Using Query-Log Based Collective Intelligence
to Generate Query Suggestions
for Tagged Content Search

Dirk Guijt$^1$ and Claudia Hauff$^2$({\textsuperscript{2B}})  

1 Sanoma, Hoofddorp, The Netherlands  
dirk.guijt@sanoma.com  
2 Web Information Systems, TU Delft, Delft, The Netherlands  
c.hauff@tudelft.nl

Abstract. One of the standard features of today’s major Web search engines are query suggestions, which aid the user in the formulation of their search queries. Over the years, a number of different approaches have been proposed which have commonly been evaluated in the standard Web search setting. In this work, we build a query suggestion pipeline based on the collective intelligence stored in log data collected from a more constrained search engine which uses tags to index the content. This constrained environment, though large-scale, differs considerably from standard Web search with respect to its users, indexing process and Web coverage. We implement a number of suggestion approaches based on query-flow and term-query graph models and investigate to what extent they are applicable in this more constrained environment.

Keywords: Query suggestions · Query-flow graphs · Search sessions · Collective intelligence · Tags · Tagged content

1 Introduction

One of the standard features of today’s major Web search engines are query suggestions, which aid the user in the formulation of their search queries whilst typing. Those suggestions are commonly generated using the collective intelligence of the search engine users which is stored in the search engine’s query logs. Learning the behaviour of the search engine users and applying the knowledge by generating query suggestions is not an easy task, as large-scale query logs are noisy, may contain errors and logging artefacts.

In this work, we present our efforts on implementing query suggestions for startpagina$^1$, a Dutch Web search portal similar in spirit to the Open Directory Project$^2$. It is currently relying on query suggestions offered by a major

---

The work was conducted during the first author’s TU Delft Master thesis project at Sanoma.

1 http://startpagina.nl  
2 Open Directory Project (now DMOZ): http://www.dmoz.org

© Springer International Publishing Switzerland 2015  
DOI: 10.1007/978-3-319-19890-3_12
Commercial Search Engine (CSE) instead of generating its own. Since interacting with those externally generated suggestions leads to general Web search results (which often include results not indexed by startpagina) many users leave startpagina in the middle of a search session and continue with a general Web search. While search engine switching (initiated by the user) is a well-known phenomenon [25], we consider this switch to be a rather unconscious one, as the users are “switched” to a different engine without their explicit request. In order to increase user retention, we use startpagina’s query log and implement a four-step query suggestion pipeline based on state-of-the-art approaches in (1) search session splitting [16,17], (2) the classification of query reformulations [9,19], (3) the generation of query-flow graphs, and, (4) the extraction of suggestions from such graphs [7–10].

Our main research questions focus on the effects of applying provenly effective methods of generating query suggestions in regular Web search engines to a more constrained search environment, which is similar to search on a social tagging portal. In contrast to regular Web search, our experimental environment indexes its contents through tags created by human annotators. Our search engine is limited in its ability to return results for all queries, as a given query must match a tag before the associated content can be returned. It is not useful to suggest queries for which no matching tag exists and thus no content is available. In this work, we investigate whether query logs can be used to generate relevant query suggestions in this context.

We derive six different models from state-of-the-art query-flow and term-query graph models [7–10] and compare their effectiveness to the suggestions provided by the CSE (our baseline that is currently in use at startpagina). In this particular use case we are able to achieve much higher coverage, that is, the models trained on our log data can provide valid suggestions, i.e. suggestions for which startpagina has results, for a much larger number of queries.

In the remainder of the paper, we first cover related work (§ 2), provide more details about startpagina (§ 3) and describe our approach (§ 4), before presenting our experiments (§ 5) and a discussing of the lessons learnt (§ 6).

2 Background

2.1 Search Sessions

Deriving search sessions, i.e. chronologically ordered sets of queries by a single user with a common search goal (within a certain time interval), from query logs is one of the necessary pre-processing steps to generate query-log based query suggestions. Search session splitting has been tackled by various researchers in the past, including [15–17]. Most works rely heavily on lexicographical and temporal properties; it was shown that those low-level features achieve nearly the same effectiveness for search session splitting as more complex semantic or result-set based features [16].