDS IN THEIR NAVIGATION

Jean- Yves Delort  
LIRMM  
Montpellier 2 university  
delort@lirmm.fr

## ABSTRACT

Adaptive navigation systems are tools intended to assist users when they search the Web. Thus, they often have to know the users' current information needs (IN). However, they suffer from the lack of information available about users' IN and the lack of user feedbacks. Accordingly, they have to seek for clues about users' IN in their navigation behaviours. This paper puts forward a method that aims at discovering clues about users' IN in the content of the documents they access without requiring their feedbacks. The paper also presents an incremental method to detect the users' shifts of focus. Eventually, details of an experimentation with real users are outlined and results show that the proposed approach has a good ability to detect the users' shifts.

## KEYWORDS

Implicit interests, information needs, shifts of focus.

## 1. INTRODUCTION

Adaptive navigation systems (ANS) are tools intended to assist users when they search the Web. Thus, they often have to know the users' current information needs (IN). However, they suffer from the lack of information available about users' IN and the lack of user feedbacks. Accordingly, they have to seek for clues about users' IN in their navigation behaviours. Unfortunately, little information is available because the users do not provide the ANS with *explicit information*. First, they generally do not give feedbacks and second, users generally do not specify the *words* that correspond to their current information needs<sup>1</sup>.

As a consequence, most ANS must rely on implicit methods that look for clues about the users' feelings and IN without requiring their feedbacks. There is extensive literature on the characterization of user's IN from her/his behavior. The issue is particularly important for hypertexts, where researchers have tried to understand the nature of linking and define it with respect to needs the user's IN. For instance, typed links (De Rose 1991, Nanard 1991) allow to know the users' intentions from their interactions with an hypertext. However, first, typed links do not offer an easy way to understand the navigation as a whole and, second, many hypermedia applications, like the Web, do not

---

<sup>1</sup>Except when they query web search engines, but then, they barely use 2-3 words which is hardly enough to make a precise representation of their IN.

contain typed links. A common assumption is to consider that linking a page comes down to an implicit vote in favor of it. Various other heuristics have been put forward to figure out the users' interests from their behaviors : For instance (Lieberman 1995) argues that the time spent on a webpage is connected to the user's interest or (Claypool 2001) asserts that the amount of content read on a page (computed from the scroll activity) is bounded to the users' current interests.

Our research is based on the following idea: the pieces of information which are common to most pages accessed during the same search activity contain relevant clues.

This paper organizes as follow: First, the concept of clue about information needs and the clue enumeration issue are presented. Then, a method to extract the relevant clues about IN from a navigation without user feedback is outlined. Finally, the paper summarizes the results of an experimentation using the proposed method to detect the users' shifts.

## **2. RELATED WORK**

Many researches tackle with the issue of detecting signs of user's preferences in his/her behavior in a hypermedia system. For instance, (Claypool 2001, Lieberman 1995) consider that the time spent on a webpage is connected to the user's interest. (Claypool 2001) also asserts that the amount of content read on a page (computed with respect to the scroll) is bounded to the user's current interest. Based on the up- to- down- left- to- right way of reading, (Lieberman 1995) also suggests that the position of the clicked link on the displayed page says that links located more on the left and/or on the top of the page were judged not interesting by the user. Conversely, studies by (Fuller 1996, Konstan 1997) have proven that the time spent on a document is not a good predictor of the user's interest which partly contradicts (Claypool 2001, Lieberman 1995).

As some recent applications tend to show, the content of the accessed documents seems to be a promising way to extract useful information about the user's IN. For instance, in (Davison 2002) different prefetching algorithms based on a similarity value between the text surrounding the links clicked by the user and the content of previously seen documents are proposed. Beltagy et al. (2001) have put forward an open hypermedia system (OHS) based on the content of the documents accessed by the user to choose the links that will be inserted into the next page she/he will see. One of the important strength of their approach is that it does not require the user feedback to work. However, a limitation is that only the last document accessed by the user is taken into account to suggest him/her the links. In (Widyantoro 1999) user IN are distinguished between short and long term interests. They are represented by means of interesting topic-related categories which are learnt using the content of the current accessed document and the user's feedback on the page. Zhu et al. (2003) have proposed a system that is able to predict if the content of the current page satisfies the user needs. Their method aims at detecting common features between the accessed documents but these features are independant from the textual content and also uses the user feedbacks.

