Product Ad Recommendation System for Web Pages

Sanjana Maiya
University of California, Santa Cruz
smaiya@ucsc.edu

ABSTRACT

Contextual Advertising refers to the placement of advertisement within a web page based on keywords extracted from the web page. It is essential that the ad displayed on the page is relevant to the page content, both for user experience, click and revenue generated. This relevance is decided based on the match between the contents of the page and individual advertisements.
In this project, a system to recommend Amazon product ads for news articles is developed. The system has two sub-systems, a keyword extraction sub-system to extract keywords from the page and a search sub-system to "search and retrieve" ads from the corpus of all product ads by using the extracted keywords as the query. The project is evaluated based on the relevance of the ads to the news articles.

1. INTRODUCTION

Today, online advertising supports a large part of the web ecosystem. In 2014, the total internet advertiser spend was around 120 billion U.S. dollars, putting the Web in second place after only television in terms of ad spend.

Content targeted ad systems analyze the content of a web page, such as a news article, and find prominent keywords in the page. These keywords are then used by the system to find ads which are good matches for the keywords selected. The selected ads are displayed to the user, and the relevance of the ad determines whether the user clicks on the ad or not.

Choosing appropriate keywords is important to make sure that the user sees ads of interest and hence does not get discouraged from visiting the content pages in the future. From the advertiser's perspective, displaying relevant ads is equally important since it directly impacts revenue. Therefore, displaying highly relevant ads is a big win for content creators, web users and advertisers

Google AdSense, Yahoo! Bing Network Contextual Ads, Microsoft AdCenter are some of the most popular content targeted adnetworks.

In this project, we will implement an end-to-end content targeted ad-system, which extracts keywords from news articles and suggests relevant product ads from Amazon to the user. We will evaluate this system at two levels : how well the keywords have been extracted, and how relevant the ads are to the news articles.

2. RELATED WORK

The work [1] by Ribeiro-Neto et. al on contextual matching examines several techniques for matching web pages with ads based on keywords extracted. Ads and web pages are represented as vectors in a vector space model and the matching is done using several strategies based on cosine similarity. Further, since there might be a discrepancy between the vocabulary used in the web page and ad, the vector used to represent the web page is expanded with terms from other similar pages, thus improving overall precision.

[2] is a follow up of the above method and significantly improves precision by learning to advertise using Genetic Programming.

[3] "A Semantic Approach to Contextual Advertising" expands on the search of ads in the vector space by adding classes as extra dimensions, so that the ads topically match the web page contents. The semantic match of the pages and the ads is performed by classifying both into a common taxonomy.

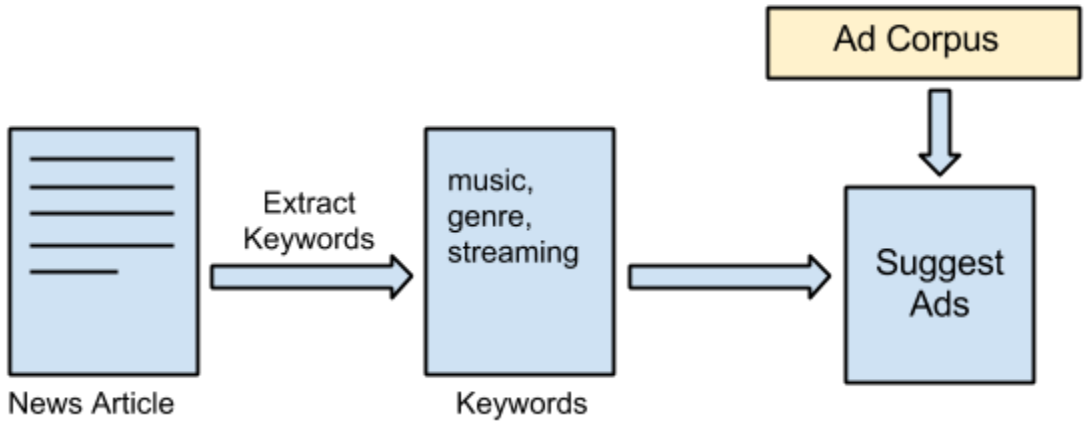
[4] is a follow up paper of the above, describes how matching of ads to web pages can be improved significantly by augmenting the ad-page scoring function with extra parameters from a logistic regression model on the words in the pages and ads. A key property of the proposed model is that it can be mapped to standard cosine similarity matching and is suitable for efficient and scalable implementation over inverted indexes. The model parameter values are learnt from logs containing ad impressions and clicks, with shrinkage estimators being used to combat sparsity.

