Intent-Based User Segmentation with Query Enhancement

Wei Xiong, Information Systems Department, New Jersey Institute of Technology, Newark, NJ, USA
Michael Recce, Information Systems Department, New Jersey Institute of Technology, Newark, NJ, USA
Brook Wu, Information Systems Department, New Jersey Institute of Technology, Newark, NJ, USA

ABSTRACT

With the rapid advancement of the internet, accurate prediction of user’s online intent underlying their search queries has received increasing attention from online advertising community. This paper aims to address the major challenges with user queries in the context of behavioral targeting advertising by proposing a query enhancement mechanism that augments user’s queries by leveraging a user query log. The empirical evaluation demonstrates that the authors’ methodology for query enhancement achieves greater improvement than the baseline models in both intent-based user classification and user segmentation. Different from traditional user segmentation methods, which take little semantics of user behaviors into consideration, the authors propose a novel user segmentation strategy by incorporating the query enhancement mechanism with a topic model to mine the relationships between users and their behaviors in order to segment users in a semantic manner. Comparing with a classical clustering algorithm, K-means, the experimental results indicate that the proposed user segmentation strategy helps improve behavioral targeting effectiveness significantly. This paper also proposes an alternative to define user’s search intent for the evaluation purpose, in the case that the dataset is sanitized. This approach automatically labels users in a click graph, which are then used in training an intent-based user classifier.

Keywords: Behavioral Targeting, Classical Clustering Algorithm, Online Advertising, Query Enhancement Mechanism, User Segmentation

INTRODUCTION

Online advertising spending has been increasing at an unprecedented pace over the past decade. In order to increase the effectiveness of targeting advertising, models are built based on user’s web activities, such as search queries, to personalize advertisements. There are hundreds of companies and many different approaches (e.g., context, social, cookie-based, etc.) developed to improve targeting advertising. The largest internet companies, such as Google, Facebook, and Yahoo, are all advertising companies. Data from search activities, web surfing and social

DOI: 10.4018/ijirr.2013100101
connections are all mined to optimize online advertising effectiveness.

With the rapid advancement of the World Wide Web (WWW), accurately predicting user’s online intent underlying their search queries has played an important role in satisfying user’s online experience. It helps advertisement campaign to target more relevant users, content publishers to recommend web content, search engines to return personalized results, and many other service providers to facilitate user’s online experience. For instance, a user with a travel plan in his mind would have a higher probability of clicking a flight advertisement. Thus from a perspective of a flight advertiser, identifying users who are likely to travel could help target ads delivery and increase effectiveness. Similarly, if a content publisher knows a user’s online intent, it can recommend relevant content to match user’s interests.

Assume a user who issued queries such as “best carry-on luggage” and “foreign transaction fees”. From the observation of these user queries, it can be inferred that this user is probably planning an overseas trip and may have the intent to purchase a flight. Thus, it is an opportunity both for advertisers to deliver flight advertisements, and also for other online service providers to offer travel related services.

This study is focused on capturing relevant users based on their online intents. We explore this problem in three major aspects, which can be summarized as follows:

- **Representing a user’s online intent**: A user’s online intent is modeled based on the user’s online behavior, such as the search queries issued by the user or the web pages viewed by the user;
- **User classification**: For advertisers who are interested in users who have a specific intent, a user can be classified as either having or not having this intent. Therefore, a good intent representation strategy should be able to effectively differentiate users based on their online intents;
- **User clustering**: It would be also interesting to investigate how much intent-based user clustering could help behavioral targeting by grouping similar users into segments according to their online intent.

As a rich source of information on web searchers’ behavior, query logs have been utilized by advertising companies to deliver personalized advertisements and leveraged by researchers to tackle other application problems, such as query suggestion. To carry out research on behavioral targeting, it is desirable to have golden standard datasets, which contain both query logs and ad click information. This type of datasets is used by advertising companies to train and test a model that predicts user’s ad click behavior. However, they are not available in academic community, which makes conducting research in this area difficult.

The publicly available query logs are small, dated, and sanitized, since search engine companies are reluctant to release complete query log data. In the past decade, web search has grown at an unprecedented pace. Typical queries issued by users contain very few terms. In an empirical study (Jansen, Spink, & Saracevic, 2000), about 62% of all queries contain one or two terms, and fewer than 4% of the queries have more than six terms. On average, a query only contains 2.21 terms, which can carry only a small amount of information about the user. The tendency of users to use short and ambiguous queries makes it difficult to fully describe and distinguish a user’s intent. For instance, the user intent behind query “Steve Jobs” will be represented as two terms in the bag of words model: “Steve” and “Jobs”, along with their weights in the feature space, which could describe an intent of a user who is either interested in the person “Steve Jobs” or looking for a job.

Another important aspect of user’s search query is that, the volume of queries is huge and follows the Zipf’s law, where a small amount of queries appear very often while most queries appear only a few times. This makes the query feature space sparse and hence could undermine a classifier’s performance in predicting future unseen data. For example, “laptops” and “cameras” are frequent queries and there
www.igi-global.com/chapter/bates-berrypicking-model-1989-2002-2005/127126?camid=4v1a