## **3. DEFINITIONS**

Navigation interactions are a type of browser interactions the user makes in order to see a different page from the one displayed in the current browser window. Examples of

navigation interactions are: clicking on the back button, submitting a form or following a link. Examples of non- navigation interactions are: adding a bookmark or searching a pattern of words in the page. In the sequel, *an interaction* refers to a navigation interaction. An interaction  $i$  is characterized by the pair  $(t,d)$  where  $t$  is the time when the user did the interaction and  $d$  is the *associated document* consequence of the interaction  $i$ .  $d$  is to be considered as a set of words. Choosing  $d$  as a crisp set comes down to represent it in the vector- space model (Salton 1975) with the boolean weighting system (assigns 1 if the word is present, 0 otherwise). However, numeric values (e.g. the frequency of the words in the documents) can also be taken into account representing  $d$  as a fuzzy set (fuzzy sets offer the possibility to handle non- boolean values and extend the classic crisp set operators).

An *interaction stream*  $S$  is a strictly ordered set of interactions by a given user with respect to the time attribute. Thus an interaction stream is a sequence of pages *of any kind* accessed by a user with his browser. The size of an interaction stream is the number of interactions it contains. The set  $P(S)$  will refer to the powerset of  $S$ .

## 4. CLUES

In this section, the concept of clue about a user's IN is defined in the context of a certain kind of navigation. Then, a couple of examples are given before it is formalized.

Marchionini (Marchionini 1995) distinguishes between two main styles of information-seeking behaviors: During a *search activity* (SA) the user has in mind specific features or characteristics of the objects that will be used to satisfy her/his IN. It differs from a *browsing activity* which is defined by an effort to explore an area in the information space that is likely to contain relevant information and the borders of which are fuzzy.

In this paper, we focus on Marchionini's first kind of user's behavior, the search activity and we assume that in this context, the words which are common to most pages accessed during the same SA contain relevant clues about the user's IN. Clearly, the characteristics of the users' IN are used to specify and guide the navigation (and thus the interactions). But conversely, as the Bates's evolve-berrypicking model shows (Bates 1989), users' interactions are also likely to make their current IN evolve: as a consequence of their viewing the intermediary result sets, features they have in mind of their current IN may be changed, removed or added.

Let us explain our hypothesis with two examples. First, assume that the user's goal is to find when the Eiffel Tower was built in Paris. Then, he may go on his favorite search engine and make a query such as `"eiffel tower paris year building"`. Because most Web search engines use simple word matching, the user will probably access result pages containing all these words. Additional terms not explicitly given are likely to be also found in the contents of the result pages, such as *France* or *International Exposition*<sup>2</sup>. As a consequence, though the user has made a query with few terms about his IN, implicit relevant pieces of information can be drawn from the common features of the accessed documents. Suppose now that the user wants to buy a plane ticket for the IADIS Conference on the Internet and the Web. Then, she/he may go on various airline company websites in order to compare their prices. He may never explicitly mention his IN.

---

<sup>2</sup>The Eiffel Tower was built as a focal point for the Paris International Exposition of 1889.

However, in the content of many of the accessed pages, terms such as *airlines, Madrid, Spain, prices, tickets* are likely to be found.

Formally, a clue about an information need corresponds to a pair  $(I,W)$  where  $I$  is a set of interactions and  $W$  is the (not-empty) intersection of the content of their associated documents.

## 5. THE CLUE ENUMERATION ISSUE

The clue enumeration issue refers to the problem of enumerating all the possible clues in an interaction stream. In this section the issue is explained and the concept of relevant and irrelevant clues about IN are defined.

A simple enumeration of all the clues is not possible because the solution set size could be up to  $2^n$  for a stream of size  $n$ . However, many of the clues in an interaction stream may not discriminate the user's current IN. Such a clue would be for instance a clue the interaction set attribute of which belongs to more than one search activity. A clue  $c=(I,W)$  will be said *irrelevant* if  $I$  does not belong to a unique search activity.

