Detecting Search Sessions Using Document Metadata and Implicit Feedback

Tomáš Kramár
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology
Bratislava, Slovakia
kramar@fiit.stuba.sk

Mária Bieliková
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology
Bratislava, Slovakia
bielik@fiit.stuba.sk

ABSTRACT

It has been shown that search personalization can greatly benefit from exploiting user's short-term context – user's immediate need and intent. However, this requires that the search engine must be able to divide user's activity into segments, where each segment captures user's single goal and focus. Several different approaches to search session segmentation exist, each considering different features of the queries, but it may be helpful to also consider user's implicit feedback on the search results clicked in response to the query. We propose a method for segmenting queries into search sessions which is based on document metadata and incorporates implicit feedback. Our approach also considers multitasking, where user shifts her current interest, but afterwards proceeds with the original task. We evaluated our approach on manually segmented query log and compared the results of our approach with results from other methods and showed that using implicit feedback can improve the performance of the segmentation task.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Query formulation, Selection process, Search process

General Terms

Information retrieval

Keywords

personalization, search, short-term contexts, search sessions, search session segmentation

1. INTRODUCTION

Web contains an ever-growing amount of documents. Accessing these document poses great difficulties, especially after the rise of Web 2.0 where users have been given the ability to create the content, which allows for more social-based approaches to personalization [1], but also contributes to information overload. According to Technorati1 , a service which tracks user-generated media, the content on the Web is growing with a pace at 2 blog posts per second, and this number does not include the growth of other content.

Search engines play a crucial role in accessing this amount of content. Users interact with search engines by entering few keywords, which describe their intent and expect the machine to provide a list of relevant documents. This model has several known disadvantages:

- the number of keywords is usually low, typically 1-3 keywords [17] and this often leads to ambiguity and unclear intent;
- many of the words are ambiguous; a word "jaguar" can refer to an animal, a car and even has less-known meanings such as a game console or German battle tank;
- the queries are almost never accurate [9], they are either too generic or too specific, but almost never exactly aligned with the specific intent the user has in mind.

The combined impact of these problems leads to a conclusion that finding the relevant document when we do not have enough information about it is indeed a difficult task, both for the user and the search engine.

To mitigate this problem, several approaches to search personalization have been researched, each with the ultimate aim to help users find the relevant content, without trying to change how humans think, or work.

There is relevance feedback, query expansion, search intent detection, alternative ranking schemes and many others. These techniques leverage and act upon some form of search context. Generally, the term context refers to attributes of the environment [7], such as location, time, or weather, but in the domain of personalized search the term is commonly used to describe user's needs, goals and intents (e.g. [21, 27]). Based on the time span that is used to build the search context, the context may be long, or short-term.

Long-term search context is composed of the goals and intents that can be recognized by observing the complete user activity, beginning with the first known information about the user and her activity.

Short-term search context is composed of the goals and intents that the user has in the moment of search. These represent the current focus and are obtained by observing the user activity beginning in a recent point of time.

To be able to use short-term search context a personalization system must know the exact moment the user changes her intent, so that it can start and use a new context. The task of detecting this change is referred to as search session detection (segmentation). The term search session was never formally defined in the literature and its meaning differs in different works. In this work, we assume that search session is a sequence of search related actions with the single underlying informational intent, similarly to [22].

The goal of search session segmentation is to partition the stream of user queries into segments of queries, where each segment is the search session, i.e., holds the condition that all queries that it contains are related to a single underlying goal.

Several existing approaches to search session segmentation exist, but they have various disadvantages. When a segmentation approach acts solely on the features provided by the query itself, the amount of understanding of the underlying intent is quite limited. Therefore, many approaches also consider the documents that were clicked in response to the query, but these approaches do not evaluate user's feedback that is implicitly left in each document. Many existing approaches also do not consider interruptions in Web browsing and multitasking (i.e. having multiple intents).

In this work we aim to contribute to the area of search session segmentation by matching the queries with the metadata of the documents clicked from the search results to get better insight into the purpose of the query by aggregating more data than only the query itself provides. We also evaluate the level of page usefulness for the particular query by collecting and analyzing the implicit feedback indicators that the user provides for each page view. Our approach also considers user interruptions and is able to separate intermingled sessions and reconnect interrupted sessions.

The paper is structured as follows. Section 2 describes the related work done in the area of search session segmentation. In Section 3 we describe our data collection methodology. Section 4 describes the proposed method for search sessions segmentation. Experimental results are described in Section 6.

2. REFERENCES