Chapter XX
Query Log Analysis for Adaptive Dialogue-Driven Search

Udo Kruschwitz
University of Essex, UK

Nick Webb
SUNY Albany, USA

Richard Sutcliffe
University of Limerick, Ireland

ABSTRACT

The theme of this chapter is the improvement of Information Retrieval and Question Answering systems by the analysis of query logs. Two case studies are discussed. The first describes an intranet search engine working on a university campus which can present sophisticated query modifications to the user. It does this via a hierarchical domain model built using multi-word term co-occurrence data. The usage log was analysed using mutual information scores between a query and its refinement, between a query and its replacement, and between two queries occurring in the same session. The results can be used to validate refinements in the domain model, and to suggest replacements such as domain-dependent spelling corrections. The second case study describes a dialogue-based question answering system working over a closed document collection largely derived from the Web. Logs here are based around explicit sessions in which an analyst interacts with the system. Analysis of the logs has shown that certain types of interaction lead to increased precision of the results. Future versions of the system will encourage these forms of interaction. The conclusions of this chapter are firstly that there is a growing literature on query log analysis, much of it reviewed here, secondly that logs provide many forms of useful information for improving a system, and thirdly that mutual information measures taken with automatic term recognition algorithms and hierarchy construction techniques comprise one approach for enhancing system performance.
INTRODUCTION

The Web is growing at an incredible speed and has become an active research area in its own right (Spink & Jansen, 2004). Search engines such as Google (Brin & Page, 1998) enable users to process, access and navigate vast amounts of information. Such engines are built upon the well-established principles of Information Retrieval (IR) (Baeza-Yates & Ribeiro-Neto, 1999). While an IR system takes as input a user query and returns a ranked list of documents considered relevant to it, a Question Answering (QA) system goes one stage further and returns an exact answer extracted from one of the documents. Since its adoption at the Text REtrieval Conference (TREC) (Voorhees, 1999), the Cross Language Evaluation Forum (CLEF) (Magnini, Romagnoli, Vallin, Herrera, Peñas, Peinado, Verdejo & de Rijke, 2003) and the National Test Collection for Information Retrieval (NTCIR) (Sasaki, Chen, Chen & Lin, 2005), in concert with targeted funding under the Advanced Research Development Agency (ARDA) Advanced QUestion Answering for INTelligence (AQUAINT) program, QA has developed rapidly to the stage at which commercial systems such as Qristal are beginning to appear (Laurent, Séguelà & Nègre, 2006).

A considerable amount of the work in IR and QA has been devoted to the retrieval of results for individual queries. Increasingly, however, users need Interactive Information Systems (IIS) capable of converging on a person’s information need by stages, using methods such as Interactive QA (Webb, 2006; Webb & Webber, 2008; Small, Strzalkowski, Liu, Ryan, Salkin, Shimizu, Kantor, Kelly, Rittman & Wacholder, 2004) and dialogue driven search (Kruschwitz, 2003; Kruschwitz, 2005; Kruschwitz & Al-Bakour, 2005). Traditional artificial dialogue systems already allow users to interact with simple, structured data such as train or flight timetables (Zue, Glass, Goodine, Leung, Phillips, Polifroni & Seneff, 1990; Goddeau, Brill, Glass, Pao, Phillips, Polifroni, Seneff & Zue, 1994; Allen, Schubert, Ferguson, Heeman, Hwang, Kato, Light, Martin, Miller, Poesio & Traum, 1995; Aust, Oerder, Seide & Steinbiss, 1995). Such models make extensive use of corpora containing both Human-Computer (H-C) and increasingly Human-Human (H-H) interactions (Hardy, Biermann, Inouye, Mckenzie, Strzalkowski, Ursu, Webb & Wu, 2004). Such corpora can be used to study and capture the phenomena, vocabulary and style of such interactions and hence to develop appropriate machine models.

By contrast, IR and QA systems often operate in much wider domains for which appropriate corpora are not available. As a result, query logs are potentially an extremely valuable resource for increasing our understanding of the complex interactions involved and hence in developing more sophisticated systems. Logs contain a huge amount of information but effective methods for extracting it are only now being developed.

In this chapter we will focus on interactive systems which retrieve information via a dialogue with the user. We will first discuss previous work on log analysis, on interactive information systems and on the use of such analysis to improve interaction. We will then present two case studies which show how an analysis of query logs can be used to improve the underlying domain model and the model of interaction, and hence the quality of interaction in a system. The first study deals with UKSearch (Udo Kruschwitz Search) which performs IR working on a University intranet and is aimed at faculty, staff and students. The second focuses on HITIQA (High Quality Interactive Question Answering) which performs QA over Web documents and is aimed at benefiting the Intelligence community, and those with an analytical approach to information. Finally we will draw the findings of both studies together to identify how log analysis can be used to improve interactive systems in future and to establish what challenges must be overcome in the process.
Related Content

Study on Secure Dynamic Covering Algorithm for E-Logistics Information in a Cloud Computing Platform
[www.igi-global.com/article/study-on-secure-dynamic-covering-algorithm-for-e-logistics-information-in-a-cloud-computing-platform/188381?camid=4v1a](www.igi-global.com/article/study-on-secure-dynamic-covering-algorithm-for-e-logistics-information-in-a-cloud-computing-platform/188381?camid=4v1a)

Using Watermarking Techniques to prove Rightful Ownership of Web Images
[www.igi-global.com/article/using-watermarking-techniques-prove-rightful/55382?camid=4v1a](www.igi-global.com/article/using-watermarking-techniques-prove-rightful/55382?camid=4v1a)

Message-Based Routing in Mobile Networks
[www.igi-global.com/chapter/message-based-routing-mobile-networks/53050?camid=4v1a](www.igi-global.com/chapter/message-based-routing-mobile-networks/53050?camid=4v1a)

Text Summarization Based on Conceptual Data Classification
[www.igi-global.com/article/text-summarization-based-conceptual-data/2616?camid=4v1a](www.igi-global.com/article/text-summarization-based-conceptual-data/2616?camid=4v1a)