Automatically Mining Facets for Queries from Their Search Results

Abstract:

We address the problem of finding query facets which are multiple groups of words or phrases that explain and summarize the content covered by a query.We assume that the important aspects of a query are usually presented and repeated in the query's top retrieved documents in the style of lists, and query facets can be mined out by aggregating these significant lists. We propose a systematic solution, which we refer to as QDMiner, to automatically mine query facets by extracting and grouping frequent lists from free text, HTML tags, and repeat regions within top search results. Experimental results show that a large number of lists do exist and useful query facets can be mined by QDMiner. We further analyze the problem of list duplication, and find better query facets can be mined by modeling fine-grained similarities between lists and penalizing the duplicated lists.

Introduction:

We address the problem of finding query facets. A query facet is a set of items which describe and summarize one important aspect of a query. Here a facet item is typically a word or a phrase. A query may have multiple facets that summarize the information about the query from different perspectives. Table 1 shows sample facets for some queries. Facets for the query "watches" cover the knowledge about watches in five unique aspects, including brands, gender categories, supporting features, styles, and colors. The query "visit Beijing" has a query facet about popular resorts in Beijing (tiananmen square, forbidden city, summer palace, ...) and a facet on travel related topics (attractions, shopping, dining, ...). Query facets provide interesting and useful knowledge about a query and thus can be used to improve search experiences in many ways. **First**, we can display query facets together with the original search results in an appropriate way. Experimental results show that quality of query facets mined by QDMiner is good. We find that quality of query facets is affected by the quality and the quantity of search results. Using more results can generate better facets at the beginning, whereas the improvement of using more results ranked lower than 50 becomes subtle. We find that the Context Similarity Model outperforms the Unique Website Model, which means that we could further improve quality of query facets by considering context similarity of the lists during ranking the facets and items. The remainder of this paper is organized as follows. We briefly introduce related work in Section . Following this, we propose QDMiner in Section . We discuss evaluation methodology.