In [5] a method for phrase extraction is described that uses a variety of features to determine the importance of phrases for displaying ads. The system is trained with pages that have been hand annotated with important phrases. The learning algorithm takes into account features based on tf-idf, html meta data and query logs to detect the most important phrases. During evaluation, each page phrase up to length 5 is considered as potential result and evaluated against a trained classifier.

3. METHODOLOGY

The project is implemented in two parts : The keyword extraction from the web page, which in this case is a news article, followed by the matching of keywords with the ad corpus to come up with top relevant ads.

Figure 1: System diagram



3.1 Keyword Extraction

In order to extract keywords from news article, an unsupervised approach is taken. Here, the keyword extraction algorithm is given the news article in text form, without any other information, and the keywords are extracted automatically. One of the reason for taking an unsupervised approach is that the news article dataset was completely unlabelled, and also spanning several categories like sports, entertainment, politics, business and technology.

The keyword extraction algorithm is built on top of RAKE [6] in Java. The features of the algorithm are as listed:

1. The text article is split into sentences, where punctuation act as sentence boundaries.
2. The algorithm removes certain words using a list of stop words and discards them as unimportant. Also, stopwords are used as phrase boundaries. This helps generate candidates that consist of one or more non-stopwords.
3. For each candidate phrase, the score is calculated by the summing the score of each word in the phrase. The words are scored using the term frequency and the length of the candidate phrase in which they occur.
4. Part-of-speech (P.O.S) tagging is done for all the sentences in the news article, and words which are not nouns or adjectives are discarded.
5. Keywords which are of length one are also discarded.

The result of the algorithm is a list of phrases, and a score associated with each phrase. We will choose the top scoring phrases, and append them as long as we do not have at least 10 keywords which best represent the news article

With the extracted keywords, we move on to the second step. The keywords will be used as a query to search from a corpus of indexed ads.

3.2 Ad Retrieval and Scoring

Around 19.4 million Amazon product ads are indexed and saved in an inverted index. For each set of keywords, the index is queried to retrieve the upto 3 top scoring advertisements. Apache Lucene[9] is used as a framework for the indexing and retrieval. While matching ads, the following are considered:

1. Removal of stop words, a default list provided by Lucene
2. A list of sensitive stopwords is maintained, with words like kill, death, rape etc. If the keywords query has any of these words, then no ad is returned for the news article.
3. The scoring makes use of term frequency and inverse document frequency.
4. The number of words in the query which occur in the document is used.
5. The query is normalized based on its length
6. If the resulting score is lesser than a threshold decided through cross-validation, then we assume that the ads retrieved are not relevant, and no ads are displayed for the news article.

We had initially considered the price of the article which scoring the ad, however, the idea was discarded since it degraded the relevance of the ads.

4. DATASETS

The project uses 2 datasets - news articles for which ads are recommended, and a set of products that will be used as ads.

1. 2200 news articles are taken from BBC [7], and are from various categories such as sports, technology, entertainment, politics and business. All these articles are in text format and were initially unlabeled. 200 articles were selected from at random from each category (a total of 1000 articles) and posted on Amazon Mechanical Turk to be labelled with keywords.

2. 19.4 million Amazon products [8] were used as the ad corpus. Each ad is an Amazon product (title + category) from one of 25 categories like Books, Electronics, Home and Kitchen, Beauty and Musical Instruments.

5. EVALUATION

5.1 Evaluation of keyword extraction

To evaluate the keyword extraction from news articles, 100 articles were selected and the extracted keywords were compared against the annotated keywords for the articles.

The measures used were Precision, Recall and f1 measure where :

Precision = (totalRelevantKeywordsRetrieved / totalKeywordsRetrieved)

Recall = (totalRelevantKeywordsRetrieved / totalRelevantKeywords)

f1Measure = 2 * (Precision * Recall) / (Precision + Recall)

The results (see Table 1) showed a surprisingly low precision and recall. The results from Turk keyword labelling and our keyword extraction did not agree very much.

