Available online at www.sciencedirect.com
ScienceDirect
Simulation Modelling Practice and Theory 16 (2008) 519–538
SIMULATION MODELLING PRACTICE AND THEORY
www.elsevier.com/locate/simpat

A Monte-Carlo simulation application for automatic new topic identification of search engine transaction logs ☆

Seda Ozmutlu *, Huseyin C. Ozmutlu [1], Buket Buyuk [2]

Uludag University, Industrial Engineering Department, School of Engineering and Architecture, Gorukle, Bursa 16059, Turkey

Received 4 April 2007; received in revised form 13 February 2008; accepted 14 February 2008
Available online 20 February 2008

Abstract

One of the most important dimensions of Web user information seeking behavior and search engine research is content-based behavior, and limited research has focused on content-based behavior of search engine users. The purpose of this study is to perform automatic new topic identification in search engine transaction logs using Monte-Carlo simulation. Sample data logs from FAST and Excite are used in the study. Findings show that Monte-Carlo simulation for new topic identification yields satisfactory results in terms of identifying topic continuations; however, the performance measures regarding topic shifts should be improved.

Keywords: Monte-Carlo simulation; Information science; Search engine user behavior; New topic identification

1. Introduction

Search engines are the most important tools for reaching information over the Web. It is important to successfully capture the information-seeking behavior of search engine users to develop effective information retrieval algorithms. However, it is a real challenge to interpret user information-seeking behavior, since people have different and changing information needs, and they utilize different information searching strategies to solve their information problems [15]. Agichtein et al. [2] state that accurate modeling and interpretation of user behavior has important applications to ranking, click spam detection, Web search personalization and other tasks. In another study, Agichtein et al. [1] demonstrate that incorporating user

☆ This research has been funded by TUBITAK, Turkey and is a National Young Researchers Career Development Project 2005: Fund Number: 105M320: "Application of Web Mining and Industrial Engineering Techniques in the Design of New Generation Intelligent Information Retrieval Systems".
* Corresponding author. Tel.: +90 224 429 2085; fax: +90 224 429 2079.
E-mail addresses: seda@uludag.edu.tr (S. Ozmutlu), hco@uludag.edu.tr (H.C. Ozmutlu), buketbuyuk@uludag.edu.tr (B. Buyuk).
[1] Tel.: +90 224 429 2082; fax: +90 224 429 2079.
[2] Tel.: +90 224 429 2077; fax: +90 224 429 2079.

1569-190X/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.simpat.2008.02.005

behavior data into information retrieval algorithms can significantly improve ordering of top results in real Web search setting.

One of the most important dimensions of Web searching behavior is content-based behavior, which means the information seeking-behavior or pattern of Web users related to content. Currently, Web search engines are not designed to differentiate according to the user's profile and the user's context. Exploiting the user's interest and context in various topics has the potential to improve Web retrieval systems [14,54]. One of the main elements of content-based behavior is new topic identification. New topic identification is discovering when the user has switched from one topic to another during a single search session to group sequential log entries that are related to a common topic [16]). In order to find useful patterns in user sessions, it is necessary to group the queries on the transaction logs into clusters. After the query clusters have been identified, the common usage patterns can be discovered by statistical tools [19]. Also, if the Web search engine is aware that the user's new query is on the same topic as the previous query, the Web search engine could provide the results from the document cluster relevant to the previous query. Alternatively, if the user is on a new topic, the Web search engine could resort to searching other document clusters. Consequently, Web search engines can decrease the time and effort required to process the query and increase the quality of the results. Other implications of session and new topic identification in terms of personalized Web services, Web caching systems and Web site design, are well-documented by Huang et al. [20].

Besides providing better results to the user, custom-tailored graphical user interfaces could be offered to the Web search engine user, if topic changes were estimated correctly by the Web search engine. Ozmutlu et al. [38] mention that users interested in different topics could benefit more from such IR systems designed according to their searching needs. Had topic identification been successfully performed, sophisticated graphical user interfaces could be offered by search engines that can help users (a) to coordinate multiple topics into effective queries, i.e., search histories, various thesauri or keyword generation tools, (b) provide the ability to create multiple sets of working notes related to different or related search topics, i.e., sketching and note creation tools, (c) enable Web users to submit and track multiple queries concurrently on different or related topics, (d) allow for searching multiple Web search engines or collections concurrently on multiple topics, (e) enable the reformulation of multiple queries on different or related topics, and facilitate task switching, i.e. allowing the tracking, storing and manipulating of retrieved results and printouts related to different topics over multiple searches, (f) review search histories from various searches and topics and provide the ability to create clusters of retrieved information related to different or related topics.

There are many studies on new topic identification, session identification, query clustering and text categorization, provided in detail in the related studies section of the paper. Most of these studies analyze the queries semantically and are focused on interpretation of keywords or understanding the topic or the contents of the query. Semantic analysis of queries is an important line of research, but, it is a complicated task; hence its current success is ambiguous [35–37]. Even Support Vector Machines, currently known as the most successful method for text categorization, still cannot perform semantic analysis efficiently [35]. One promising approach is to use non-semantic methodologies to the problem of query clustering or new topic identification in a user search session. In such an approach, queries can be categorized in different topic groups with respect to their statistical characteristics, such as the time intervals between subsequent queries or the reformulation of queries. Considerable performance has been achieved in automatic new topic identification; however, room for improvement still exists, and there is a need to experiment new methodologies.

In addition, the previous approaches used for new topic identification are quite complicated. Monte-Carlo simulation has less computational burden than the other approaches used for new topic identification, and also provides the opportunity to make an educated guess of topic shifts and continuations. Therefore, in this study we chose to apply Monte-Carlo simulation to see whether it improves the results of new topic identification without increasing the computational effort required. The conditional probabilities used in the Monte-Carlo simulation is calculated using samples from real transaction logs and simulation is used to identify topic changes on a separate dataset from the same search engine. The success of Monte-Carlo simulation is determined comparing its results to that of the human expert and to the other new topic identification schemes. We initially present the literature review related to topic identification, followed by the description of the methodology, results and the conclusion.