Qualitative data analysis on contextual advertisements

Bayyavarapu Kranthi Kumar
bkranthikumar.15.cse@anits.edu.in
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

Dadi Sharmila
sharmila.dadi29@gmail.com
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

Buraga Srinivasa Rao
bsrinivasarao.15.cse@anits.edu.in
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

Gedda Poojitha
gpoojitha.15.cse@anits.edu.in
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

ABSTRACT

The internet advertising market is growing day by day much faster than any other advertising platform. The technology and methods used for advertising online go more and more towards automated processes that analyze the content of the page and then matches the ads with these parameters. Advertisers in web pages like Facebook and LinkedIn have some preferred set of users they wish to reach by showing their ads. Some sites offer fine-grained sets of user characteristics– including their career, wealth, education information, etc- that advertisers can specify for targeting their audience, and each of these characteristics requires different amounts of money for targeting. Generating contextual advertisements for websites is a problem of growing interest. The recent $240 million investment by Microsoft in Facebook clearly reaffirms the opportunity in contextual advertising for online social networks. Content-contextual advertisement programs such as Google AdSense and Yahoo Contextual Match work by automatically spotting keywords in web pages and displaying ads based on the keywords. The displayed ads are also referred to as Contextual Advertisements. Contextual advertising allows advertisers to display ads relevant to the content of the website rather than using the data about the visitor. Contextual ads can complement the content. Contextual ads are very effective for particular kinds of content. Some studies indicate they can increase click-through rate by 63%. These ads are generally represented by a URL along with a textual description that is also used to match the ad with the target page. User activity on venues such as forums, marketplaces, and groups on social networking sites are excellent targets for monetization.

Keywords— Content-based advertising, Contextual advertising, RAKE, Scrappy, Advertising

1. INTRODUCTION

The advertisements are generated based on spotting the keywords and the websites visited by the user. This is known as content-based advertisements. As the advertisements are generated based on the history of visited websites by the user, we analyze the particular user from his login session and till the log out session. This comes under the domain of Data Analytics since we need to examine the data of a particular user. The process of obtaining raw data and converting it into information useful for decision-making by users is Data Analysis.

Qualitative Data Analysis (QDA) is the processing of the qualitative data that have been collected into some form of explanation. QDA is usually based on an interpretative philosophy. The idea is to analyze the meaningful and symbolical content of qualitative data.

Contextual advertising allows advertisers to display ads relevant to the content of the website rather than using the data about the visitor. The idea is not completely new; before the advent of the Internet, contextual advertising was widely used in magazine and newspaper ads. With heavy reliance on personal data came a bigger awareness of how advertisers were collecting and using this data and reluctance of users to online ads, which resulted in the rise of ad blockers. As a result, we are finding a revival of contextual ad targeting today—methods immune to the GDPR changes, and almost just as effective. Contextual advertising works by scanning every URL of the website using a web crawler and when a user visits a given page, the information about the URL is passed in the request to the ad server. The ad server can match with data collected by the crawler for the specific URL and can match relevant keywords.

2. RELATED WORK

Our paper is closely related to the previous works on contextual advertising. In an influential paper Targeted Advertising for Online Social Networks by Pinaki Mitra and Kamal Baid, it can be observed that targeted advertising can help advertisers mitigate product market competition and increase advertiser profits in a competitive industry.

The research on online advertising in general and contextual advertising, in particular, is emerging with the growth of the Internet. The most basic approach is syntactic matching that estimates the ad relevance by analyzing the similarity of word features between ads and pages. A notable work by Ribeiro-Neto *et al*. studies a number of matching strategies using extracted keywords as well as the traditional text matching with the bag-of-words representation and cosine similarity. Those authors also point out the problem of low intersection between vocabularies used for pages and ads (vocabulary impedance). To solve the problem, they expand the page vocabulary by adding words from similar pages with appropriate weights.
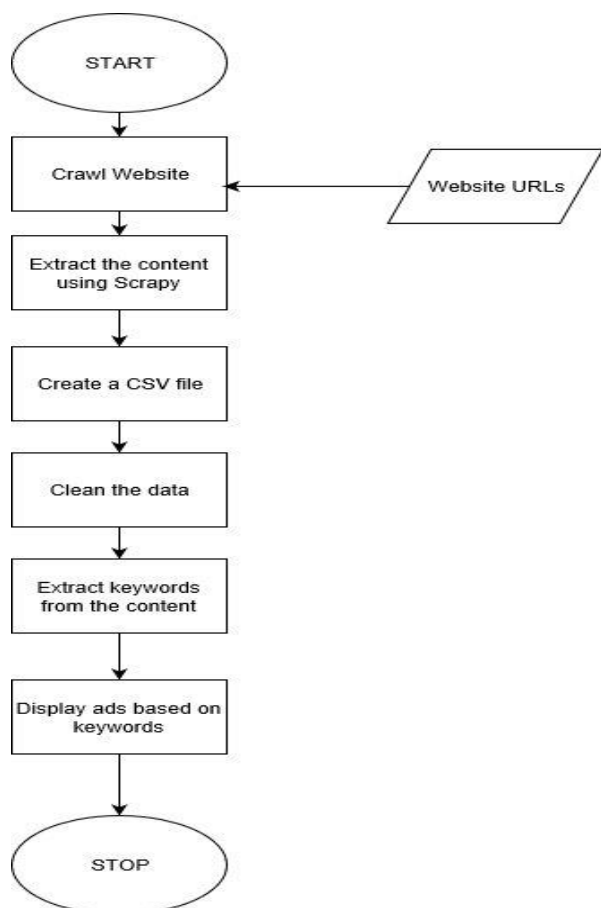
3. PROPOSED WORK

