

# Multiword Keyword Recommendation System for Online Advertising

Stamatina Thomaidou  
Department of Informatics  
Athens University of Economics and Business  
Athens, Greece  
E-mail: thomaidous@aueb.gr

Michalis Vazirgiannis  
Department of Informatics  
Athens University of Economics and Business  
Athens, Greece  
E-mail: mvazirg@aueb.gr

**Abstract**— As search engines, social networks, and the World Wide Web become more popular and widely used, online advertising turns into a very profitable industry. Individuals and companies promote their products or services in search engines through textual ads, alongside the organic search results triggered by a specific query. For this purpose, advertisers must create advertising campaigns. The development of these campaigns is a laborious task involving significant human resources and expertise. In this paper we propose a system for multiword keyword recommendations in the context of developing a web advertising campaign in a semiautomatic manner. Given a landing page, the system extracts relevant terms consisted of two or three words to match a potential search query. Furthermore, it proposes the most relevant keywords and other suggested terms that do not exist in the landing page text using search result snippets. In addition, we present blind testing experiments on real world data indicating that our approach outperforms prominent existing industrial solutions in most of the cases.

**Keywords**—online advertising; sponsored search; textual advertising; keyword selection

## I. INTRODUCTION

### A. Online Advertising

The advertising industry is rapidly changing as companies, individuals, and advertisers increasingly understand the usefulness and value of the World Wide Web as an integral part of our lives. Online advertising is gaining acceptance and market share while it has evolved into a \$10 billion industry [14, 15]. It is more targeted than traditional advertising means, as the audience is already searching for what the advertiser wants to sell with targeted queries. This is translated to better return of investment (ROI), since the advertising budget spent is directly reaching potential customers and not uninterested audience. Moreover this form of advertising is significantly cheaper than traditional advertising because of its targeted nature. Other important benefits are the immediate publishing of information that the internet provides, good conversion tracking, and finally, the fact that even an “offline” potential buyer researches through search engines the products that they want to purchase.

Sponsored search advertising is the primary source of income for search engines [15, 16]. Few carefully selected paid

advertisements are displayed alongside organic search results. Pay per click (PPC) is an internet advertising model used on websites, where advertisers pay their host only when the ad is clicked. In the case of search engines, advertisers typically bid on keyword phrases relevant to their target market.

### B. Rationale

There is a fine but important distinction among placing ads reflecting the query intent and placing unrelated ads: users may find the former beneficial as an additional Web navigation facility while the latter are likely to annoy the searchers and disturb the user experience [1]. Keyword selection is a cornerstone process in web advertising campaign development, considering that search engine providers, advertisers, and consumers want the best combination of presenting, promoting, and discovering the proper advertisements.

In this paper we propose a system for automated keyword extraction and suggestion in the context of web advertising campaigns. The advantages of our method are: a. to optimize human resource effort and b. to improve quantity, quality and variety of proposed keywords. Next, we present related work, the system design and articulation, experimental results and comparison to competitive systems. The paper concludes with a summary and suggestions for plans and ideas for future work.

## II. RELATED WORK

### A. Query log mining for keyword generation

The areas of sponsored search, textual advertising and keyword research tend to be more focused on automatic extraction, suggestion, and expansion of keywords. An advertising campaign might involve a wealth of landing pages. The manual selection of even a small set of keywords is quite laborious, a fact which lead to the recent launch of commercial tools that produce keyword sets directly from a landing page. There exist different techniques for keyword generation. Search engines use query log based mining tools to generate keyword suggestions. In this way, they focus on discovering co-occurrence relationships between terms and suggest similar keywords. They start from an initial key phrase and they are based on past queries that contain these search terms. Google AdWords Keyword Tool [2] exploits this ability and presents frequent queries for the seed set of words.

