query history and click-through data are often used to model the users interests, as for example in [27]. Although it is shown that this can improve retrieval quality, such data is often sparse and additional information about the user can let us better understand the user's interests and information needs. Teevan et al [31] make use of a much richer user representation by utilizing a Desktop index which indexes files on the user's hard drive, e-mails, visited web pages and so on. However, this approach treats web documents as common documents and does not take advantage of the characteristics and structure encapsulated within a web page. In this paper, we make use of the richness of a user's complete browsing history, but we also exploit the specific characteristics and structure of web pages. Next to that, we also attempt to apply more advanced NLP techniques to these web documents and investigate whether the noisy nature of web pages influence has a negative effect on this. We find that this approach and taking advantage of web document structure both visibly improve retrieval quality.

Once a user's browsing behavior and interests have been learned, they can be used to re-rank the results returned by a search engine. For instance, a person who is a web developer issues the search query "Ajax". It is likely that both the search results about web development and the user's profile will contain words that are indicative of web development. These results can then be promoted to the top of the ranking based on similarities between the user profile and the relevant search results. In other words, we attempt to increase the chance of getting a relevant result near the top of the ranking.

Previous research [32] suggests that such profile based personalization may lack effectiveness on unambiguous queries like "london weather forecast" and therefore no personalization should be attempted for these queries. However, if this or a related query has been issued by this user before, we could detect any preference for certain weather forecast websites by using the user's URL history, which can also be deducted from the browsing history. We therefore expand upon a method which successfully incorporates a user profile and URL history into a personalized search framework [8]. However, there still exist scenarios in which search personalization might not help or even harm, for example when a query can not be personalized or when a user's information need changes over time. Our method keeps this in mind by also assessing a personalization strategy that is not too aggressive and still allows potentially less relevant results in at a slightly lower rank.

Profile Based Approaches

Some personalization techniques [31, 19, 28] are based on a user profile that expresses the individual's interests and browsing behavior. This can be done explicitly through information provided by the user, which we will not consider due to the extra effort involved on the user's behalf. Various methods and a wide range of information sources have also been proposed to learn a user's profile implicitly without any user effort.

In [7], the user profile is inferred from the entire search history and is used to model long term user interests. Shen et al [23] make use of the recent search history to model the short term user interests, in which session boundaries are used to define short term search history. These methods often suffer from data sparsity and very frequently re-rank results relying on a very limited amount of data.

Other methods have attempted to incorporate more information about the user by using the full browsing history [28, 17]. The Curious browser, a web browser developed to record a user's explicit relevance ratings of web pages and browsing behavior when viewing a page, such as dwell time, mouse clicks and scrolling behavior, is described in [4]. The most promising profile based approach was suggested by Teevan et al. [31]. They use a rich model of user interests, built from both search-related information, previously visited web pages and other information about the user (e.g. documents on their hard drive, e-mails etc.) to re-rank web search results within a relevance feedback framework. In doing this, they obtain a significant improvement over default web ranking. We compare our method to this approach in section 5 and show significant improvement in retrieval performance. The method described in [1] is based on solely using a user's desktop information.

Concerning the model used to describe a user, user interests can be represented as a set of keyword vectors [6], a set of concepts [16], an instance of a predefined ontology [9, 24, 18, 30] or a hierarchical category tree based on ODP and corresponding keywords [15, 2]. In this paper, we will focus on modeling users through a vector of weighted terms.

Click-through Based Approaches

A different range of personalization strategies utilize URL and click-through data from past queries. In [12], user click-through data is collected as training data to learn a retrieval function, which is used to produce a customized ranking of search results that suits a group of users' preferences. In [28, 26, 29], the click-through data collected over a long time period is exploited through query expansion to improve retrieval accuracy. The most promising URL and click-through based method seems to be PClick, or person-level re-ranking based on historical clicks, as suggested in [8]. If a query is issued by a user for the second time, pages that have been clicked during the first search for this query are promoted to the top of the ranking. A disadvantage of this approach is that it can only be applied to repeated queries. The method suggested in this paper will incorporate both a user profile and a user's URL and click-through history. We compare our approach to the PClick method in section 5, and find obtain significant improvements.

Commercial Personalization Systems

Recently, personalized search has also been made available in some of the mainstream web search engines including Google[2] and Yahoo!. These appear to use a combination of explicitly and implicitly collected information about a user. They allow the user to build a profile of themselves by selecting categories of interests and custom tailor the results delivered to the user based on that. However, in practice we expect few users to provide this explicit information. Implicitly, users' search and click-through history is also used personalize results, with results closer to that geographical location of the user additionally favored. However, as the details of these methods and algorithms are not publicly available, we only compare our approach to the default search engine ranking and not the personalized version.

[2]http://googleblog.blogspot.com/2009/12/personalized-search-for-everyone.html