Table 1 : Keyword extraction evaluated against amazon turk labels

Precision	13.22
Recall	18.37
f1 measure	15.38

However, running our end-to-end system on these news articles and eyeballing the recommended ads indicated that our keywords were pretty good, and the low agreement between turk labelled keywords and the ones recommended by our algorithm hinted at a problem with the turk labelled data.

To verify how well the articles had been labelled by Turks, we took a manual look at the results from Amazon Turk, and found them to be much below our expectations. Many of the articles were labelled with completely irrelevant keywords. Due to time and fund constraints, the news articles could not be re-labelled. We also decided not to pursue the approach of using this labelled data for training a supervised learning model for keyword extraction, as the data was not of good quality. However, to make sure that the keyword extraction from our system was better than the annotated keywords, the annotated keywords were used as a baseline in the end-to-end evaluation.

5.2 Evaluation of ads recommended

The second evaluation was to test how well the ad system had performed end-to-end. The same hundred articles used in the keyword extraction test were used for suggesting upto 3 ads for each article. Three ads are suggested since a web page may have multiple ad slots. Three sets of results were obtained. The first set of results was retrieving ads using title of articles as keywords. This served as our first baseline. Titles of articles usually have important keywords since they serve as summaries of articles. The second set of results were ads which were retrieved using the Turk labelled keywords (section 5.1). This was the second baseline, and we expected to better this as discussed in section 5.1. The third set of results was from our keyword extraction algorithm, where keywords extracted were used to retrieve ads. The evaluation was done using the following metrics:

1. **Ad Coverage** : The total number of articles for which ads were recommended. This metric indicates the ability of the system to recommend ads. Since the system recommends up to 3 ads, the ad coverages for first, second and third ad are considered. For example, if the Ad coverage for the second ad is 65%, it means that we retrieved the second ad for only 65% of news articles.
2. **Precision** : Amongst the ads that were displayed, the percentage of ads that were relevant. This evaluation is done for the first, second and third ad recommended. Overall precision is calculated

over all three ads, where overall precision = (number of relevant ads displayed / total number of ads displayed)

Amazon Mechanical Turk was used in order to evaluate the relevance of ads for news articles. The users were asked to classify up to 3 ads as relevant/not relevant for each news article.

Baseline results:

The ads extracted using the title of the article as keywords were used as the first baseline for evaluation. The metrics for titles as keywords are as follows:

Table 2: Ad Coverage and Precision for Baseline 1 (using title as keywords)

Ad Coverage for Ad 1	100%
Ad Coverage for Ad 2	98%
Ad Coverage for Ad 3	98%
Precision for Ad 1	30%
Precision for Ad 2	19.38%
Precision for Ad 3	18.36%
Overall Precision	22.63%

The ads extracted using the labelled keywords were used as the second baseline for evaluation. The metrics for the labelled keywords are as follows:

Table 3: Ad Coverage and Precision for Baseline 2 (labelled keywords)

Ad Coverage for Ad 1	79%
Ad Coverage for Ad 2	65%
Ad Coverage for Ad 3	56%
Precision for Ad 1	63.29%
Precision for Ad 2	49.23%
Precision for Ad 3	55.35%
Overall Precision	56.5%

