Introduction to Computational Advertising

Evgeniy Gabrilovich   **Vanja Josifovski**   **Bo Pang**

Yahoo! Research

701 First Avenue

Sunnyvale, CA 94085, USA

{gabr,vanjaj,bopang}@yahoo-inc.com

1 Introduction

Web advertising is the primary driving force behind many Web activities, including Internet search as well as publishing of online content by third-party providers. Even though the notion of online advertising barely existed a decade ago, the topic is so complex that it attracts attention of a variety of established scientific disciplines, including computational linguistics, computer science, economics, psychology, and sociology, to name but a few. Consequently, a new discipline — Computational Advertising — has emerged, which studies the process of advertising on the Internet from a variety of angles. A successful advertising campaign should be relevant to the immediate user's information need as well as more generally to user's background and personalized interest profile, be economically worthwhile to the advertiser and the intermediaries (e.g., the search engine), as well as be aesthetically pleasant and not detrimental to user experience.

2 Content Overview

In this tutorial, we focus on one important aspect of online advertising that is relevant to the ACL and HLT communities, namely, contextual relevance. There are two main scenarios for online advertising, as advertisers might request to display their ads for a query submitted to a Web search engine, or for a Web page that the user reads online. The former scenario is called sponsored search, since ads are matched to the Web search results, and the latter — content match, as ads are matched to a larger amount of content. It is essential to emphasize that in both cases the context of user actions is defined by a body of text, which could be quite short in the case of sponsored search or fairly long in the case of content match. Consequently, the ad matching problem lends itself to many NLP methods, but also poses numerous challenges and open research problems in text summarization, natural language generation, named entity extraction, computer-human interaction, and others.

At first approximation, the process of obtaining relevant ads can be reduced to conventional information retrieval, where we construct a query that describes the user's context, and then execute this query against a large inverted index of ads. We show how to augment the standard information retrieval approach using query expansion and text classification techniques. First, we demonstrate how to employ a relevance feedback assumption and use Web search results produced by the query. We also go beyond the conventional bag of words indexing, and construct additional features by classifying both the input context and the ad descriptions with respect to a large external taxonomy. A third type of features is constructed from a lexicon of named entities obtained by analyzing the entire Web as a corpus.

We present a unified approach to Web advertising, which uses the same underlying infrastructure to handle both sponsored search and content match scenarios. The last part of the tutorial will be devoted to recent research results as well as open problems, such as automatically classifying cases when no ads should be shown, handling geographic names (and more generally, location awareness), and context modeling for vertical portals.