

# Multiword Keyword Recommendation System for Online Advertising

Stamatina Thomaidou<sup>1</sup> Michalis Vazirgiannis<sup>2</sup>

<sup>1</sup>Department of Informatics, Athens University of Economics and Business  
co-financed by ESF and NSRF Program Heracleitus II

<sup>2</sup>Department of Informatics, Athens University of Economics and Business  
Laboratoire d'informatique (LIX) École Polytechnique France  
partially supported by the DIGITEO Chair grant LEVETONE in France

International Conference on Advances in Social Network Analysis and Mining  
Kaohsiung, Taiwan 2011

# Outline

- 1 Online Advertising
- 2 Related Work
- 3 System Description
- 4 Experiments and System Evaluation
- 5 Conclusions and Future Work

# Introduction

*Online advertising* is a form of promotion that uses the Internet and World Wide Web for the expressed purpose of delivering marketing messages to attract customers.

Benefits:

- More targeted than traditional means - Better ROI
- Immediate publishing of information
- Good conversion tracking
- Purchase offline but in most cases research online first (ROPO)

# Introduction

- Textual ads - Two main channels for distributing such ads:
  - ① Sponsored search (or paid search advertising) places ads on the result pages of a Web search engine, where ads are selected to be relevant to the search query
  - ② Content match (or contextual advertising) places ads on third-party Web pages
- All major Web search engines (Google, Microsoft, Yahoo!) support sponsored ads and act simultaneously as a Web search engine and an ad engine
- Pricing Models: Pay-per-click (PPC), Pay per action (PPA), Pay-per-impression (PPI)

# Rationale

- Keyword selection - cornerstone process in web advertising campaign development
- Propose a system for automated keyword extraction and suggestion in the context of web advertising campaigns
- Optimize human resource effort and improve quantity, quality and variety of proposed keywords

This system was developed in the context of an overall *automated solution for creating and optimizing a Google AdWords campaign*

# Commercial tools and research literature approach

## *Commercial tools*

- Manual selection of keywords is quite laborious → commercial tools produce keyword sets directly from a landing page
- Search engines use query log based mining tools to generate keyword suggestions
- Taking into consideration traffic reports → terms that occur frequently in advertisers search logs which are likely to be **expensive** or **general**

## *Research systems*

- One approach: 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.
- Another common approach: Start with Keyword Extraction then expand terms...

# Keyword Generation from a Landing Page

- We chose Google Adwords as the testing campaign platform → Modules for Google Adwords API
- Keyword Generation Procedure
  - Keyword Extraction
  - Keyword Suggestion
- Output: Sets of multiword keywords (n-grams) for each landing page
- We give it as input data to the next process of the proposed system responsible for *automated campaign creation, optimization, and management*

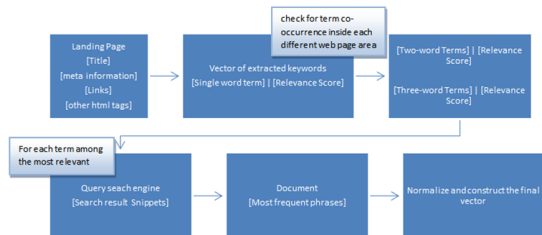


Figure: System Procedures

# Keyword Extraction

- Extract useful information from HTML document (landing page) → Java using JDK SE 6 on the Eclipse IDE
- Pre-processing step
  - HTML content of each landing page is parsed, stop words are removed and the content is tokenized and lowercased
  - Jericho HTML Parser: java library allowing analysis and manipulation of parts of an HTML document
  - KEA data file for english stopwords / Lucene GreekAnalyzer.java for greek stopwords
- Assign weights to important tags
  - Importance of tags according to where web page designers choose to place most important information

Table: Tag Weights

Element	Assigned Weight
<title>	50
meta keywords	40
meta description	40
anchor text	30
<h1>	30
<b>	10
other	1



## Term Scoring (1)

*Like a tf-idf scheme but...*

For each term  $l_j$  in the tokenized output, we compute a weight associated with the term for each occurrence inside a specific tag, e.g. the occurrence of a term inside bold tags  $\langle b \rangle$  :

$$w_{jtag} = weight_{tag} * f_{jtag} \quad (1)$$

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

## Term Scoring (2)

Then, we compute the special weight of each term as the sum of all  $w_{jtag}$  weights:

$$special\_weight_j = \sum w_{jtag} \quad (2)$$

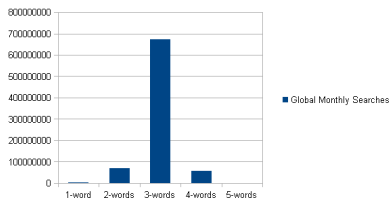
Relevance score of each term is computed:

$$relevance\_score_j = \frac{special\_weight_j}{MAX\_WEIGHT} \quad (3)$$

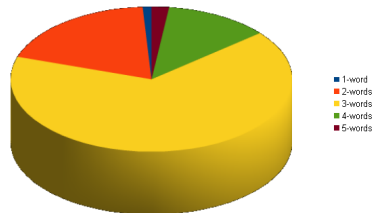
## Multiwords

*Which is the "optimal" phrase length?*

Figure: Keyword searches for a car rental company



(a) GMS



(b) Frequency of discrete keywords

## Co-occurrence

### *Construct term co-occurrence matrix*

- Top  $N$  words with high relevance scores are ranked in descending order
- Define the meaning of co-occurrence as follows: if  $word_i$  and  $word_j$  appear in a same unit which is predefined (i.e. HTML tag-defined area), then they co-occur once, and  $freq_{i,j}$  should be added one. It is obvious that the matrix is symmetrical, so  $freq_{i,j}$  is equal to  $freq_{j,i}$
- Multiply with the proper tag weight
- Extract three-word terms → 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
- By gathering all terms, we construct the extracted keywords vector

## Boosted Score for multiwords

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:

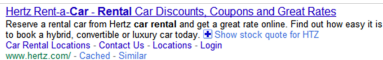
$$boosted\_score_j = relevance\_score_j * k^{noOfWords} \quad (4)$$

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.