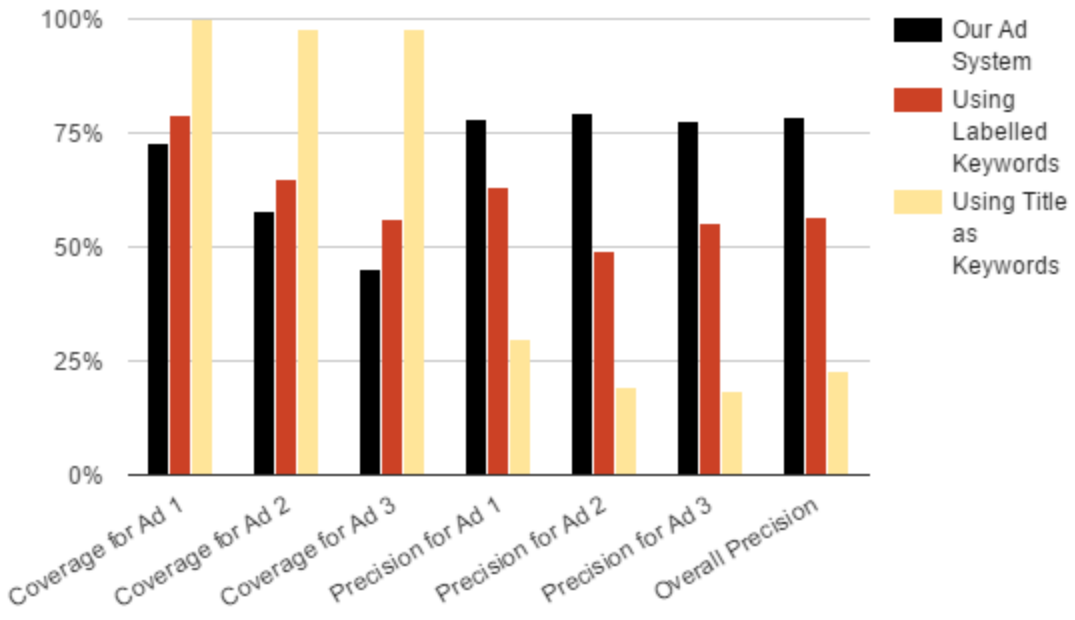
Results using end-to-end Ad recommender system

Table 4: Ad Coverage and Precision for our system

Ad Coverage for Ad 1	73%
Ad Coverage for Ad 2	58%
Ad Coverage for Ad 3	45%
Precision for Ad 1	78.08%
Precision for Ad 2	79.31%
Precision for Ad 3	77.77%
Overall Precision	78.40%

The results (see Fig 2) from our ad system (Table 4) are significantly better than the baseline (Table 2,3) which uses the labelled keywords for retrieving ads. The coverage of ads has dropped, which is not a bad thing, since it suggests that ads which are irrelevant are discarded. We are better off not displaying ads than displaying ads which are completely irrelevant to the news article and turn off users. The precision numbers are much better than the baselines, indicating that of the ads which are displayed, the relevance is high.

Figure 2: Evaluation - Coverage and Precision



6. CONCLUSION AND FUTURE WORK

The relevance of ads displayed in web pages is a key factor for the success of contextual ads. In this report, we have developed a system to recommend Amazon product ads for news articles. We extract keywords from news articles and retrieve ads based on the extracted keywords. The results of the system are evaluated end to end, and the system achieves a precision of around 78% while recommending ads for the articles.

Future work involves accurate labelling of keywords for a subset of news articles, so that supervised learning can be done to improve the keyword extraction process. Also, user specific data from browser logs and Amazon product purchases can be used in order to suggest more relevant ads.

7. REFERENCES

[1] B. Ribeiro-Neto, M. Cristo, P. B. Golgher, and E. S. de Moura. Impedance coupling in content-targeted advertising. In SIGIR ’05: Proc. of the 28th annual intl. ACM SIGIR conf., pages 496–503, New York, NY, 2005. ACM.

[2] A. Lacerda, M. Cristo, M. A. G., W. Fan, N. Ziviani, and B. Ribeiro-Neto. Learning to advertise. In SIGIR ’06: Proc. of the 29th annual intl. ACM SIGIR conf., pages 549–556, New York, NY, 2006. ACM.

[3] A. Z. Broder, M. Fontoura, V. Josifovski, and L. Riedel. A semantic approach to contextual advertising. In SIGIR, pages 559–566, 2007.

[4] Deepayan, C., Deepak, A., & Vanya, J. (2008) Contextual Advertising by Combining Relevance with Click Feedback. In the proceedings of the 17th International Conference on World Wide Web, pp. 417 - 426. ACM Press

[5] Wen tau Yih, Joshua Goodman, and Vitor R. Carvalho. Finding advertising keywords on web pages. In WWW ’06: Proc. of the 15th international conference on World Wide Web, pages 213–222, New York, NY, USA, 2006. ACM Press.

[6] RAKE algorithm : Rapid Automatic Keyword Extraction (RAKE) algorithm as described in: Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010)

[7] BBC dataset: D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006.

[8] Amazon dataset: Image-based recommendations on styles and substitutes J. McAuley, C. Targett, J. Shi, A. van den Hengel SIGIR, 2015

[9] Apache Lucene: https://lucene.apache.org/