Large-scale Contextual Query-to-Ad Matching and Retrieval System for Sponsored Search

[Abstract]

Ricardo Baeza-Yates, Nemanja Djuric, Mihajlo Grbovic,
Vladan Radosavljevic, Fabrizio Silvestri
Yahoo Labs

ABSTRACT

Semantic embeddings of words (or objects in general) into a vector space have proven to be a powerful tool in many applications. In this talk we are going to show one possible application of semantic embeddings to sponsored search. Sponsored search represents the major source of revenue for web search engines and it is based on the following mechanism: each advertiser maintains a list of keywords they deem of interest with regards to their business. According to this targeting model, when a query is issued, all advertisers with a matching keyword are entered into an auction according to the amount they bid for the query, and the winner gets to show their ad, usually paying the next largest bid (this is called second price). The main challenges is that a query may not match many keywords, resulting in lower auction value, lower ad quality, and lost revenue for both, advertisers and publishers. We address them by applying semantic embeddings to this problem by learning how to project queries and ads in a common embedding, thus sharing the same feature space. The major novelty of the techniques we show is that learning is done by jointly modeling their content (words in queries and ad metadata), as well as their context within a search session. This model has several advantages and can be applied to at least three tasks. First, it can be used to generate query rewrites with a specific bias towards rewrites able to match relevant advertising. Second, it can be used also to retrieve for a given a query a set of relevant ads to be sent to the auction phase. Third, given an ad we are able to retrieve all the queries for which that ad can be considered relevant. The major advantage of learning both content and context embeddings is in the fact that a context-based model may suffer from coverage issue: if a query or an ad does not appear in the training set it cannot be treated by the model; content-based embeddings instead can be used to also build models capturing similarities between content, e.g. for a query not appearing in the model built we may capture some of its sub-queries by using content vectors.

Another very interesting characteristic of this method is that all the tasks mentioned above are basically solved by means of a simple K-nearest neighbor search over the set of vectors in the embedding. The method has been trained up to 12 billion sessions, one of the largest corpora reported so far. We report offline and online experimental results, as well as post-deployment metrics. The results show that this approach significantly outperforms existing state-of-the-art, substantially improving a number of key business metrics.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms

Algorithms, Experimentation, Economics.

Keywords

Sponsored search, semantic embeddings, query similarity, query rewriting.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2744112.