Keywords generated by taking into consideration traffic reports are limited to words that occur frequently in advertisers search logs which are likely to be expensive (in terms of CPC) because of their competitive nature among a large amount of advertisers. Furthermore, in several cases they are not so relevant, because this technique favors more general terms and not specific keywords that the advertiser would ideally choose to match his text ad. Consequently, this approach results increased bids for the suggested keywords to compete efficiently for a high ranking among other text ads. However, increasing the bid to achieve a high position in the sponsored results does not guarantee a profit increase as general terms result to high click-through rates but at the same time to low conversion rates [4].

Other commercial tools determine an advertiser's top competitors and then actively search for the keywords they are targeting. After a period of time, lists of targeted keywords that are competitive for pay per click advertising are automatically generated. This also may result to a recommendation set of keywords which are likely to be expensive.

### B. Specific keywords for search engine advertisements

As mentioned above, advertisers want to match their text ads with more relevant and specific user search queries in the purpose of attracting more targeted audience to not only click their ads but also make a conversion. The objectives of seed set expansion for key phrases are to find out more relevant, nonobvious and well-focused keywords. A. Joshi and R. Motwatni [3] defined a new measure called nonobviousness. While evaluating their technique they defined nonobvious term as a term not containing the seed keyword or its variants sharing a common stem.

The challenge of generating keywords for advertising purposes has attracted significant scientific attention. Studies concerning keyword selection can be classified into keyword extraction oriented techniques, into "synonymous" words suggestion methods, and into important recent research studies focused on achieving a proper combination of both efforts.

TermsNet and Wordy authors [3, 4] in their methodologies exploit the power of search engines to generate a huge portfolio of terms and to establish the relevance between them. After selecting the most salient terms of the advertiser's web page they query search engines with each initial seed term. With their methods they find other semantic similar words. Wordy system [4] proposed single word terms for each seed keyword.

The above methods can be considered as corpus (or domain) independent, as the systems directly extract keywords from the documents - in the mentioned cases from web pages and pages from search results - without any previous or background information [5]. Corpus dependent approach requires a large stack of documents and predetermines phrases to build a prediction model. S. Ravi et al. [6] propose a generative model within a machine translation framework so the system translates any given landing page into relevant bid phrases. They first construct a parallel corpus from a given set of bid phrases  $b$ , aligned to landing page keywords  $l$ , and then learn the translation model to estimate  $\Pr(l|b)$  for unseen  $(b,l)$  pairs. This approach performs very efficiently but depends on

the chosen domain and data that the human decision factor may affect.

## III. SYSTEM DESCRIPTION

### A. Keyword Generation from a Landing Page

This system was developed in the context of an overall automated solution for creating and optimizing a Google AdWords campaign. Our proposed method begins with generating the appropriate keywords for a given landing page. We call this procedure *keyword generation* step and it contains two subtasks: *keyword extraction* from the landing page and *keyword suggestion* for each extracted keyword phrase. This component aims at proposing valid and representative keywords. As a final step, the resulted output from these procedures will be given as input data to the next process of the proposed system responsible for campaign creation, optimization, and management [17].

### B. Keyword Extraction

In this process, we follow the corpus independent approach in order to rely solely on the given landing page document. In web page structure, text field holds the main meaning. According to vector space model, each web page can be seen as a document and text must be segmented as many weighted keywords which all together hold the semantics of a document [7]. After segmentation of text, the result will be a bundle of keywords and each keyword is called a term in a document. Then each of these terms must be weighted properly to assure that terms with higher semantic meaning and relevance to our page have larger weight. The weight of given term is calculated in the following equation called *tf-idf* scheme after all the documents are processed:

$$w_{i,j} = tf * idf = \frac{freq_{i,j}}{Max_i freq_{i,j}} * \log\left(\frac{N}{n_i}\right) \quad (1)$$

Because of the form of the examined document, which in our case is a single landing web page (a single HTML document) is that we have no longer the  $N$  documents and  $n_i$  parameter. Thus, we apply a single document keyword extraction method.

