Big Data For Query Expansion: A Comparison of Content-based versus Social-based Keywords

Jeff Harwell¹, Christopher Pentoney¹, Gondy Leroy, PhD¹,²
¹Claremont Graduate University, Claremont, CA; ²The University of Arizona, Tucson, AZ

Abstract

As the amount of online information continues to grow, search becomes an increasingly vital user activity. However, it is difficult to motivate users to generate an appropriate, detailed query so query expansion is commonly used to bolster user’s searches. We compare two query expansion methods. The first expands a query using previously submitted queries, the social approach, implemented using Google Search Autocomplete. The second uses phrases as they are found in text, the content approach, implemented using the Google Web Corpus. The query datasets consist of 207,638 tri-gram and 232,981 bi-gram queries from the AOL 500K User Session Collection search corpus that are expanded using both approaches. We use term specificity and term ambiguity, calculated using WordNet, to evaluate the effect on the queries. We evaluate the effect on the query results by calculating domain diversity for a 1% sample of the queries and complete this analysis by calculating edit distance. On average, expansion terms suggested using the social approach were significantly (p < .01) less ambiguous and more specific (p < .01) than terms suggested using the content approach. The search results for the 1% sample of expanded queries showed a small but significantly lower number (p < .01) of distinct domains returned for the social approach. Finally, edit distance showed that query expansion significantly impacts the results returned by the search engine compared to the unexpanded query. This study is the first in a series to understand how modern search engines influence our information retrieval and the breadth, depth and completeness of information available to the community.