Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 42 (2006) 751–768
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Analysis of multiple query reformulations on the web: The interactive information retrieval context

Soo Young Rieh [a,*], Hong (Iris) Xie [b,1]

[a] School of Information, University of Michigan, 550 E. University Avenue, Ann Arbor, MI 48109-1092, United States
[b] School of Information Studies, University of Wisconsin, Milwaukee, P.O. Box 413, Milwaukee, WI 53201-0413, United States

Received 15 February 2005; accepted 19 May 2005
Available online 6 July 2005

Abstract

This study examines the facets and patterns of multiple Web query reformulations with a focus on reformulation sequences. Based on IR interaction models, it was presumed that query reformulation is the product of the interaction between the user and the IR system. Query reformulation also reflects the interplay between the surface and deeper levels of user interaction. Query logs were collected from a Web search engine through the selection of search sessions in which users submitted six or more unique queries per session. The final data set was composed of 313 search sessions. Three facets of query reformulation (content, format, and resource) as well as nine sub-facets were derived from the data. In addition, analysis of modification sequences identified eight distinct patterns: specified, generalized, parallel, building-block, dynamic, multitasking, recurrent, and format reformulation. Adapting Saracevic's stratified model, the authors develop a model of Web query reformulation based on the results of the study. The implications for Web search engine design are finally discussed and the functions of an interactive reformulation tool are suggested.

Keywords: Query reformulation; Interactive information retrieval; Web searching; Web search engine

1. Introduction

People have used Web search engines for 10 years now, and in consequence they have become familiar with the simplicity of Web searching. The interface design of Web search engines appears to be fairly standardized: a search box for 20–30 characters and a search button next to it. A common belief is that a Web

* Corresponding author. Tel.: +1 734 647 8040; fax: +1 734 764 2475.
E-mail addresses: rieh@umich.edu (S.Y. Rieh), hiris@uwm.edu (H. Xie).
[1] Tel.: +1 414 229 6835; fax: +1 414 229 6699.

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2005.05.005

search requires only that users type in one or several keywords representing their information need, click on a search button, and wait for mere seconds for thousands of search results. Before entering keywords into the search box, however, there is a critical Web searching step: a query must be formulated. Query formulation requires two kinds of mappings: a semantic mapping of the vocabulary users employ in articulating the task onto the system's vocabulary and an action mapping or set of actions recognizable by a search system (Marchionini, 1995). Semantic mappings are declarative and action mappings are procedural in nature. While action mappings may be relatively easy in Web searching, semantic mappings remain difficult because people need to specify not something they do know but something they do not (Belkin, 1980; Taylor, 1968).

Information retrieval is an interactive and iterative process. Previously, Swanson (1977) pointed out the essential role of the trial-and-error process in information retrieval. According to Swanson, an initial request is a guess about the attributes of desired documents, after which the response of the IR system is employed to revise the initial guess for another try. Efthimiadis (1996) identifies two query formulation stages as the initial query formulation stage in which the search strategy is constructed and the query reformulation stage in which the initial query is adjusted manually or with the assistance of a system. It is often argued that query reformulation is not any easier than initial query formulation given that information retrieval (IR) systems provide very little assistance. Users enter the keywords they know in their initial query. If the initial query does not return the expected search results, users must then submit their second best keywords. This reformulation process can be even more frustrating and complex than the initial formulation because users often experience difficulty in incorporating information from previously retrieved documents into their queries (French, Brown, & Kim, 1997).

Despite the perception that Web searching is simple and easy (Fast & Campbell, 2004), approximately half of all Web users find they must reformulate their initial queries: 52% of the users in the 1997 Excite data set and 45% of the users in the 2001 Excite data set (Spink, Jansen, Wolfram, & Saracevic, 2002) in fact made modifications to their initial query. Several studies have investigated patterns of query reformulation on the Web (e.g., Bruza & Dennis, 1997; Lau & Horvitz, 1999; Spink, Jansen, & Ozmultu, 2001); however, none went on to examine query reformulations beyond the level of descriptive analysis.

This research therefore aims to identify the patterns of multiple query reformulations that focus on sequences of query reformulation per search session. This article enlarges the scope of a preliminary study by the authors (Rieh & Xie, 2001) through the addition of data sets and enhanced data analysis. The article also differs from the previous paper in applying models of interactive information retrieval (IR) (e.g., Belkin, 1993, 1996; Ingwersen, 1992, 1996; Saracevic, 1996, 1997). The framework of interactive IR models allows investigation of query reformulations in terms of interacting with the system, interpreting search results, and shifting strategies beyond comparison of search queries requested and search results received. The study thus has three objectives:

1. characterizing the facets of query reformulation in Web searching;
2. identifying the patterns of multiple query reformulation in terms of sequences; and
3. exploring the ways in which search engines can support query reformulation more effectively in Web searching.

2. Theoretical framework: models of interactive information retrieval

Since the 1990s the literature on IR interaction has been growing, and a number of interactive IR models have been proposed as alternative models for traditional IR models. The traditional model represents IR as the matching or comparing of elements and processes between system and user, but the weakness of this model is that the interaction is not directly depicted. Yet any observation of an IR session reveals the