Towards Building a New Age Commercial Contextual Advertising System

James Miller, University of Alberta, Edmonton, Canada
Abhimanyu Panwar, University of Alberta, Edmonton, Canada
Iosif Viorel Onut, IBM Canada, Ottawa, Canada

ABSTRACT

Advertising via the Internet is a significant industry; however, in many ways, the industry is still in its infancy and still requires significant refinement to achieve its full potential. In contextual advertising (CA), the ad-network places ads related to the content of the publishers’ webpages. In this article, the authors introduce an approach to implement a CA system for an ad-network. Their contributions are threefold: First, they propose schemes to prepare feature vectors of a webpage for the purpose of classification by its subject. To do so, the authors extract information from its peer webpages as well. Secondly, they prepare a suitable taxonomy from ODP. This taxonomy fulfils the requirements of a CA system such as broad coverage of semantically relevant topics etc. Thirdly, they conduct experiments on the proposed CA system architecture. The results establish the competence of the proposed approach. The authors empirically establish that the scheme which extracts information from the intersection of cues from web accessibility and search engine optimisation, of the target webpage provides the best accuracy among all the CA systems.

KEYWORDS

Classifiers, Contextual Advertising, Feature Vector, Inlinks, Outlinks, Webpage Classification

1. INTRODUCTION

The Web has become a popular venue to advertise. The online revenues have been steadily increasing at a staggering rate of 20% each year. The fact that total online revenues worldwide has risen to the tune of US $117 billion (statista.com, 2013) exemplify the significance of online advertising. A portion of this revenue comes from contextual advertising (Chakrabarti et al., 2008). Contextual advertising (CA) is essentially a mode of targeted advertising where the advertisement (ad) shown to the user is relevant to the webpage’s content. For instance, if a user is browsing a webpage about pizza, then the ads shown on the webpage may be of local pizza vendors. Webpage owners provide space on the webpage at primary locations to display such content-related ads. Users click on such ads and get directed to the ad webpage. Doing so not only brings revenues to the webpage owner but increases user experience as well, thereby giving rise to a win-win situation for both of the parties involved (Broder et al., 2007). In the current form of contextual advertising, when a webpage is being rendered in the user’s browser, the webpage requests ad(s) from an ad-providing entity known as ad-network. This request may contain a webpage URL and other associated information. The ad-network selects
ads from its ad-repository by analyzing the request and responds by delivering content related ads. Therefore, it is of the utmost importance for the ad-network to select the most optimal content-related set of ads for the requesting webpage.

To match the ads with the webpage, an approach based on matching keywords between a webpage and that of the ad has been suggested (Ribeiro-Neto et al., 2005; Yih et al., 2006) in early work on CA. But this approach may lead to a bad selection of ads. A webpage about the damages of oil spills showing ad of an oil company is such an example. Such types of situations tarnish the reputations of both the webpage owner as well as the advertiser. To overcome these limitations, a semantic approach has been suggested (Broder et al., 2007, Anagnostopolulos et al., 2007, Armano et al., 2011; Lee et al., 2013). In this approach, a webpage and ads are classified into the nodes of a taxonomy. “Top matching” ads are retrieved to display on the webpage. The taxonomy used in CA is a hierarchical classification of a wide range of topics. If ads cannot be matched to the specific topic of the webpage, then ads are delivered on the generalized subject of the webpage. For example, if the subject of the webpage is swimming, then ads related to sportswear also serve the purpose of CA. Therefore, the semantic approach based ads are optimally related content-wise with the webpage.

For a semantic based CA system, a robust classification system resides at its core. The ad-network must classify the webpage based on its content’s subject into the nodes of the taxonomy of topics. The more accurately and specifically a webpage is classified, the better the ads will be retrieved from the ad-repository.

In today’s information savvy times, there are webpages on the web possibly on every topic and subject imaginable. Therefore, the taxonomy of a CA system must cover a wide range of topics and subjects. Preparing such a kind of taxonomy and populating it with example webpages requires substantial human effort. ODP (dmoz.org 2015) is a publicly available taxonomy which covers a large number of topics and has been used in CA research (Lee et al., 2013, Vargiu et al., 2013). It has been actively maintained for the last 15+ years by human editors, hence it is a good resource to be exploited in the implementation of a CA system. However due to the skewed distribution of webpages and other properties of ODP, it is not a good candidate to be used in its original form. In this study, we provide a methodology to prepare a well-suited taxonomy for a CA system from ODP.

Accurate classification of a webpage based on its subject into the nodes of the taxonomy is essential for a CA system to function optimally. The classification of a webpage usually takes place in three steps.

The first step is to collect information about its subject matter from the resources available. A webpage is a semi-structured document and its contents are written within HTML tags. In this study, we exploit this property to extract relevant information and discard the noise. It has been suggested that peer webpages such as inlinks and outlinks also carry relevant information about the webpage (Qi and Davison 2009). We empirically determine the effects of the inclusion of information from the peer webpages on the accuracy of the classification.

The second step in the classification of a webpage is the method of transforming the webpage into a feature vector. Converting a text corpus prepared from the information sources of a webpage into a feature vector can lead to large dimensions of feature vectors. We implement feature selection techniques to select the most relevant features for the purpose of classification.

The third step is the classification algorithm or the classifier. It takes the feature vector of a webpage as input and predicts the class of the webpage. We implement state of the art classification algorithms on the taxonomy prepared from the ODP and empirically determine the best performing combination of the choices for the three steps of the classification for a CA system.

We propose a novel architecture to implement a CA system for the ad-network. It involves two stages. The first stage is offline. This stage is completed prior to launching the services of the ad-network. In this stage, a classifier is trained on a well-formed subset of ODP. This classifier is stored by the ad-network. The second stage is the online stage, where the ad-network responds to the requests of the webpage for contextually related ads. The ad-network loads the classifier prepared in stage one.
Related Content

Assessing the Usefulness of Testing for Validating and Correcting Security Risk Models Based on Two Industrial Case Studies
Integrated Requirement and Solution Modeling: An Approach Based on Enterprise Models
www.igi-global.com/chapter/integrated-requirement-solution-modeling/23785?camid=4v1a

Extending the ORM Conceptual Schema Language and Design Procedure with Modeling Constructs for Capturing the Domain Ontology
www.igi-global.com/chapter/extending-orm-conceptual-schema-language/23783?camid=4v1a

Quality-Driven Model Transformations: From Requirements to UML Class Diagrams
www.igi-global.com/chapter/quality-driven-model-transformations/26834?camid=4v1a