Fig. 1: System Architecture

There are two modules in our system.
3.1. Web crawling
The first module is for extracting the content from the website. The content includes title, description and the URL of a particular webpage. We will implement the crawling process for a particular section of the website. The scrape library is used for web crawling. The input given to the crawler is a particular webpage from where we want to start extracting the data. The output produced by the crawler is a CSV file with title, description and the URL of the webpage as columns. With these contents, the CSV file is created and is used for the next module.

3.2. Keyword Extraction
The second module of our system is to extract keywords from a particular CSV file. We implement RAKE algorithm to extract keywords.

Working of RAKE algorithm includes the following steps: Split the document into an array of words, breaking it at word delimiters (like spaces and punctuation). For each candidate keyword, add the word scores of its corresponding words to find the candidate keyword score.

Split the words into sequences and break each sequence at a stopword. Each sequence is now a "candidate keyword". Calculate the "score" of each individual word in the list of candidate keywords. This is calculated using the metric: degree (word)/frequency(word). Take the first one-third highest scoring keywords from the list of candidate keywords as the final list of extracted keywords.

4. RESULTS
The goal of our experiments is to show up the essential features of contextual advertising. Our project displays three advertisements for a webpage. The first one is a static advertisement generated based on the content of the webpage. The second is an advertisement generated using the history of the user of that website. The third advertisement is displayed on a particular webpage if the content given by the client relates to the content on that page. The crawler which is implemented using scrape library gives the output in a CSV file with the title, description and the URL of the particular webpage as columns. The RAKE Algorithm gives the most relevant keywords for the webpage identified as the output. The high-scored keyword for the complete website is taken as the default advertisement i.e., if the user has not reached through any webpage, the default advertisement is shown. As soon as the user visits the particular webpage, that relevant ad is shown at the left side of the webpage.

5. FUTURE WORK AND CONCLUSION
One more problem in content-contextual advertising is vocabulary impedance. The problem occurs because an ad generator can't contain all the keywords in its databases. When a keyword is purchased by a company, that keyword and corresponding ads are put into a database. A company can purchase a limited number of keywords. So in this case, even if we provide informative keywords to ad-generator, we get unrelated ads. Solving this problem is also a part of future work.

Rake algorithm can be improved more as the keywords are generated using only the description of the article whereas more effective keywords can be generated using the relationship between the title of the article and the description of the article. The crawler built in this project works only on a given website and ads are displayed using users visit history only on this website. It can be improved such that ads are displayed on a website using the user's complete history.

6. REFERENCES
[1] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," Journal of Computer-Mediated Communication, vol. 13, no. 1, pp. 210–230, November 2007.

[2] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, and Z. Chen, "How much can behavioural targeting help online advertising?" in WWW, 2009, pp. 261–270.

[3] D. Freitag, "Information extraction from HTML: Application of a general machine learning approach," in In Proceedings of the Fifteenth National Conference on Artificial Intelligence. AAAI Press, 1998, pp. 517–523.

[4] P. Turney and C. Canada, "Extraction of key phrases from the text: Evaluation of four algorithms," National Research Council, Institute for Information Technology, Tech. Rep., 1997.

[5] H. Li, D. Zhang, J. Hu, H.-J. Zeng, and Z. Chen, "Finding keyword from online broadcasting content for contextual advertising," in ADKDD '07: Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising. New York, NY, USA: ACM, 2007, pp. 55–62.

[6] "Yahoo term extraction service. http://developer.yahoo.com."

[7] M. Craven, D. Dipasquo, D. Freitag, A. Mccallum, T. Mitchell, and K. Nigam, "Learning to extract symbolic knowledge from the World Wide Web." AAAI Press, 1998, pp. 509–516.

[8] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," in Proceedings of the 27th annual meeting on Association for Computational Linguistics. Morristown, NJ, USA: Association for Computational Linguistics, 1989, pp. 76–83.