



PERGAMON

Information Processing and Management 38 (2002) 727–742

www.elsevier.com/locate/infoproman

**INFORMATION
PROCESSING
&
MANAGEMENT**

Combining evidence for automatic Web session identification

Daqing He¹, Ayşe Göker^{*}, David J. Harper²

*School of Computer and Mathematical Sciences, The Robert Gordon University, St. Andrew Street,
Aberdeen AB25 1HG, Scotland, UK*

Accepted 25 October 2001

Abstract

Contextual information provides an important basis for identifying and understanding users' information needs. Our previous work in traditional information retrieval systems has shown how using contextual information could improve retrieval performance. With the vast quantity and variety of information available on the Web, and the short query lengths within Web searches, it becomes even more crucial that appropriate contextual information is extracted to facilitate personalized services. However, finding users' contextual information is not straightforward, especially in the Web search environment where less is known about the individual users. In this paper, we will present an approach that has significant potential for studying Web users' search contexts. The approach automatically groups a user's consecutive search activities on the same search topic into one session. It uses Dempster–Shafer theory to combine evidence extracted from two sources, each of which is based on the statistical data from Web search logs. The evaluation we have performed demonstrates that our approach has achieved a significant improvement over previous methods of session identification. © 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Session identification; Search context; Dempster–Shafer theory; Web user logs

1. Introduction

With the rapid expansion of the Internet, information retrieval tasks are increasingly performed using Web search engines, which do not have the help of human intermediaries, in contrast to the case in traditional retrieval environments. Unfortunately, from a retrieval perspective, the Web is a vast heterogeneous database covering a large variety of topics at different depths. This poses a

^{*} Corresponding author. Tel.: +44-1224-262713; fax: +44-1224-262727.

E-mail addresses: dqh@scms.rgu.ac.uk (D. He), asga@scms.rgu.ac.uk (A. Göker), djh@scms.rgu.ac.uk (D.J. Harper).

¹ Tel.: +44-1224-262708; fax: +44-1224-262727.

² Tel.: +44-1224-262706; fax: +44-1224-262727.

challenge to the performance of Web retrieval systems in the absence of search intermediaries. A search intermediary was able to establish the context of a user's search for information, and hence advise and guide a user when searching. A search context usually contains the information about the user's search task, the topic areas that the user is and has been interested in, and the features of the search engine. It has been argued forcefully that exploiting the user's context has the potential to improve Web retrieval systems as more information is available about a user and his/her information need (Croft, 1984; Goker, 1997; Talja, Keso, & Pietilainen, 1999).

In this paper, we will present an approach that has great potential for studying Web users' search contexts. Our approach automatically groups a user's consecutive search activities on the same search topic into one session. It uses Dempster–Shafer theory to combine evidence extracted from two sources, each of which is based on the statistical data from Web search logs. The evaluation we have performed demonstrates that our approach has achieved a significant improvement over previous methods of session identification. In addition, further application areas of our approach include other types of Web user studies, such as modelling and learning about Web surfing behaviour.

Before presenting the detail of combining evidence for grouping consecutive search activities into one session, firstly in Section 2, we want to explain why we are interested in session identification, and why session identification has great potential for studying Web users' search contexts.

2. Session identification and contexts

Web users' behaviour is an important resource for inferring users' contextual information. Most studies about Web users' behaviour have been performed through analysing Web user logs (Cooley, Mobasher, & Srivastava, 1999; He & Goker, 2000; Jansen, Spink, Bateman, & Saracevic, 1998; Silverstein, Henzinger, Marais, & Moricz, 1999; Tauscher & Greenberg, 1997). This is because studying the logs is probably the only available method to obtain a large amount of data related to user search behaviour with a rather cheap cost, even though the logs lack general information about users. For the same reason, our studies are based on analysing Web user logs.

When discussing the information retrieval process, often the focus is on the individual activities such as formulating queries, searching document collections and presenting returned documents. However, there are cases where we need to go beyond analysing these individual activities in isolation, and consider the groups of these activities. Spink, Wilson, Ellis, and Ford (1998) show that nearly 60% of users had conducted more than one information retrieval (IR) search for the same information problem. In their research, they refer to the process of repeatedly searching over time in relation to a specific but possibly evolving information problem as the *successive search phenomenon*.

Contextual information plays a more important role in the study of successive searches than that of isolated searches since the contexts behind a series of successive searches are probably closely related to each other, if not the same. However, finding contextual information is a difficult task even for successive searches, especially if the searches are launched on the Web. Previous studies have demonstrated that less information is available about the users and their information needs on the Web, not to mention the fact that Web searches are shorter and search statements

contain less terms than their counter parts in traditional IR searches (Han, Goker, & He, 2001; Jansen et al., 1998).

