Journal of Visual Languages and Computing 31 (2015) 230–236

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

journal homepage: www.elsevier.com/locate/jvlc

Understanding user intent on the web through interaction mining ☆

Loredana Caruccio *, Vincenzo Deufemia, Giuseppe Polese

Department of Computer Science, University of Salerno, 84084 Fisciano (SA), Italy

ARTICLE INFO

Available online 30 October 2015

Keywords:
User intent understanding
HCI features
Web search
User behavior mining
Query classification

ABSTRACT

Predicting the goals of internet users can be extremely useful in e-commerce, online entertainment, and many other internet-based applications. One of the crucial steps to achieve this is to classify internet queries based on available features, such as contextual information, keywords and their semantic relationships. Beyond these methods, in this paper we propose to mine user interaction activities to predict the intent of the user during a navigation session. However, since in practice it is necessary to use a suitable mix of all such methods, it is important to exploit all the mentioned features in order to properly classify users based on their common intents. To this end, we have performed several experiments aiming to empirically derive a suitable classifier based on the mentioned features.

1. Introduction

During an Internet navigation session the user performs several actions that can provide hints on his/her future activities. Being able to capture and interpret the hidden goals behind such actions can provide organizations with a competitive advantage. For instance, e-commerce organizations might predict user needs, and advertise the products that users will most likely buy, also through the mining of previous customers purchase strategies [1]. Thus, multimedia catalogues, web and information retrieval systems need to embed search engines capable of capturing user intent, which is the focus of user intention understanding (UIU) research area [2].

Many approaches for user intent understanding are based on the analysis of search behaviors [3–5], such as clicked URLs [6] and submitted queries. Most of them aim to capture semantic correlations among search behaviors of the same user, in order to let search engines produce customized results for each individual user.

Other studies analyzed user interactions with *Search Engine Result Pages* (SERPs) to infer their intent [7–10]. However, by limiting the analysis to results contained in a SERP, such methods ignore many important interactions and contents visited from such results. For this reason, some approaches to user behavior analysis focus on user interactions with web pages to infer clues on their interest and satisfaction with respect to the visited contents [11–13]. Following this trend, in this paper we define a new model for UIU analyzing both interactions with SERP results and those on the visited web pages. The interaction features considered in the proposed model are local page level statistics, that is, they are fine-grained and refer to portions rather than the whole web pages. This provides the basis for a more promising prediction of the user intent, since several experiments with eye-trackers revealed that users analyze web pages by sections, overlooking those of low interest [14].

☆ This paper has been recommended for acceptance by Henry Duh.

* Corresponding author.
E-mail addresses: lcaruccio@unisa.it (L. Caruccio), deufemia@unisa.it (V. Deufemia), gpolese@unisa.it (G. Polese).

http://dx.doi.org/10.1016/j.jvlc.2015.10.022
1045-926X/

Other than interaction features, the proposed model considers additional features, such as query keywords and contextual information, all feeding a classification algorithm to understand user intent. The classification process uses a two-level taxonomy in which the first level defines *navigational*, *informational*, and *transational* types of queries [15], where the last two are further decomposed in the second level [16].

We also provide experimental results highlighting the efficiency of the proposed model for query classification. The proposed set of features has been evaluated with several classification algorithms. To this end, in order to more precisely compare the achieved results, and detect the most promising features, we have introduced a metric to evaluate the performances of the different classifiers.

The rest of this paper is organized as follows. In Section 2, we provide a review of related work. Then, we present the model exploiting interaction features for UIU in Section 3. Section 4 describes experimental results. Finally, conclusions and future work are given in Section 5.

2. Related work

As said above, many approaches for user intent understanding analyze search behaviors of users while they navigate and submit queries through the web [3–5].

In the early '90s, a pioneer study on search behaviors focused highlighted three browsing strategies [17]: *scan browsing*, in which new information is scanned based on its relevance to changing tasks, representing transient browse goals; *review browsing*, in which, with respect to scan browsing, the scanned information is also reviewed and integrated; finally, *search-oriented browsing*, in which the new information is scanned, reviewed, and integrated based on its relevance to a fixed task.

Morrison et al. proposed three taxonomic classification schemes based on user responses to web activities that significantly impacted on their decisions and actions [18]. In particular, they formalized the main questions users ask themselves before starting a search session: *why*, *how*, and *what*, which represent the primary purpose of the search, the method used to find the information, and the content of the searched information, respectively, yielding three different taxonomies.

Sellen et al. extended previously defined taxonomies by extensively monitoring user search activities [19]. They ended up with a classification dividing web activities into six categories, in which two new types were introduced: *transacting* and *housekeeping*. The first concerns using the web to execute secure transactions targeted at products and services, such as ordering a product or filling out a questionnaire. The second concerns using the web to check or maintain the accuracy and functionality of web resources.

A taxonomy focusing on search queries has been defined by Broder [15], who identified the following three classes of queries based on user's intent: *navigational*, aiming to reach a particular web site, *informational*, aiming to collect information from one or more web pages, and *transactional*, aiming to perform some web-mediated activities, that is, to reach a web site where some service is offered, and from which further interactions are expected.

Kang et al. focused on analyzing two types of search activities [9]: *topic relevance*, that is, searching documents guided by a given topic, of informational type, and *homepage finding*, aiming to search main pages of several types of navigational web sites. Starting from common information used by Information Retrieval (IR) systems, such as web page content, hyperlinks, and URLs, the model proposes methods to classify queries based on the two categories mentioned above.

Agichtein et al. proposed a predictive model derived from real case studies, which is based on the analysis and the comprehension of user interactions during web navigation [7]. The model tries to elicit and understand user navigation behaviors by analyzing several activities, such as clicks, scrolls, and dwell times, aiming to predict user intention during web page navigation. Moreover, the study proposes to analyze features that are used to characterize the complex interactions following a click executed on a result page.

Lee et al. proposed a feature based model for the automatic identification of search goals, focusing on navigational and informational queries [10]. The model has been developed starting from experimental studies on real user navigation strategies, which have primarily revealed the possibility of effectively associating most queries to one of two categories defined within the taxonomy. They observed that queries not effectively associable to a category are usually related to few topics, such as proper nouns or names of software systems. More specifically, the model proposes two features: *past user-click behavior* to infer users intent from their past interactions with results, and *anchor-link distribution*, which uses possible targets of links sharing the same text with the query.

3. A model for user intent understanding

In this section we describe the model and the features used for the classification process. The model of this work is based on the model proposed in [12].

3.1. A two-level taxonomy for web queries

During a web search the user has a specific goal, generally described by a textual query, and classifiable in a taxonomy. In what follows, we introduce the two-level taxonomy that will be used in the proposed approach for classifying user queries, which is shown in Fig. 1. It synthesizes concepts defined in the taxonomies proposed in [15,16], which have been refined based on the analysis of the query set used in our experiments.

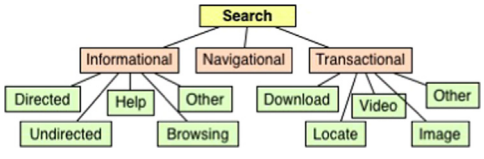

Fig. 1. Two-level taxonomy.