CEA : A Content-Based Algorithm to Detect Users' Shifts of Focus on the Web

Jean-Yves Delort — Bernadette Bouchon-Meunier — Maria Rifqi

Laboratoire d'Informatique de Paris 6
Université Pierre et Marie Curie
8, rue du Capitaine Scott
75015 Paris, France

{Jean-Yves.Delort,Bernadette.Bouchon-Meunier}@lip6.fr

RÉSUMÉ. *This paper addresses the issue of detecting the users' changes of their current information needs or* shifts of focus*. The proposed approach is to find relevant clues about the user's current interests from the content of the accessed pages. First, the concept of clue about user's information needs is formalized. Secondly, a clue extraction algorithm in a stream of accessed pages is presented. Finally, this paper discusses the result of a Web navigation survey intended to characterize user's shifts of focus.*

ABSTRACT. *Dans ce papier un nouvel algorithme permettant de détecter les variations dans les besoins courants d'information des utilisateurs est présenté. L'approche proposée s'appuie sur l'extraction d'indices pertinents sur les besoins courants d'information de l'utilisateur à partir du contenu des documents auxquels il accède. Dans un premier temps le concept d'indice est expliqué puis formalisé. Ensuite les principes de CEA, notre algorithme d'extraction d'indice dans un flux de documents accédés sont présentés. Enfin les résultats d'une étude expérimentale visant à détecter les changements d'intérêts courants d'utilisateurs sont discutés.*

MOTS-CLÉS : *Navigation, hypermedia, besoin d'information.*

KEYWORDS: *Navigation, hypermedia, information needs.*

1. Introduction

Adaptive Navigation systems (ANS) are intended to support user navigation in a hypermedia system. ANS are able to adapt to users' behaviors, *i.e.* their past and current preferences, information needs and navigation styles. On the Internet, ANS suffer from the difficulty of understanding users' current information needs and the frequent users' shifts of focus (hereafter shifts). Lately, ANS using the content of the accessed documents by the users have been proposed [ELB 01, FIN 01, DAV 02]. Each of these works concluded that relevant information about the user's current information needs could be extracted from the content of the accessed documents. In many information-seeking models, interactions are indeed directly connected to users' information needs. According to Marchionini [MAR 95], the characteristics of the users' current information needs are used to specify and guide the search process and thus, the interactions. But conversely, as demonstrated in the Bates' evolve/berrypicking model [BAT 89], users' interactions are also likely to make their current information needs evolve : as a consequence of their viewing the intermediary result sets, features they have in mind of their current information needs may be changed, removed or added. Thus the content of the documents accessed during a search activity is likely to 1) contain clues about the user's current information needs, 2) to trigger shifts of his attention and 3) to trigger changes in his current information seeking strategy.

This paper makes the following contributions. First, the concept of clue of information needs is introduced and formalized as a set of relevant pieces of information taken from the common content of the accessed documents during a search activity. Secondly, the architecture of *CEA*, an incremental algorithm intended to extracted clues about user's information needs is outlined. Finally, preliminary results of an experimentation intended to capture user's shifts using CEA are given and discussed.

2. Related work

The concept of search activity involves both cognitive and affective factors. Search activities have been described with respect to each of these dimensions. They have often been characterized in terms of kind of goal [SAL 90, WIL 81]. For instance, in [SAL 90], a search activity can be driven by a single goal, or it can be driven by a set of goals emerging during the search activity, and at the very least it can be serendipitous. Marchionini [MAR 95] defines two main styles of information-seeking behavior : during a searching activity (resp. a browsing activity) the user has in mind specific features (resp. general features) or characteristics of the objects that will be used to satisfy his information needs. He stresses that these features are used to guide and specify the search activity. Bates [BAT 89] emphasises that users' interactions are also likely to make their current information needs evolve : during the search process, features they have in mind of their current information needs may be changed, removed or added.