Our work addresses the problem of paucity of information about users' search contexts within the Web. An individual search activity may be informative sometimes, but a collection of search activities provides much more information about the search topic and the context, especially if they are organised according to their chronological order and are related to the same search topic. We believe that, it is likely that consecutive search activities related to one topic share the same context. It is, therefore, reasonable to say that the information about search topics is an important constituent of the context behind users' searches. This motivates us to *group together search activities related to the same search topic and treat them as a whole during the process of identifying the search contexts*. By *activities*, we refer to the search related actions that take place during the course of information retrieval. These include forming or reforming a query (more precisely a search statement), browsing the results, providing relevance judgements, and so on.³

However, we have to be cautious when grouping search activities. Research on query expansion has found that the performance of the retrieval system could suffer if too much unrelated information is added into the queries (Harman, 1992). As using context to enhance retrieval usually involves refining queries with contextual information, the information from unrelated topics has to be eliminated when constructing the context.

Goker and McCluskey (1991) reported that frequent users in a bibliographic information retrieval system usually have 2–3 topics that they search around. Web users might have a different number of topics, but it is not unreasonable to assume that their searches too will tend to be clustered around their interests or requirements (consciously or unconsciously), and these clusters are probably mixed chronologically. To avoid applying irrelevant contextual information to the retrieval process, there should be mechanisms to mark the boundaries between activities from different topics.

Usually, there are clear signs of the start and the end of sessions in the searches performed in a library with the help from a human intermediary, or performed on an OPAC system with login/logout facilities. These signs help to ease the problem of obtaining information crossing topics in traditional retrieval systems. However, as He and Goker (2000) point out, there has not been a consistent definition of sessions in Web searches, and there is no clear session delimiter in Web searches either, especially in Web search logs. Therefore, we think that avoiding contextual information being obtained from different topic areas becomes an important issue in the approach of enhancing Web retrieval systems with contextual information.

We define a sequence of activities that are related to each other not only through an evolving information need at a deeper, conceptual level but also through close proximity in time as a *session*. If we view a user with an interest in a particular topic as acting in a *role*, we could anticipate that activities in the same session are more likely to correspond to this role. This claim is reasonable, particularly since the retrieval process can be viewed as an interactive problem solving task with a particular goal. Whether there are connections between roles and contexts of different information needs is an open question, but it is reasonable to believe that there are contextual

³ We think it is a little misleading to refer to all activities as queries. This could create a confusion between the initial search statement and any following browsing activities, for example.

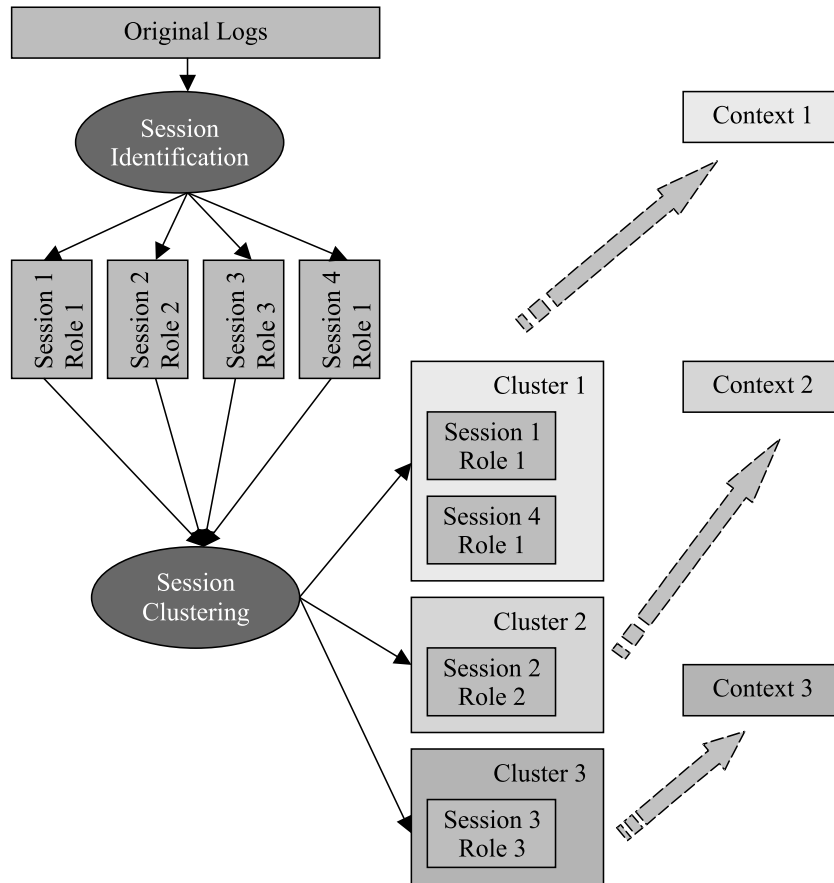


Fig. 1. A scenario of applying the session identification results in inferring the contexts of Web retrievals.

connections among different searches from the same user acting the same role. The contextual information of one search can be helpful for subsequent searches as long as the roles behind them are the same.

As depicted in Fig. 1, our definition of session provides a link between explicit, easy-detected, chronologically occurring search activities and implicit, less definite, slow-evolving contexts. By identifying the session boundaries, we can make sure that the information collected from one session is within the same context, which provides a good foundation for inferring and applying the context. In addition, sessions appearing in different time but with the same context can also be clustered together to provide an even larger collection of the contextual information.