Accordingly, the key issue in the clue enumeration problem is to choose conditions for a subset of interaction  $I$  to be a part of a unique search activity. A condition for a set of interactions to belong to the same search activity is a criterion that states whether these interactions could or could not have been done to achieve the same goal. In the simplest case, we can consider a binary criterion<sup>3</sup>. Formally, given an interaction stream  $S$ , the condition for a subset of interactions to be part of a unique search activity can be formalized by a boolean function  $F$  on the powerset of  $S$ . Thus, given  $F$  and  $S$ , the set of *relevant clues* is the set of clues  $R$  defined as follow :

$$R = \{c=(I,W) \text{ such that } c \text{ is a clue in } S \text{ and } F(I)=1\}$$

Let us consider an example. Assume that  $S=\{i_1,\dots,i_7\}$  is an interaction stream containing two successive search activities such that the four first interactions,  $\{i_1,\dots,i_4\}$  correspond to the first search activity and the three remainings,  $\{i_5,i_6,i_7\}$  correspond to the second one. Documents associated with  $S$  are described by subsets of words of  $\{a,b,c,d,e,f,g\}$  according to the distribution of table 1.

Table 1 : Distribution of words in the documents

	a	b	c	d	e	f	g
$i_1$	x	x		x		x	
$i_2$	x			x	x		
$i_3$	x	x	x				
$i_4$	x	x	x				
$i_5$					x	x	
$i_6$				x	x	x	x
$i_7$		x					x

In table 1, an "x" means that the document in column contains the word in row. The set of clues drawn from the stream  $S$  contains elements such as  $\{(i_1,i_3,i_4),(a,b)\}$  and  $\{(i_5,i_6),(e,f)\}$  which are relevant. On the other hand,  $\{(i_1,i_2,i_6),(d)\}$  and  $\{(i_2,i_5,i_6),(e)\}$  are irrelevant because their interaction set attributes contain interactions that did not occur during the same search activity. Thus, these clues should be discarded.

---

<sup>3</sup>The problem would however benefit from the introduction of fuzzy values which would handle in a more suitable way situations where, for instance, the goal of a search activity is modified.

## 6. THE CLUE EXTRACTION ALGORITHM

The *Clue Extraction Algorithm* (CEA) is an incremental algorithm for finding clues in an interaction stream  $S$ . The purpose of the CEA is to extract set of words that belong to interaction occurring *in a row or almost in a row*. Due to size limitation the details of the algorithm are not published in this paper. However, in the sequel we present its principles.

The CEA takes into account the fact that the same search activity can contain documents that have none common features with their previous and following neighbours in the interaction stream. Then, it introduces a value  $N$  that represents the maximal gap size between two interactions having at least one feature in common.

Let us come back to the previous example and choose  $N=1$ , then the clues found by the CEA are (sorted with respect to the time attribute of their first interaction in the stream) given in table 2.

Table 2 : Example of clues found by the CEA

id		$i_1$	$i_2$	$i_3$	$i_4$	$i_5$	$i_6$	$i_7$
1	{a, b, d, f}	x						
2	{a, d, e}		x					
3	{a, b, c}			x				
4	{a, b, c}				x			
5	{e, f}					x		
6	{a, d, e, f, g}						x	
7	{b, g}							x
8	{a}	x	x	x	x			
9	{a, b}	x		x	x			
10	{d}	x	x					
11	{c}			x	x			
12	{e, f}					x	x	
13	{g}						x	x

In table 2, each line corresponds to a different clue. An ``x'' means that the associated document of the interaction in column contains the set of words in row. The upper part of the array displays the single- element clues and the lower one contains the clues the size of the interaction set attributes is greater than one. These clues show the relationships between the interactions. We can see that the remaining clues belong to either one or the other of the two search activities. There is no clue the interaction set attribute of which overlaps the two search activities.

## 7. APPLICATION

In this section we present the details of an experimentation intended to test the efficiency of the clues for a specific issue for many adaptive navigation systems. We consider the task of detecting the frequent users' shifts of interest or shifts of focus in a Web navigation. Note that this task is harder than detecting ``user sessions''. A user session is the sequence of pages accessed by a user on a website in order to achieve a given task. Existing method are usually taking into account only time-based criteria (He 2000). Accordingly, they are not enough sensitive to detect the user real shifts. In addition they cannot untangle parallel search activities. To address this problem, we have assumed that

the content of the accessed documents could help to improve the results of existing methods.