The expression "user session" is often used to refer to relations between the users' information needs and their interactions : it represents a set of accessed pages related to the same search activity. Researchers have proposed time-based detection processes of session boundaries [COO 99, DEL 02]. The user's search tactics are different from his shifts of interest : the former correspond to "moves made to further a search" [BAT 79] and user's information-seeking strategies are made up of them. However, a search tactic does not necessarily implies a shift in the user's current information needs. Users' information needs have also been studied with respect to their search query to web search engines [CAC 01, JAN 98, LAU 98] or their behavior in Web directory [CAC 01]. For instance, analyzing web search queries recorded by a popular search service, Lau and Horvitz [LAU 98] identified seven classes of shifts, and proposed a Bayesian-network model predicting the next user shift with respect to his previous queries. Research in different ANS fields (e.g. link generation [ELB 01], recommender systems [MOB 00] or prefetching [DAV 02]) have used the content of the current or previously accessed documents to get relevant clues about the user's current information needs. In [WID 99] user information-needs 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.

Research in information-seeking behavior in digital librairies is very active and concepts of search activity, information needs and information seeking-strategies and interaction have been widely studied. However, if existing work uses successfully the content of accessed documents for different kinds of application, the reason of these successes has not been explained.

3. Background

General definitions as well as the concept of clue are formalized in this section.

3.1. *Definitions*

Navigation interactions are a type of browser interactions the user does 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 [SAL 75] 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. The size of an interaction stream is the number of interactions it contains. The set $P(S)$ will refer to the powerset of S.

3.2. *Clues*

As pointed out earlier, the content of the current or previously accessed documents is likely to contain relevant clues about the user's current information needs. On the basis of this observation, we argue that during a search activity (in the sense of Marchionini), the content of the accessed documents often contains terms and pieces of information relevant or informative to the information needs associated to the current search activity. In the sequel, a search activity should be taken in the sense given by the Marchionini's definition given in the previous section.
Assume that the user's goal is to find when the Paris Eiffel Tower was built. Then the previous hypothesis suggests that the content of accessed documents during the search activity should frequently contain terms such as "Eiffel Tower" and "Paris". Furthermore, additional terms discovered during the search activity and related to the information needs are also likely to appear frequently, for instance "France" or "International Exposition"[1]. As a consequence of the previous observation, if few terms are used to deal with the current information needs or if some terms usually appear with others then the documents accessed during the same search activity are likely to contain the same words. For instance, in the previous example, assume that the terms "Paris" and "Eiffel Towel" appear in most of the accessed documents. Then because the words "Paris" and "Eiffel Towel" occurred frequently they are likely to give us clues about the user's current information needs. Assume that the pages containing "Paris" often contain the word "France" and that most of the pages containing the words "Eiffel Tower" also contain the word "France". Then, another clue about the current information needs is "France". Formally, a clue about an information need corresponds to a pair (I, W) where I is a set of interactions and W is the intersection of the content of their associated documents :

DEFINITION 1 (clue). *Let S be the interaction stream. A* clue in the stream S *is a pair (I, W) such that :*

1) $I \in P(S) \setminus \{\emptyset\}$, where $P(S)$ refers to the powrset of S,

2) $W = \bigcap_{i_k \in I} d_k$ where d_k is the associated document of k^{th} interaction i_k in the stream S,

3) $W \neq \emptyset$.

[1] The Eiffel Tower was built as a focal point for the Paris International Exposition (public exhibition) of 1889.

4. The clue enumeration issue

The clue enumeration issue refers to the problem of enumerating all the possible clues in an interaction stream. A simple enumeration to find all the pairs can contain up to 2^n clues for a stream of size n. However, many of the clues found are likely to be irrelevant because they do not correspond to a user's realistic search activity. A clue $c = (I, W)$ will be considered as an irrelevant clue if I does not belong to a realistic 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 realistic search activity. A condition for a set of interactions to belong to the same search activity is a criterion that says 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. However, the problem would 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.