Our approach of session identification combines statistical data from two sources to automatically identify session shifts in Web search logs. By *session shift*, we refer to a *session change happening between two consecutive activities from the same user*. When there is no session change between two consecutive activities, we refer to this as a *session continuation*.

Fig. 2 illustrates the concepts along with the session identification. The two new concepts, the *time gap between activities (GBA)* and *time interval (TI)*, will be described in Section 3.3.

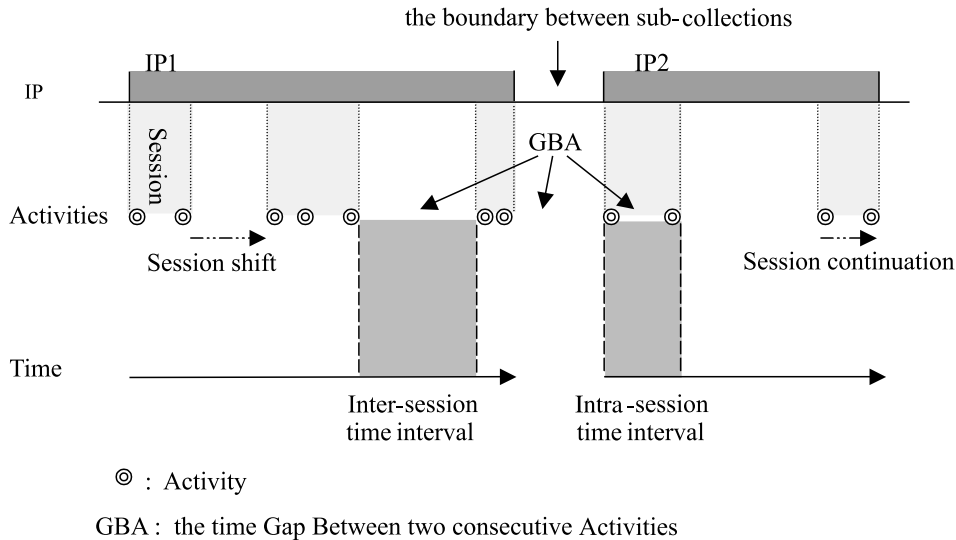


Fig. 2. An illustration of concepts along with session identifications.

The remaining parts of this paper describe our recent work on automatic session identification. We first describe our methodology for detecting session shifts followed by a brief review of related work. We then discuss the evidence combination approach in detail. Finally, we present some results of applying the evidence combination method in our experiments and refer to some future work.

3. Methodology

3.1. The Reuters log collection

The log collection used in this study are transaction records of the Web searches initiated by Reuters (www.reuters.com) Intranet users and we refer it as the Reuters log collection. The search engine used is a local version of AltaVista (www.altavista.com) with both simple and advance search facilities. The time range of the collection was from 15:27:21 on 30th March 1999 to 09:06:27 on 7th April 1999. There are 1440 unique IP addresses but only 1357 of them have their search activities included in the collection,⁴ which consists of 9534 search activities. Each transaction record contains the following three fields:

- *IP address*: an IP address associated with the machine that the activity is initiated from.
- *Time of day*: measured in days, hours, minutes and seconds.

⁴ This is because the search activities of the removed IPs were initiated prior to the starting time of the collection, so only the information pertaining to these search activities were recorded. Hence, these IP addresses are not used in our analysis.

- *CGI command*: a command line to call a CGI program for the search. It includes information about the query terms, the search method (i.e. simple search or advance search), the page number (i.e. subsequent pages for the same search), and so on.

In the Reuters log collection, the search activities were ranked according to the time they were initiated. To facilitate our analysis of the logs, we reordered the activities according to, firstly, the IP addresses and, then, the initiating time when the IP address is the same. The whole collection is an agglomeration of these many sub-collections of activities from the same IP addresses.

At the time this work was undertaken, we had another log collection in addition to the Reuters one. This collection contains transaction records of the Excite search engine (www.excite.com). It has more search activities and a much wider range of IPs, but it covers just 30 min starting from midnight, 9th April 1997. Because our study of successive searches and session identification requires a collection with reasonable time range of users' search activities, we choose the Reuters log collection rather than the Excite one. Of course, we acknowledge that Reuters Intranet users may only represent a small portion of Internet users, and to this extent, future work includes an analysis of larger user population.

During our previous studies on the Reuters logs, we had three search experts examine the logs and mark the places where they thought a session shift occurred. To reduce the misjudgements due to a possible lack of information about the users' information needs, we asked people from Reuters to help us to resolve ambiguities in identifying the transaction of sessions. For example, Fig. 3 shows a session shift between the third and fourth search activities from a particular IP address.

We divided the Reuters logs into two sections. The first part has been used for obtaining statistical data for session identification. It contains 4860 query activities from 661 IP addresses. The other part which contains 4674 query activities from 696 IP addresses was used for testing the approach. In this paper, they will be referred to as *the training collection* and *the test collection*.