As a preprocessing step, the HTML content of each landing page is parsed, stop words are removed and the content is tokenized. For this process, our system uses the Jericho HTML Parser [8] which is a Java library allowing analysis and manipulation of parts of an HTML document, including server-side tags, while reproducing verbatim any unrecognized or invalid HTML. It also provides high-level HTML form manipulation functions.

Next, for each word  $l_j$  in the tokenized output, we compute a weight associated with the word for each occurrence inside a specific tag, e.g. the occurrence of a word inside bold tags <b>:

$$w_{jtag} = weight_{tag} \times f_{jtag} \quad (2)$$

where  $weight_{tag}$  is a special weight assigned to each different kind of HTML tags and  $f_{jtag}$  is the frequency of the word inside the specified tag.

The weight of each tag is assigned according to its importance inside the HTML document. We can set higher values on *important tags* such as <title>, meta keywords, meta description, anchor text, <h1>, <b>.

Then, we compute the special weight of each word as the sum of all  $w_{jtag}$  weights for this word.

$$special\_weight_j = \sum w_{jtag} \quad (3)$$

Finally, the relevance score of each word is computed:

$$relevance\_score_j = \frac{special\_weight_j}{MAX\_WEIGHT} \quad (4)$$

where MAX\_WEIGHT represents the maximum special weight that a word could have inside the HTML document. MAX\_WEIGHT can be different for each HTML document because some of them may not have links or bold tags, etc. So, the MAX\_WEIGHT is simply the sum of all the tag weights of the specific page. In Table I we propose the assignment of tag weights following an approach that ranks the importance of these tags according to where web page designers choose to place the most important information on their website.

TABLE I. TAG WEIGHTS

<i>Element</i>	<i>Assigned Weight</i>
<title>	50
Meta keywords	40
Meta description	40
Anchor text	30
<h1>	30
<b>	10
other	1

Unimportant words occurring on the page can be filtered out using a threshold on the relevance score. While single words frequently have a broad meaning, multiword phrases are more specific and thus can be more representative as advertising keywords [9]. A typical query, especially while searching for a product or service, varies from 1 up to 3 words. For that reason, from the extracted single word terms we pull together possible combinations of two-word phrases inside the given landing page. Next, in order to construct word co-occurrence matrix, the top N words with high relevance scores are ranked in descent order. Then we define the meaning of *co-occurrence* as follows: if  $word_i$  and  $word_j$  appear in a same unit which is predefined, then they co-occur once, and  $freq_{i,j}$  should be added one [7]. It is obvious that the matrix is symmetrical, so  $freq_{i,j}$  is equal to  $freq_{j,i}$ . The predefined unit for our case is each different area inside an HTML document, defined by HTML tags. As we explained above, we must multiply with the proper tag weight.

Finally, we consider the most salient co-occurring two-word terms above a certain threshold and follow the same process, searching for new co-occurrence with each unique single word term. In this way, we extract three-word terms. By gathering all terms, we construct the extracted keywords vector. In order to *boost* three-word terms first, two-word terms second and single word terms third, we modify their relevance score with the following factor:

$$boosted\_score_j = relevance\_score_j * k^{noOfWords} \quad (5)$$

where k is a free parameter (in our experiments we set it as  $k=100$ ) and *noOfWords* is the number of words composing a term.

### C. Keyword Suggestion

From the previous step of keyword extraction we have already extracted the initial keywords. These will be the seed keywords for the additionally suggestions. Initially, as this procedure begins, the set of suggestions is empty. We provide as input the extracted keywords from the landing page. For each given seed keyword, the keyword is entered as a query into a search engine API. We use for this purpose Google JSON/Atom Custom Search API [10] to achieve into making these queries programmatically. With this API, developers can use RESTful requests to get search results in either JSON or Atom format. The API will return a set of short text snippets, snippets that are relevant to the query and thus to the keyword.