Formally, given an interaction stream S, the condition for a subset of interactions to be part of a realistic 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 such that :

$$R = \{c\ :\ c \text{ is a clue in } S \text{ and } F(c) = 1\}$$

Let us see an example. Assume that $S = \{i_1, .., i_7\}$ is the interaction stream of a user who has done two search activities. Assume that we know that the four first interactions correspond to the first search activity and the three remaining correspond to the second one. Associated documents to S are described by a subset of words of $\{a, b, c, d, e, f, g\}$ such as :

	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

where 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 $(\{i_1, i_2, i_6\}, \{d\})$ and $(\{i_5, i_6\}, \{e, f\})$ the interaction set attributes of which contain elements that occurred during the same search activity. Thus, our hypothesis suggests that they are likely to be relevant or informative about the information needs associated to the user search activity. On the other hand $(\{i_1, i_2, i_6\}, \{d\})$ or $(\{i_2, i_5, i_6\}, \{e\})$ have interaction set attributes that contain elements that did not occur during the same search activity. Thus these clues should be discarded and conditions to extract only relevant clues are required.

4.1. *Realistic search activity conditions*

We have introduced two kinds of realistic search activity conditions. Let $S = \{i_1, .., i_n\}$ be an interaction stream. The likelihood for a subset of interactions to belong to the same search activity can be modeled through a neighborhood condition. A *neighborhood condition* is a predicat that says, if a subset of interactions I could occurred during the same search activity. It has the fundamental property that if it is true for I then it is also true for any subset $J \subseteq I$. It can be formalized by a boolean measure m defined on the powerset of S. For example, suppose that the neighborhood condition implies that the time span between two consecutive interactions in a subset of interactions never exceeds a fixed threshold of N_1 seconds. If $J = \{j_1, .., j_k\} \subseteq S$ is a subset of interaction of size k (with $k \geq 1$), let us assume that there exists a permutation σ of $(1, .., k)$ such that $J = \{j_{\sigma(1)}, ..j_{\sigma(k)}\}$ is an interaction stream. Then $m(J) = 1$ *iff* :

$$\forall j,\ 1 \leq j < k \quad t_{\sigma(j+1)} - t_{\sigma(j)} \leq N_1$$

Let us take $1 \leq k \leq n$ and denote by C_k the set of elements of $P(\{i_1, .., i_k\})$ which can still be part of a future search activity after interaction i_k. An *ending condition* is intended to state whether a set of interactions belongs to C_k or not. An ending condition has two fundamental properties :

1) At time k, interaction $\{i_k\}$ belongs to C_k,

2) $\forall I \in C_{k+1}$, $I \setminus \{i_{k+1}\} \in C_k$. In other words, the set of future possible search activities only depends on the set of future possible search activities before the last interaction occurred.

A very simple example of ending function is to consider that the only subset of interactions containing an interaction occurred after the last N interactions (with $N \geq 1$) are likely to be in a current or future search activity. Let $S = \{i_1, .., i_n\}$ be an interaction stream and $1 \leq k \leq n$, $K = max(0, k - N)$ then :

$$V(k) = \{X \in P(S)\ :\ \{i_K, i_{K+1} .., i_k\} \cap X \neq \emptyset\}$$

4.2. *Clue Extraction Algorithm*

We have designed the Clue Extraction Algorithm which is an incremental algorithm for finding clues about in an interaction stream. The condition measure F for a subset of interactions to be part of realistic search activity is supposed to be given by a neighborhood condition m (thus $F(I) = m(I)$). CEA extracts the active clues from a stream which are considered to be a good characterization of relevant clues and are defined with respect to an ending condition. Due to space limitation the algorithm is not given here but an extensive description of the algorithm is given in [DEL 03].