3.2. Our method

Previous research has pointed out that the study of Web users is difficult for several reasons. Firstly, it lacks adequate background information about Web users and, secondly, the population and the diversity of Web users are very large (Han et al., 2001; Jansen et al., 1998). The facts related to the first point are: *the queries are short and most users do not register so there is hardly any knowledge of their background* (Han et al., 2001; Jansen et al., 1998). The second point is supported by the fact that there are millions of Web users from all over the world who can access

IP1	06/Apr/1999:14:37:05	official+reuters+font
IP1	06/Apr/1999:14:37:14	reuters+font
IP1	06/Apr/1999:14:40:30	branding
----- Session Shift -----		
IP1	06/Apr/1999:14:56:09	TIP
IP1	06/Apr/1999:14:56:24	technical+information+page

Fig. 3. Samples from the Reuters logs. The actual IP address is not presented for reasons of privacy, and only query parts of the CGI commands are listed for the simplicity of the display.

the search engines at any point in time with a variety of methods (Kehoe, Pitkow, & Morton, 1999). Both of these make many studies of Web users through the direct involvement of human subjects either too expensive to operate or too narrow to be representative of Web users in general.

Our study, therefore, relies extensively on the analysis of transaction logs of search engines. Although Web user information is limited to that recorded in the logs, this nevertheless provides a means to examine a great number of Web users cheaply. We anticipate that often there will be a limited resource of information pertaining to Web users and an irregularity due to their various backgrounds, regions and methods of accessing search engines, so we adopt a statistical approach rather than a hard coded rule based one.

3.3. Time interval

We call the time difference between two consecutive search activities the GBA. The duration of a GBA is called the TI of the GBA. Previous work has shown correlation between the time interval of GBAs and session shifts (Goker & He, 2000; He & Goker, 2000). That is, the longer the time interval is, the higher the likelihood that there is a session shift happening in the GBA.

To use time intervals as a piece of evidence in session identification, a threshold is predefined so that when the time interval of a GBA is larger than the threshold, the GBA is marked to contain a session shift. For simplicity, we call the threshold the *session interval* to demark session identification (He & Goker, 2000).

In this paper, however, we want to use time interval in a different way. Instead of trying to find one particular session interval that would identify most session shifts, we calculate *the probability of a particular GBA containing a session shift given its time interval*. Such information then can be combined with other probabilities to identify session shifts.

We divided the time span of the training collection into seven sections (see Table 1). Six of them cover the time intervals up to and including 30 min, i.e. the first is from 0 up to and including 5 min, the next from 5 to 10 and so on. The seventh section contains all the time intervals longer than 30 min. This arrangement of sections is due to the results of our previous studies which indicate that most time intervals are smaller than 15 min and 10–12 min is a more likely time span to find the best session interval for session identification (Goker & He, 2000).

Table 1
Statistical data based on time interval

TI sections (min)	Intra-session F_{TI}	Inter-session F_{TI}	$P(TI)$	$P(\text{shift} TI)$	$P(\text{contin} TI)$
0–5	3264	173	0.8185	0.0503	0.9497
5–10	108	31	0.0332	0.2230	0.7770
10–15	36	25	0.0145	0.4098	0.5902
15–20	30	19	0.0117	0.3878	0.6122
20–25	15	15	0.0071	0.5000	0.5000
25–30	13	11	0.0057	0.4583	0.5417
30+	108	351	0.1093	0.7647	0.2353
Total	3574	625			

The first column of Table 1 shows the time spans of the seven time interval sections. The units are minutes. The second and third columns show the frequency of time intervals between two activities within a session (intra-session) and across two sessions (inter-session), respectively. The column of $P(\text{TI})$ displays the distribution of time intervals in each of the seven sections regardless of whether a GBA is within a session or across sessions. The last two columns present the conditional probabilities (using Bayes theorem on training collection) of having a session shift (i.e., $P(\text{shift}|\text{TI})$) or a session continuation (i.e., $P(\text{contin}|\text{TI})$) given the evidence of a time interval belonging to a particular section, respectively. If the reader calculates the total number of intra-plus inter-session TIs in the last row, s/he will find that the number is equal to the difference between the total number of activities in the training collection and the number of IP addresses in the training collection. This is due to the fact that the collection is an agglomeration of many individual sub-collections.

3.4. Search patterns

Many sessions contain several activities. If looking at the relation between two consecutive activities a_i and a_{i+1} from the same user, we can summarise some mutually exclusive classes of search patterns that users could have used in their searches. We associate these patterns with a GBA between two consecutive activities from the same IP to facilitate the session identification.

- *Browsing*: the second activity a_{i+1} requests for another set of results on the same query.
- *Generalisation*: the second activity a_{i+1} is on the same topic as the first one a_i , but seeking more general information.
- *Specialisation*: the second activity a_{i+1} is on the same topic, but seeking more specific information.
- *Reformulation*: the second activity a_{i+1} is on the same topic, but at least part of both queries in the two activities are different.
- *Repetition*: the second activity a_{i+1} is the same as the first one, but the pattern is not browsing.
- *New*: the second activity a_{i+1} is on different topics.
- *Relevance feedback*: the second activity a_{i+1} is generated by the system when the user selects the choice of “related pages”.
- *Others*: activity a_i does not contain a query so cannot be allocated to any of the search patterns above.