From the response data we retrieve *feed/entry/summary/text()* which is a string type property indicating the *snippet* of the search result and *feed/entry/title/text()* which is a string type property indicating the *title* of the search result. The top 30 results are downloaded and loaded in Apache Lucene Library [11], which we use it for implementing indexing and query support in our system. Each extracted term from the previous step which was a seed for the query has now a set of results which we use as a document in the Lucene index. Each set of title and snippet results that were retrieved after a seed query represents this document for Lucene indexing.

In this step, we parse the resulted document and construct a new vector of words. Based on the Lucene scoring method we can find single word and two-word terms that have the most occurrences inside the document and thus are kept as the most relevant for the specific seed query. Each of these terms is representing a new query  $q$ .

The score of query  $q$  for document  $d$  correlates to the cosine-distance or dot-product between document and query vectors in a Vector Space Model (VSM) of Information Retrieval. Again, we sort in descent order the new queries based on this score and we create a vector of suggested keywords and their scores for each of the seed terms.

Before we place our output as an integrated input vector for the next component, we normalize scores and use again a specified threshold for keeping only the most salient terms.

#### IV. EXPERIMENTS AND SYSTEM EVALUATION

Although the actual performance evaluation of this system could be achieved by running a developed web advertising campaign for a period of several weeks, we evaluated at a first glance our method using human ranking for resulted keywords following a blind testing protocol.

##### A. Description of Experimental Results

The landing pages for our experiments were taken from different thematic areas, promoting several products and services. The categories were:

1. hardware product
2. corporate web presence optimization service
3. gifts
4. GPS review
5. hair products
6. vacation packages
7. web design templates
8. car rental services

To compare our system results, we used other competitive keyword suggestion tools: Google Keyword Suggestion Tool [2], Alchemy API [12] and from Google AdWords API the RelatedToUrl method [13]. AAC (Automatic Advertising Campaign) stand as the acronym for our system. We present in Table II an example of each system keyword recommendations for websites of category (2).

##### B. Human Evaluation Methodology

We constructed a dataset of each method generated keywords in order to start a blind experiment evaluation. Eleven researchers and informatics postgraduate students provided judgments for each system output regarding keyword relevance, specificity and non-obviousness using a scale of 1-5. Test measures were defined as follows:

- a) *Relevance*: The relevance of keywords related to each landing page
- b) *Specificity*: How general or specific were the generated keywords
- c) *Nonobviousness*: How usual and repeated or nonobvious were the generated keywords related to the category and advertising form of each landing page

In Figures 1, 2 and 3 we present the evaluation results.

TABLE II. RESULTS FOR A SAMPLE LANDING PAGE (CORPORATE WEB PRESENCE OPTIMIZATION)

Landing Page: atticom.gr/en/- Top-20 Results			
Google Keyword Suggestion	Alchemy	AAC	RelatedToURL
search engine placement search engine optimization seo seo optimization company	competent technical staff search engine optimization world class researchers	text mining services personnel services web services web pages corporate reputation mining	seo company seo company seo company what is seo what is seo what is seo

Landing Page: atticom.gr/en/- Top-20 Results			
Google Keyword Suggestion	Alchemy	AAC	RelatedToURL
seo company seo companies website optimization company search engine placement companies seo optimization services best seo company what is seo seo search engine search engine placement company search engine optimization firms local search marketing local search engine optimization google search optimization website optimization services top seo companies local search engine marketing small business seo	online advertising services Our team competitive international context efficient solutions internet marketing art research innovative advertising engine corporation image international	reputation mining data development web advertising marketing search engine search engine optimization company vision personnel vision personnel services web pages design news contact change introducing mining marketing search mining services services web reputation mining corporate reputation development web web advertising	seo optimization company seo optimization company seo optimization company search engine placement