This section organises as follow: First, the section summarizes the details of the survey. Second, it deals with the evaluation issue. Then, an alternate approach to this issue that considers the problem in a text segmentation framework is introduced. Eventually, results are presented and discussed.

## 7.1 Survey

The survey involved ten web-skilled participants, all of them researchers or PhD students at the CS laboratory of the University Paris 6. The participants were provided a questionnaire with 17 information seeking problems involving only SA. User's shifts are defined here as the times when they end up searching for a question and start searching for a new one. Thus, interactions that occurred between two consecutive shifts were intended to satisfy the same IN. These shifts can easily be identified because they correspond to the transition from one question to another in the search process. Note that, *in real life*, user's shifts could be harder to detect if her/his IN were to change but still kept on being related to the same topics. Questions in the questionnaire are designed so that they force the users to experience different information-seeking strategies :

- Some questions were difficult, if not impossible, to answer just using a search engine. The interest of these questions is to see if relevant clues can be extracted from documents which have not been accessed by means of user queries. For instance: "What is the exhibition schedule of the New York Museum of Modern Art?"
- Some questions required to see at least more than one page to be answered. Such questions were intended to see if relevant clues can be extracted with search strategies combining smaller search strategies. For instance, "Which one of these two cities, Patras (Greece) and Vidin (Bulgaria) has the larger population?"
- Some questions were really hard to answer, so that few participants could answer them. The interest of these questions was to require really complex search strategies when the user tries to address the question from different points of view. For instance, "What is the ranking of the top five personal computer resellers in the EU market?"

Users' interaction streams were collected thanks to a modified version of the Mozilla web browser, however they kept their profile being thus able to use their favorites and history. HTML tags in the documents were removed and the remaining texts were filtered with a stopword list and stemmed with the Porter's stemming algorithm.

## 7.2 Evaluation issue

Here we present the evaluation measures. They depend on the possible shifts. If we know the time of the first  $t_f$  and last  $t_l$  interaction of a search activity, then three cases can be considered :

- Only one possible shift is detected at time  $t$  with  $t_f \leq t \leq t_l$  (relevant shift). It means that the algorithm has discovered that the user has started a new search activity.
- Two or more possible shifts are detected at times within  $t_f$  and  $t_l$  (duplicated shift). It means that the algorithm has detected more search activities than what really occurred.
- None possible shift is detected (missed). It means that the algorithm hasn't been able to detect that the user has started a new search activity.

Formally, a possible shift is a *relevant shift* as soon as it occurs during a search activity. A possible shift is a *duplicated shift* if a previous relevant clue was found within the same search activity boundaries (duplicated shifts could be said false-positive). A search activity is *missed* if no relevant shift was found within its boundaries. Let  $P, R, D$  be the numbers of possible, relevant and duplicated shifts respectively, and  $M$  be the number of missed search activities, then <sup>4</sup> :

$$P = R+D$$

$$M=17- R$$

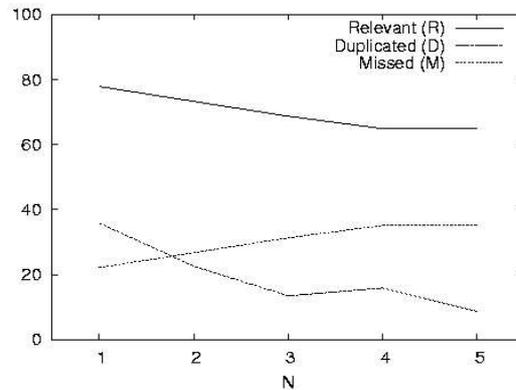
### 7.3 Results and discussion

In average, it took about 58 minutes to the participants to answer the questionnaire. During this time, they performed an average of 192 interactions. Thus, the average number of interactions to answer a question was 11.3 though two questions required about 20 interactions each. The ratio of search engine pages (result pages, cached documents and homepages) over the total number of accessed documents represents 40%. Interactions of the participants' interaction streams were manually annotated according to the definition of shift given above.