Since a session is associated with queries from the same role on the same topic, theoretically only GBAs marked with *New* should be related to a session shift, and all other search patterns should be associated with two activities from the same session.⁵ Therefore, information of search patterns would appear to be ideal evidence for session identification. However, due to the lack of adequate information in the logs, there are errors in the process of automatic assignments of search patterns. For example, new activities can easily be mixed up with reformulation activities.

⁵ We acknowledge that it is not that obvious for the *Others* pattern. For simplicity, we treat it the same as other patterns except *New*.

Algorithm Search Pattern Identification:

Input: Activities A_i with query Q_i and page number P_i , and A_{i+1} with query Q_{i+1} and page number P_{i+1}

Local: $B = \{t \mid t \in Q_i \wedge t \in Q_{i+1}\}$ // terms in common;
 $C = \{t \mid t \in Q_i \wedge t \notin Q_{i+1}\}$ // terms which appear in Q_i only;
 $D = \{t \mid t \notin Q_i \wedge t \in Q_{i+1}\}$ // terms which appear in Q_{i+1} only;

Output: Search Pattern SP

begin

$SP = \textit{Others}$; // default value

if ($P_i + 1 == P_{i+1}$) **then** $SP = \textit{Browsing}$;

elseif ($Q_i == \phi$) **then** $SP = \textit{Others}$;

elseif ($Q_{i+1} == \phi$) **then** $SP = \textit{Relevance Feedback}$;

elseif ($Q_i == Q_{i+1} \wedge SP \neq \textit{Browsing}$) **then** $SP = \textit{Repetition}$;

elseif ($B \neq \phi \wedge C \neq \phi \wedge D = \phi$) **then** $SP = \textit{Generalisation}$;

elseif ($B \neq \phi \wedge C = \phi \wedge D \neq \phi$) **then** $SP = \textit{Specialisation}$;

elseif ($B \neq \phi \wedge C \neq \phi \wedge D \neq \phi$) **then** $SP = \textit{Reformulation}$;

elseif ($B = \phi$) **then** $SP = \textit{New}$;

end

Fig. 4. The algorithm used for identifying the search patterns.

Therefore, we assign a probability value to each type of search pattern when it is used in automatic session identification to indicate the chance of containing a session shift given the class.

We have developed an algorithm that can automatically assign a search pattern class to a given GBA. It analyses the query part and the page number part of the CGI command line of each search activity if two consecutive activities are from the same IP address. Most of the time, the algorithm can identify the search pattern by examining the terms in the two queries. These terms are not stemmed because we did not find a significant difference between using original terms and using stemmed terms. The detail of this algorithm can be found in Fig. 4.

Although this algorithm seems primitive, it has achieved high accuracy in assigning the *Browsing*, *Generalisation*, *Specialisation*, *Repetition* and *Relevance feedback* patterns when it is applied to the training collection, but is less accurate when handling *Reformulation*, *New* and *Others*.

Table 2 shows the results of this module applied to the training collection. The second column displays the numbers of GBAs within a session that are assigned to each search pattern, whereas

Table 2
Statistical data based on search pattern

SP classes	Intra-session F_{SP}	Inter-session F_{SP}	$P(SP)$	$P(\text{shift} SP)$	$P(\text{contin} SP)$
Browsing	2024	0	0.4820	0	1
Generalisation	121	0	0.0288	0	1
Specialisation	286	0	0.0681	0	1
Reformulation	280	15	0.0703	0.0508	0.9492
Repetition	404	0	0.0962	0	1
New	312	605	0.2184	0.6587	0.3413
Relevance feedback	9	0	0.0021	0	1
Others	138	5	0.0341	0.0350	0.9650

the third column presents similar information but for inter-session GBAs. The column of $P(\text{SP})$ shows the distribution of GBAs in each search pattern among all the GBAs that have been given a pattern. The last two columns display the probabilities of having a session shift $P(\text{shift}|\text{SP})$ and that of a session continuation $P(\text{contin}|\text{SP})$ given each pattern, respectively.

In Section 5, we will discuss how these two sources of evidence, namely “*time interval*” and “*search pattern*” evidence can be combined. Prior to that, however, we will present some related work in the literature.

4. Related work

The related work lies in several areas of Web retrieval research. These include studies on users’ retrieval activities, on users’ navigation behaviours, on search patterns and on evidence combination. We are going to review the first three areas in this section, and leave the last one in Section 5 after the brief description of Dempster–Shafer theory.

Studies on users’ retrieval activities through analysing Web logs explicitly or implicitly group all activities for one user (if registered) or one IP address into a session (Jansen et al., 1998; Jones, Cunningham, & McNab, 1998). The appropriateness of grouping these activities under one session is debatable in our situation, particularly where the time span is large. Additionally, one could argue that the final cut-off point for logs is usually arbitrary and the logs could just as well be split into several different batches.

