Towards better contextual advertising

Evgeniy Gabrilovich
Yahoo! Research

Abstract:

This talk will examine two issues in contextual Web advertising. In the first part of the talk, we propose a methodology for building a robust query classification system that can identify thousands of query classes with reasonable accuracy, while dealing in real-time with the query volume of a commercial web search engine. We use a blind feedback technique: given a query, we determine its topic by classifying the web search results retrieved by the query. Motivated by the needs of search advertising, we primarily focus on rare queries, which are the hardest from the point of view of machine learning, yet in aggregation account for a considerable fraction of search engine traffic. Empirical evaluation confirms that our methodology yields a considerably higher classification accuracy than previously reported. We believe that the proposed methodology will lead to better matching of online ads to rare queries and overall to a better user experience.

In the second part of the talk, we study a particular type of Web advertising, called Content Match, which given the URL of a Web page, aims to embed into the page relevant textual ads. For static pages that are displayed repeatedly, the matching of ads can be based on prior analysis of their entire content; however, ads need to be matched also to new or dynamically-created pages that cannot be processed ahead of time. Analyzing the entire body of such pages on-the-fly entails prohibitive communication and latency costs. To solve the three-horned dilemma of either low-relevance or high-latency or high-load, we propose to use text summarization techniques paired with external knowledge (exogenous to the page) to craft short page summaries in real time. Empirical evaluation proves that matching ads on the basis of Duch summaries does not sacrifice relevance, and is competitive with matching based on the entire page content. Specifically, we found that analyzing a carefully selected 5% fraction of the page text sacrifices only 1%--3% in ad relevance.