Relevance Evaluation

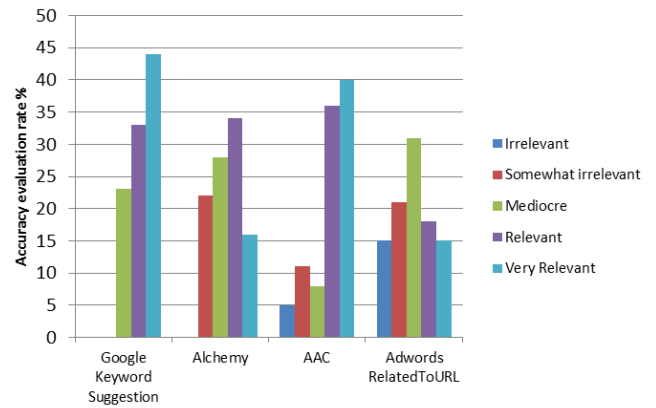


Figure 1. Answers of human evaluators reviewing the output of each system (relevance of keywords)

Specificity Evaluation

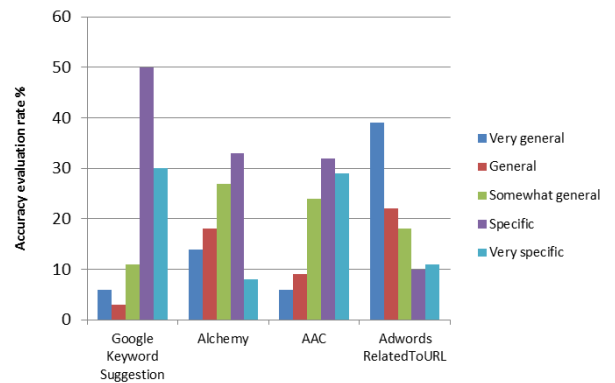


Figure 2. Answers of human evaluators reviewing the output of each system (specificity of keywords)

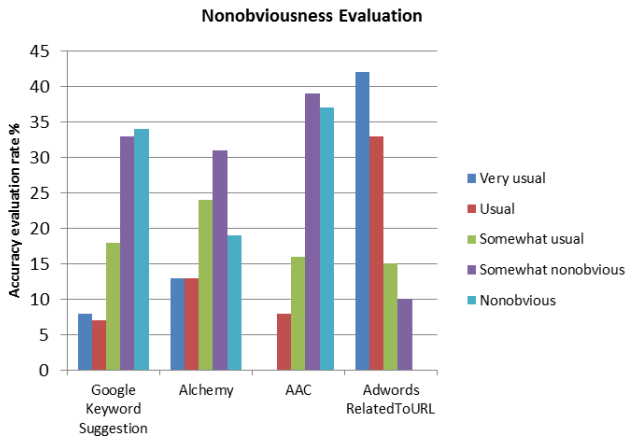


Figure 3. Answers of human evaluators reviewing the output of each system (nonobviousness of keywords)

## V. CONCLUSION AND FUTURE WORK

In this paper we proposed a system that, given a landing page in the context of online advertising for products and services promotion, automatically extracts and suggests keywords for web advertising campaigns. In this way, our contributions regarding the improvement of the advertising campaign development process consist in:

- Automating the task of finding the appropriate keywords
- Recommending multiword terms with high specificity without the need to capitalize on usage data such as query and web traffic logs
- A *fully developed system* with convincing experimentation on real world data from various thematic areas
- Experimental results indicating that our system outperforms in most cases prominent competitive industrial ones.

Using the search result snippets for the process of keyword suggestion has helped a lot to retrieve faster the proper information rather than crawling actual documents. It was also a helpful mean to keep the trends and thus retrieving trending topics at a specific time.

Future work will be conducted to enhance our system for more specific extraction using CSS analysis and structured content scraping of the landing page. Also, searching result snippets from queries on twitter search and tags can be helpful due to the compact nature of twitter messages. They can help in filtering out irrelevant or general information, while mining market trends. We also plan to evaluate keywords using information retrieval measures, such as precision and recall, adapted to the different criteria we use (relevance, specificity, non-obviousness).