Two studies about Web users’ navigation behaviour, Catledge and Pitkow (1995) and Cooley et al. (1999), adopt time interval as their means for dividing page accesses from each user into individual sessions. In their work, they call time interval *timeout*. Catledge and Pitkow (1995) found a 25.5 min time interval based on their user experiments in 1994. However, they do not refer to the reasons behind choosing that particular time interval, and in any case the users’ navigation patterns may have changed over the last seven years. More importantly, their work is about users’ navigation behaviour, and does not include activities derived from using Web search engines.

Lau and Horvitz (1999) define search patterns as *query refinement classes*. They have identified seven classes but they mix *Relevance feedback* with the class of *Others* and do not include a *Browsing* class. Their work included manually tagging classes to search activities, and did not use the classes as a resource for session identification.

5. Using evidence combining in automatic session identification

5.1. Brief description of Dempster–Shafer theory

Dempster–Shafer theory is a method of inexact reasoning. It is based on work done originally by Dempster then extended by Shafer (1976). The theory assumes that there is a fixed set of mutually exclusive and exhaustive hypotheses or propositions called the *frame of Discernment* Θ .

Each hypothesis or proposition has a degree of belief called a *basic probability assignment* or *mass function* m and has the properties:

$$m(\phi) = 0 \quad \text{and} \quad \sum_{X \subseteq \Theta} m(X) = 1.$$

Dempster–Shafer theory uses the amount of belief that has not been committed to the subsets of Θ as the basic probability assignment to set Θ . That is,

$$m(\Theta) = 1 - \sum_{X \subset \Theta} m(X).$$

The rule combining evidence is called *Dempster’s rule*. Suppose we are interested in combining evidence for hypothesis H , and we have two independent sources of evidence m_1 and m_2 , Dempster’s rule tells us that their combination can be defined as follows:

$$m_{1 \& 2}(H) = \frac{\sum_{X, Y \subseteq \Theta, X \cap Y = H} m_1(X) m_2(Y)}{1 - \sum_{X, Y \subseteq \Theta, X \cap Y = \phi} m_1(X) m_2(Y)}.$$

When there is no uncommitted confidence left for each evidence, e.g. $m(\Theta) = 0$, Dempster–Shafer theory will be equal to the probability theory of calculating the probabilities of two independent events.

5.2. Combining evidence by using Dempster–Shafer method

The evidence combining method that we propose is not new in the information processing community. Both Jose (1998) and Aslandogan and Yu (2000) have used Dempster–Shafer method in image retrieval, where the evidence from both images and captions are combined to improve the performance of their retrieval systems. *Diogenes*, the image retrieval system Aslandogan and Yu (2000) built, comfortably outperformed some well-known commercial and research prototype image search engines. Even though the Dempster–Shafer method has demonstrated its advantage, as far as we know, it has not been used in session identification yet.

We have two sources of evidence: the evidence relating to time intervals, annotated by M_{TI} , and the evidence relating to search patterns, annotated by M_{SP} . Since the analysis of time intervals does not affect the assignment of search patterns and vice versa, the independence assumption of the two pieces of evidence required by the theory holds.

When Jose (1998) and Aslandogan and Yu (2000) applied Dempster–Shafer theory in image retrieval, their task was to decide whether or not an image in a collection is relevant to the searcher’s information need. In that application, the score of an image is calculated across the collection. Therefore, it is reasonable to transfer the uncommitted belief for the relevance of an image to the judgements of other images in the collection. As a result, suppose there is a collection of n images, the frame of discernment in their approach has been defined as $\Theta = \{R_1, \dots, R_n\}$, where R_i : *Image i is relevant to the searcher’s information need*. The judgement on each image then has to wait until the scores of others images have been calculated.

The situation is different in our application. Although the judgement on whether a GBA contains a session shift or a session continuation could depend on the judgements of other GBAs, the two pieces of evidence used in our current approach relate solely to a single GBA. That is, the

judgement on one GBA in our current approach *would not affect* the chances of other GBAs containing a session shift. Therefore, it is appropriate to keep our belief of the judgement on one GBA at that GBA, and spread the belief between the commitment of the GBA containing a session shift and that of it containing a session continuation. As a result, the frame of discernment in our approach is $\Theta = \{P_s, P_c\}$, where

P_s : GBA S_i contains a session shift.

P_c : GBA S_i contains a session continuation.

This way of defining the frame of discernment allows the judgement on each GBA in our approach to be made immediately after the probability of that GBA is calculated. This has advantages in real time applications of automatic session identification, such as adaptive information retrieval systems. In these systems, the decision as to whether the current query is within the same session as previous queries has to be made on-the-spot in order to achieve adaptation.

The analysis of each evidence generates a set of conditional probabilities associated with the two propositions (see Sections 3.3 and 3.4). When converting the probabilities into the basic probability assignments, a weight W_i is used to express our confidence of using that evidence in session identification. The basic probability assignment is then the product of the conditional probability and the corresponding confidence weight W_i :

$$m_i(P_s) = P(\text{shift}|i) * W_i,$$

$$m_i(P_c) = P(\text{contin}|i) * W_i,$$

where i is TI or SP.