The CEA was then applied on each user interaction stream, in addition we removed all the clues found by the CEA having a number of interactions or a number of words lower than 4.

Figure 1 gives the average percentages of  $R, D$  and  $M$  with respect to the number of questions when the parameters  $N$  varies. The percentage of relevant shifts and missed search activities are computed with respect to the number of real shifts (17) while the percentage of duplicated shifts is computed with respect to the number of possible shifts ( $R+D$ ).

Figure 1 : CEA



At  $N$  increment, the number of relevant shifts slightly decreases. As the number of covered interactions increases, the number of possible shifts decreases and thus, the number of missed search activities increases while the number of duplicated shifts decreases. At  $s$  increment, the numbers of relevant and duplicated shifts decrease and the number of missed search activities increases. Note that, though the number of relevant shifts is in this case higher than with the CEA, the number of duplicated shifts (false detection) is also higher.

<sup>4</sup>Let us remain that 17 is the total number of questions in the questionnaire.

## 8. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a method that can give adaptive navigation systems the possibility to have a better representation of the user's current IN. A strength of the proposed approach is that it does not use the user feedback. This method has been successfully tested for the task of the online detection of user shifts. However, in a real life search context, the user shifts could be harder to detect if the user's information needs were to change but remained related to the same topics.

Future work will focus on finding new criteria to be able to cope with a broader set of web search situations: for example, when users are looking for different things at the same time using different browser windows. Furthermore, future work will study the relationships between the frequency of the words in the accessed documents and the user's IN.

## REFERENCES

- Bates M.J., 1989. The Design of browsing and Berrypicking techniques for the only search interface. *Online review*, 13(5), 407- 431.
- Brusilowsky P., 2001. Adaptive Hypermedia. *User Modeling and User-Adapted Interaction*, 11, 87-110.
- Claypool M. et al, 2001. Implicit interest indicators. *Proceedings of Intelligent User Interfaces*. p. 33-40.
- Davison. B. D., 2002. Predicting web actions from html content. *Proceedings of the Conference on Hypertext and Hypermedia*, p. 159- 168.
- De Rose S. J., 1991. Expanding the notion of links. *Proceedings of the Conference on Hypertext and Hypermedia*.
- El-Beltagy S. et al., 2001. Linking in context. *Proceedings of the Conference on Hypertext and Hypermedia*, p. 151- 160.
- Fuller R. and de Graafe, J. J., 1996. Measuring user motivation from server log. *Conference presentation for Designing for the Web: Empirical Studies*.
- He, D. and Goker, A., 2000. Detecting session boundaries from web user logs. *Proceedings of the BCS- IRSG 22nd Annual Colloquium on Information Retrieval Research*.
- Kaufmann, S., 1999. Cohesion and collocation: Using context vectors in text segmentation. *Proceedings of the Annual Meeting of the Association of for Computational Linguistics*, p. 591-595.
- Konstan, J. et al., 1997. Grouplens: Applying collaborative filtering to usenet. *Communications of the ACM*, 40(3), March, p 77- 87
- Lieberman, H., 1995. Letizia: An agent that assists web browsing. *Proceedings of the International Joint Conference on Artificial Intelligence*, p. 924- 929.
- Marchionini, G., 1995. *Information Seeking in Electronic Environments*. Cambridge University Press.
- Mobasher, B. et al., 2000. Combining web usage and content mining for more effective personalization. *Proceedings of the International Conference on E-Commerce and Web Technologies*.
- Nanard, J. and Nanard, M., 1991. Using structured types to incorporate knowledge in hypertext. *Proceedings of the Conference on Hypertext and Hypermedia*, p. 329- 343.
- Salton, G. et al, 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18, p. 613- 620.
- Schwab, I. et al. Learning user interests through positive examples using content analysis and collaborative filtering.
- Widyantoro, D. H. et al., 1999. An adaptive algorithm for learning changes in user interests. *Proceedings of Conference International on Knowledge Management*.

- White R. W. et al., 2003. Adapting to Evolving Needs : Evaluating a Behaviour- based search interface. *Proceedings of Conference International on Knowledge Management*.
- Zhu, T. et al., 2003. Learning a model of a web user's interests. *Proceedings of the International Conference on User Modeling*, 2003.