5. Evaluation

This section outlines the results of a survey intended to assess the ability of clues to detect user's shifts. Much more details on the preprocessing task and the evaluation process are given in [DEL 03]. A survey was conducted to assess the interest of clues to find relevant features of users' current information needs. In this paper the results related to the issue of detecting users' shifts are presented. Results must be considered as preliminary because the number of survey participants is low (until now 10 participants). A questionnaire with 17 information seeking problems was given to the participants. The focus was put on the shifts defined as the times when the user ends up searching for an answer to a question and starts searching for a new one. Thus, interactions that occurred between two consecutive shifts were intended to satisfy the same information needs. These shifts can easily be identified because they correspond to the transition from one question to another in the search process. In real state the user shifts could be harder to detect if the user's information needs were to change but still be related to the same topics.
Participants were asked to search on the Web answers to the given problems. Questions in the questionnaire were designed so that they involved the users in different information-seeking strategies : for instance, 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 example : "What is the exhibition schedule of the New York Museum of Modern Art ?" In average, it took about 58 minutes to the participants to answer the questionnaire. During this time, they performed 192 interactions in average. Thus, the average number of interactions to answer a question was 11.3 though two questions required about 20 interactions each. The share of search engines result pages, cached documents and homepages over the total number of accessed documents represents 40%. Interactions of the participants' interaction streams were annotated by an expert when identified as shifts.
Let F be a condition measure for a subset of interactions to be part of a realistic search activity. Each interaction can be characterized with respect to the solution set C of the clue enumeration issue. Given an interaction i_k occurring at time k, i_k is *covered* by a clue $(I, W) \in C$ if $F(I \cup \{i_k\}) = 1$. For instance, in the previous example, interactions i_1, i_2, i_3 and i_4 are covered by the clue id 8 and interactions i_3 and i_4 are also covered by the clues id 9 and 11. In other words, i_k is covered by a clue if it occurs while clues about search activities have previously been detected and are still running. An interaction i_k *belongs* to a clue $(I, W) \in C$ if $i_k \in I$. In the previous example, interaction i_2 is covered by the clue id 9 but does not belong to it.

An interaction i_k can then be described with respect to three binary attributes :

1) $hasPreviousActiveClue(i_k) = 1$ *iff* i_{k-1} is covered by a clue.

2) $hasActiveClue(i_k) = 1$ *iff* i_k is covered by a clue.

3) $overlap(i_k) = 1$ *iff* $A_k \cap A_{k+1} = \emptyset$ where A_k is the set of active concepts at time k.

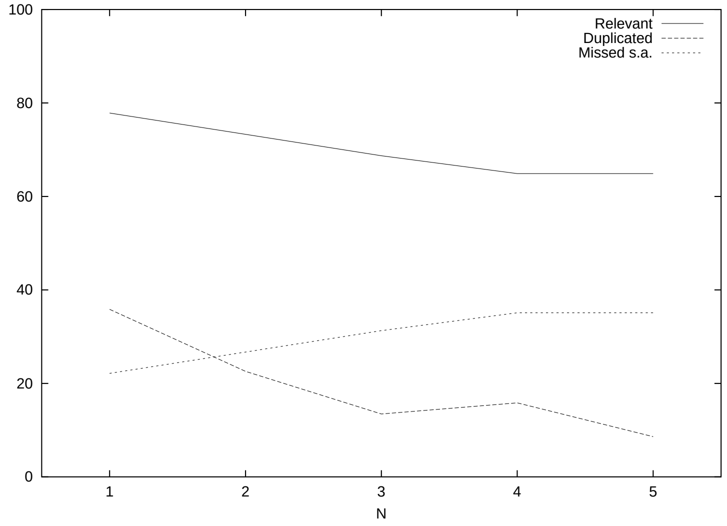

Figure 1. *Shifts detection performance*

An interaction i_k is a *possible shift* if :

$$hasActiveClue(i_k) \wedge (\neg overlap(i_k) \vee \neg hasPreviousActiveClue(i_k))$$

A possible shift is a *relevant shift* if 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 can be considered as 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[2] :

$$\begin{aligned} P &= R + D \\ M &= 17 - R \end{aligned}$$

The tested ending conditions state that the only subsets of interactions containing an interaction that occurred after the last N interactions are likely to be in a current or a future search activity (associated ending measures are formalized in a previous example). The tested neighborhood conditions state that interaction within a clue cannot be spaced of more than a maximal number of M interactions (associated neighborhood measures are formalized in the third example of neighborhood condition subsection). Results are given for $N = M$ and N varies from 1 to 5. $N = M = k$ means that after each interaction, the interaction set attributes of any element in the set of

[2] 17 is the total number of questions in the questionnaire.

active clues contains an interaction occurred among the k last interactions.
Figure 1 gives the average percentages of R, D and M (P can be inferred using the above relations) with respect to the number of questions when 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 ($R + D$). 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.
Though the users where allowed to answer the questionnaire in whatever order, the questions were related to different topic. This explain for a part, the good performance of the CEA to detect shifts. In real state the user shifts could be harder to detect if the user's information needs were to change but still be related to the same topics.