The confidence weight allows some part of probabilities being reserved in the basic probability assignment associated with Θ , and can be transferred to other evidence. The calculation of $m_i(\Theta)$ are

$$m_{\text{TI}}(\Theta) = 1 - m_{\text{TI}}(P_s) - m_{\text{TI}}(P_c),$$

$$m_{\text{SP}}(\Theta) = 1 - m_{\text{SP}}(P_s) - m_{\text{SP}}(P_c).$$

Since there are only two propositions in Θ , Dempster rule in our situation can be written as follows:

$$m_{\text{TI}\&\text{SP}} = \frac{m_{\text{TI}}(P_s)m_{\text{SP}}(P_s) + m_{\text{TI}}(P_s)m_{\text{SP}}(\Theta) + m_{\text{TI}}(\Theta)m_{\text{SP}}(P_s)}{1 - (m_{\text{TI}}(P_s)m_{\text{SP}}(P_c) + m_{\text{TI}}(P_c)m_{\text{SP}}(P_s))}.$$

To convert the combined scores into binary decisions, we introduce a threshold T_{shift} . A GBA is marked to contain a session shift when its score exceeds the threshold. The algorithm of automatic session identification now can be defined as in Fig. 5.

5.3. Training results

After the basic statistical data has been gathered, the performance then depends on three parameters: the two confidence weights W_{TI} and W_{SP} and the threshold T_{shift} .

To measure the performance of the algorithm, we have adapted a set of commonly used measures⁶ (Yang, 1999) in the context of session identification. They are:

⁶ The reference mentioned contains a survey of measures developed by other researchers.

```

Algorithm Automatic Session Identification:
begin
for (each GBA) do
  begin
    // if the GBA is between two sub-collections
    if (IP address changes) then mark IP switch;
    // else find a session shift
    elseif ( $m_{TI \&SP}(P_s) > T_{shift}$ ) then mark shift;
    // else find a session continuation
    else mark contin;
  end
end
end

```

Fig. 5. The algorithm for combining evidence in session identification.

- *Precision*:

$$P = \frac{N_{\text{shift \& correct}}}{N_{\text{shift}} + N_{\text{contin}}},$$

where N_{shift} and N_{contin} are the numbers of GBAs that are marked to have a session shift and a session continuation, respectively, whereas $N_{\text{shift \& correct}}$ represents the number of GBAs that are correctly marked to have a session shift.

- *Recall*:

$$R = \frac{N_{\text{shift \& correct}}}{N_{\text{true_shift}}},$$

where $N_{\text{true_shift}}$ means the number of GBAs marked to have a session shift by human experts.

- F_β :

$$F_\beta = \frac{(1 + \beta^2)PR}{\beta^2P + R},$$

where β is a weight to control of the emphasis on precision or recall.

We use F_β as the measure of the performance because it considers both precision and recall simultaneously. In the equation, β is set to be 1.5 as we want more emphasis on the recall measure. A high recall corresponds to less sessions being incorrectly grouped together. There are two types of error in session identification. Activities on the same topic could be wrongly divided into different sessions (a *Type A* error). Alternatively, activities on different topics could be incorrectly grouped together into one session (a *Type B* error). Since we want to infer users' contextual information from the identified sessions, *Type B* errors would do more damage than *Type A* errors.

The training was performed with a Genetic Algorithm adapted from Mitchell (1997, p. 251), in which the two confidence weights and the threshold are treated as three genes of a chromosome. The population of the chromosomes was set to be 50, the fitness function was the value of F_β measure, the number of total iterations was 500, chromosomes were probabilistically selected for crossover and the crossover rate was 0.02 with the mutation rate 0.01. Our training results show that the performance of the session identification approach achieves the best F_β measure of 0.8519 (e.g. $P = 0.6543$ and $R = 0.9840$) when the threshold $T_{\text{shift}} = 0.345$ and the two confidence weights are $W_{\text{TI}} = 0.864$ and $W_{\text{SP}} = 0.936$.

Table 3

The evaluation results of the new approach and the approach using time interval only

	Precision	Recall	Type A	Type B	F_{β}
New approach	0.5955	0.9815	396	11	0.8183
SP alone	0.5985	0.9663	385	20	0.8127
TI = 8.6 min	0.6532	0.6818	215	189	0.6727
TI = 10 min	0.6834	0.6650	183	199	0.6706
TI = 12 min	0.7063	0.6397	158	214	0.6588
TI = 15 min	0.7335	0.6162	133	228	0.6481

6. Evaluation and discussion

We evaluated the approach on the test collection to examine its performance. The results indicate that our approach performed very well in the test collection (see Table 3). The high recall and good precision values reported above in the last section also carries over to the test collection. Since our approach can provide very high value at recall, the identified sessions by this means have good accuracy to be applied in modelling Web users and their contextual information.

We also performed a comparison study between the approach of combining evidence and the approach that users' time interval only for session identification. The time intervals used in the evaluation are in the range 8.6–15 min as Goker and He (2000) find that using time intervals between 10 and 12 min can produce the lowest number of errors. Table 3 shows the results of the comparison. Clearly, the evidence combining approach performs significantly better than the method using time interval only. However, if search pattern information alone is used then the improvement is not as great. Search pattern information by itself can provide good results. In this particular case focusing on accurately identifying the *New* search activities would be an appropriate strategy. However, small improvements can nevertheless be gained by including time pattern information and there may be situations where this is important.