Finally, a further extension on our system can be the *ad creative generation* component. The creation of specialized ad

text will be based on previous work and research studies on paraphrasing methods, sentence extraction and compression, sentence and surface realizers, and text summarization. In combination with category specific templates which will be filled with the product characteristics, such as name, price, location, etc., the system will generate ad text for the advertisements of the campaign. The above characteristics will be extracted from customer's web page primarily.

## ACKNOWLEDGMENT

The research of S. Thomaidou is co-financed by the European Union (ESF) and Greek national funds via Program "Education and Lifelong Learning" of the NSRF - Program: *Heracleitus II*. Prof. M. Vazirgiannis is partially supported by the *DIGITEO Chair grant LEVETONE* in France.

## REFERENCES

- [1] A.Z. Broder, P. Ciccolo, M. Fontoura, E. Gabrilovich, V. Josifovski, and L. Riedel, "Search advertising using web relevance feedback," *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08*, 2008, p. 1013.
- [2] [adwords.google.com/select/KeywordToolExternal](http://adwords.google.com/select/KeywordToolExternal) Retrieved Feb. 25, 2011
- [3] A. Joshi and R. Motwani, "Keyword Generation for Search Engine Advertising," *Sixth IEEE International Conference on Data Mining - Workshops (ICDMW '06)*, Dec. 2006, pp. 490-496.
- [4] V. Abhishek and K. Hosanagar, "Keyword generation for search engine advertising using semantic similarity between terms," *Proceedings of the ninth international conference on Electronic commerce*, ACM, 2007, p. 94.
- [5] J. Liu, C. Wang, Z. Liu, and W. Yao, "Advertising Keywords Extraction from Web Pages," *Web Information Systems and Mining*, 2010, pp. 336-343.
- [6] S. Ravi, A. Broder, E. Gabrilovich, V. Josifovski, S. Pandey, and B. Pang, "Automatic generation of bid phrases for online advertising," *Proceedings of the third ACM international conference on Web search and data mining*, ACM, 2010, p. 341-350.
- [7] N. Zhou, J. Wu, and S. Zhang, "A Keyword Extraction Based Model for Web Advertisement," *Integration and Innovation Orient to E-Society Volume 2*, vol. 252, 2007, p. 168-175.
- [8] [jericho.htmlparser.net/](http://jericho.htmlparser.net/) Retrieved Feb.25, 2011
- [9] S. Kiritchenko and M. Jiline, "Keyword optimization in sponsored search via feature selection," *Proceedings of the ECML PKDD 2008, Workshop on New challenges for feature selection in data mining and knowledge discovery*, Citeseer, 2010, pp. 122-134.
- [10] [code.google.com/apis/customsearch/v1/overview.html](http://code.google.com/apis/customsearch/v1/overview.html) Retrieved Feb.25, 2011
- [11] [lucene.apache.org/java/](http://lucene.apache.org/java/) Retrieved Feb.25, 2011
- [12] [alchemyapi.com](http://alchemyapi.com) Retrieved Feb. 25, 2011
- [13] [code.google.com/apis/adwords/](http://code.google.com/apis/adwords/) Retrieved Feb. 25, 2011
- [14] B. Edelman, M. Ostrovsky, and M. Schwarz, "Internet advertising and the generalized second price auction: Selling billions of dollars worth of keywords," *Ariel*, vol. 02138, 2005, pp. 1-25.
- [15] Google Investor Relations [investor.google.com/earnings.html](http://investor.google.com/earnings.html) Retrieved Feb. 25, 2011
- [16] S. Yang and A. Ghose, "Analyzing the relationship between organic and sponsored search advertising: Positive, negative, or zero interdependence?," *Marketing Science*, vol. 29, 2010, p. 602-623.
- [17] K. Liakopoulos, "Automatic Advertising Campaign Development: Campaign Creation and Budget Optimization", M.Sc. Thesis, Athens University of Economics and Business, 2011.