6. Conclusions

This paper addressed the issue of detecting relevant clues about users' current information needs during their navigation on the Web. The originality of the proposed approach lies in the use of the content of the accessed documents to extract these clues. An incremental algorithm intended to extract clues from a user interaction stream was presented. Preliminary results show good ability of this approach to detect user shifts.
Clues are an efficient way to model user information needs. They are a rich source of information about the user because they link his behavior (his interactions) with the content of the accessed documents. For instance, user's interactions could be characterized in term of originality by comparing word set attributes of newly discovered clues to word set attributes of prior clues. User models based on clues can be used to study complex user's behavior such as their information-seeking strategies. Clues can also be used by ANS that require to know the current state of the user search activity and his current information needs.
Future work will focus on the relations between users' information-seeking strategies and the clues that can be extracted from their interaction streams. As mentioned a few times in this document, future work also includes a generalization of the CEA to fuzzy sets.

7. Bibliographie

[BAT 79] BATES M. J., « Information Search Tactics », *Journal of the American Society for Information Science*, , 1979.

[BAT 89] BATES M. J., « The design of browsing and berrypicking techniques for the online search interface », *Online Review*, vol. 13, 1989.

[CAC 01] CACHEDA F., VINA A., « Understanding how people use search engines : a statistical analysis for e-Business », *Proc. of e-Business and e-Work Conf.*, 2001.

[COO 99] COOLEY R., MOBASHER B., SRIVASTAVA J., « Data Preparation for Mining World Wide Web Browsing Patterns », *Knowledge and Information Systems*, , 1999.

[DAV 02] DAVISON B. D., « Predicting Web Actions from HTML Content », *Proceedings of the The Thirteenth ACM Conference on Hypertext and Hypermedia (HT'02)*, College Park, MD, juin 2002, p. 159–168.

[DEL 02] DELORT J.-Y., BOUCHON-MEUNIER B., « Link recommender systems : the suggestion by cumulative evidence approach », *Proc. of STAIRS*, 2002.

[DEL 03] DELORT J.-Y., BOUCHON-MEUNIER B., RIFQI M., « Seeking for Clues About Information Needs in WWW Navigation », *under submission*, 2003.

[ELB 01] EL-BELTAGY S. R., HALL W., ROURE D. D., CARR L., « Linking in context », *Proc. of the twelfth ACM Conference on Hypertext and Hypermedia*, 2001.

[FIN 01] FINKELSTEIN L., GABRILOVICH E., MATIAS Y., RIVLIN E., SOLAN Z., WOLFMAN G., RUPPIN E., « Placing search in context : the concept revisited », *Proc. of the 10th Int'l WWW Conference*, 2001.

[JAN 98] JANSEN B. J., SPINK A., BATEMAN J., SARACEVIC T., « Real Life Information Retrieval : A Study of User Queries on the Web », *SIGIR Forum*, vol. 32, n° 1, 1998, p. 5–17.

[LAU 98] LAU T., HORVITZ E., « Patterns of Search : Analyzing and Modeling Web Query Refinement », *Proc. of the 7th Int'l Conf. on User Modeling*, 1998.

[MAR 95] MARCHIONINI G., *Information Seeking in Electronic Environments*, Cambridge University Press, 1995.

[MOB 00] MOBASHER B., DAI H., LUO T., SUN Y., ZHU J., « Combining web usage and content mining for more effective personalization », *Proc. of the Int'l Conf. on E-Commerce and Web Technologies*, 2000.

[SAL 75] SALTON G., YANG A., WONG C., « A vector space model for automatic indexing », *Communications of the ACM*, vol. 18, 1975.

[SAL 90] SALOMON G. B., « Designing casual-use hypertext : The CHI'89 InfoBooth. », *Proceedings of CHI'90*, 1990, p. 451–458.

[WID 99] WIDYANTORO D. H., IOERGER T. R., YEN J., « An Adaptive Algorithm for Learning Changes in User Interests », *Proc. of the Conf. Int'l on Knowledge Management*, 1999.

[WIL 81] WILSON T. D., « On user studies and information needs. », *Journal of Documentation*, vol. 37, n° 1, 1981, p. 3-15.