7. Conclusion

In this paper, we have described a principled approach to identifying user sessions based on two sources of evidence *time interval* and *search pattern* obtained from analysing a large batch of Web search logs. We explained our motivation of this work, that is, to facilitate the usage of contexts behind Web users' searches. Our method has been evaluated on the test part of the Reuters log collection and has been compared with a frequently used method based on predefined time intervals. The results demonstrate that this approach can achieve the desired high recall measure accompanied by a reasonable precision measure. This means that the identified sessions provide good sources for analysing the contexts.

Further experiments are planned on an Internet search log collection to verify the performance of this method on a wider range of user population. In addition, we are investigating the use of clustering techniques to achieve a better understanding of the similarity between two consecutive activities. The last but not least step is to analyse and exploit users' search contexts based on our results of session identification, and deploy users' contexts in improving Web searches.

Acknowledgements

This work is supported in part by EPSRC GR/R11742, in which Ayşe Göker is the principle investigator. We are grateful to Reuters Ltd. for providing the Web transaction logs. Many thanks to Dr. Bill Teahan and Dr. Andrei Petrovski for their comments.

References

- Aslandogan, Y. A., & Yu, C. T. (2000). Multiple evidence combination in image retrieval: diogenes searches for people on the web. In N. J. Belkin, P. Ingwersen, & M. Leong (Eds.), *Proceedings of the 23rd annual international ACM SIGIR conference* (pp. 88–95).
- Catledge, L., & Pitkow, J. (1995). Characterizing browsing strategies in the World-Wide Web. In *3rd international World-Wide Web conference WWW95*. Available: http://www.igd.fhg.de/archive/1995_www95/papers/.
- Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems, 1*, 5–32.
- Croft, W. B. (1984). The role of context and adaptation in user interfaces. *International Journal of Man–Machine Studies, 21*, 283–292.
- Göker, A. (1997). Context learning in Okapi. *Journal of Documentation, 53*(1), 80–83.
- Göker, A., & He, D. (2000). Analysing Intranet logs to determine session boundaries for user-oriented learning. In *AH2000: Proceedings of the international conference on adaptive hypermedia and adaptive web-based systems* (pp. 319–322). Trento, Italy: Springer.
- Göker, A., & McCluskey, T. (1991). Towards an adaptive information retrieval system. In Z. W. Ras, & M. Zemankova (Eds.), *Methodologies for Intelligent Systems, 6th International Symposium, ISMIS'91* (pp. 348–357), Springer-Verlag.
- Han, S., Göker, G., & He, D. (2001). Web user search pattern analysis for modelling query topic changes. In *Proceedings of the user modeling for context-aware applications, a workshop of the 8th international conference on user modeling*.
- Harman, D. (1992). Relevance feedback revisited. In *Proceedings of the ACM-SIGIR'92*.
- He, D., & Göker, A. (2000). Detecting session boundaries from Web user logs. In *Proceedings of the BCS-IRSG 22nd annual colloquium on information retrieval research, Cambridge, UK* (pp. 57–66).
- Jansen, M. B. J., Spink, A., Bateman, J., & Saracevic, T. (1998). Real life information retrieval: a study of user queries on the web. *SIGIR FORUM, 32*(1), 5–17.
- Jones, S., Cunningham, S., & McNab, R. (1998). An analysis of usage of a digital library. In C. Nikolaou, & C. Stephanidis (Eds.), *Lecture Notes in Computer Science* (Vol. 1513). *Proceedings of the second European conference on digital libraries* (pp. 261–277). Berlin: Springer.
- Jose, J. M. (1998). An integrated approach for multimedia information retrieval. PhD Thesis, School of Computer and Mathematical Sciences, The Robert Gordon University, Aberdeen, Scotland.
- Kehoe, C., Pitkow, J., & Morton, K. (1999). GUV's 10th WWW user survey. Graphics Visualization and Usability Center, Georgia Tech Research Center, Atlanta. Available: http://www.gvu.gatech.edu/user_surveys.
- Lau, T., & Horvitz, E. (1999). Patterns of search analyzing and modeling web query refinement. In *Proceedings of the user modeling conference* (pp. 119–128).
- Mitchell, T. (1997). *Machine learning*. Singapore: McGraw-Hill.
- Shafer, G. (1976). *A mathematical theory of evidence*. Princeton, NJ: Princeton University Press.
- Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum, 33*(1), 6–12.
- Spink, A., Wilson, T., Ellis, D., & Ford, N. (1998). Modeling users' successive searches in digital environments. *D-Lib Magazine*.
- Talja, S., Keso, H., & Pietilainen, T. (1999). The production of 'context' in information seeking research: a metatheoretical view. *Information Processing and Management, 35*, 751–763.

- Tauscher, L., & Greenberg, S. (1997). How people revisit web pages: empirical findings and implications for the design of history systems. *International Journal of Human–Computer Studies*, *47*, 97–137.
- Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval*, *1*(1–2), 67–88.