# Keyword Suggestion (1)

For each given seed keyword (extracted from previous step)

- Keyword is entered as a query into a search engine API (Google JSON/Atom Custom Search API) example: "car rental"
- API returns a set of short text snippets relevant to the query in Atom format
- The top 30 results are downloaded and loaded in Apache Lucene Library



[Hertz Rent-a-Car - Rental Car Discounts, Coupons and Great Rates](#)  
Reserve a rental car from Hertz car rental and get a great rate online. Find out how easy it is to book a hybrid, convertible or luxury car today. [Show stock quote for HTZ](#)  
[Car Rental Locations](#) - [Contact Us](#) - [Locations](#) - [Login](#)  
[www.hertz.com/](http://www.hertz.com/) - [Cached](#) - [Similar](#)

Figure: Snippet

## Keyword Suggestion (2)

- 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)
- Sort in descending order the new queries based on this score and create a vector of suggested keywords and their scores for each of the seed terms
- Find new unique distinct terms and use once again co-occurrence
- Place our output as an integrated input vector  $\rightarrow$  normalize scores and use again a specified threshold for keeping only the most salient terms

# 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

We tested English and Greek Language. To compare our system results, we used other competitive keyword suggestion tools:

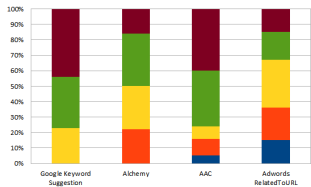
- 1 Google Keyword Suggestion Tool
- 2 Alchemy API
- 3 AAC (Automatic Advertising Campaign) stand as the acronym for our system
- 4 Google AdWords API RelatedToUrl method



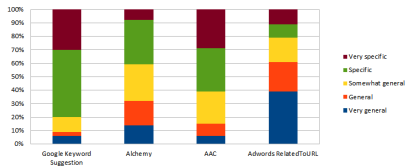
# Experimental Evaluation Methodology

- Human ranking for resulted keywords following a blind testing protocol
- Eleven researchers and informatics postgraduate students provided judgments using a scale of 1-5
- Test measures
  - 1 *Relevance*: The relevance of keywords related to each landing page
  - 2 *Specificity*: How general or specific were the generated keywords
  - 3 *Nonobviousness*: How usual and repeated or nonobvious were the generated keywords related to the category and advertising form of each landing page

# Accuracy evaluation rate



(a) Relevance



(b) Specificity

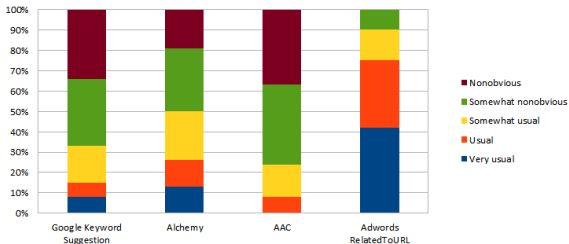


Figure: Nonobviousness

# Contributions

## Improvement of the advertising campaign development process

- 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 Keyword Suggestion...

- Faster retrieval than crawling actual documents
- Trends

## Future Work

- Structured content scraping of the landing page
  - Product attributes, descriptions, pricing, etc.
  - CSS analysis
- Keyword Evaluation
  - IR measures, such as precision and recall, adapted to the different criteria we use
- Twitter snippets from trending topics and tags

*We also working on...*

- *Automatic Ad creative Generation* Component
- Actual performance evaluation of this system could be achieved by running a developed web advertising campaign for a period of several weeks



A.Z. Broder, P. Ciccolo, M. Fontoura, E. Gabrilovich, V. Josifovski, and L. Riedel, 2008.  
Search advertising using web relevance feedback  
*In Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08*



Abhishek, V., Hosanagar, K., 2007.  
Keyword generation for search engine advertising using semantic similarity between terms  
*In Proceedings of the ninth international conference on Electronic commerce.*



Joshi, A., Motwani, R., 2006.  
Keyword Generation for Search Engine Advertising  
*Sixth IEEE International Conference on Data Mining - Workshops (ICDMW06).*



Ravi, S. et al., 2010.  
Automatic generation of bid phrases for online advertising  
*In Proceedings of the third ACM international conference on Web search and data mining.*



Bartz, K. et al., 2008.  
Natural language generation for sponsored-search advertisements  
*In Proceedings of the 9th ACM conference on Electronic commerce - EC '08. New York, New York, USA: ACM Press, p. 1.*



Liakopoulos, K., 2011.  
Automatic Advertising Campaign Development: Campaign Creation and Budget Optimization  
*M.Sc. Thesis, Athens University of Economics and Business*



S. Ravi, A. Broder, E. Gabrilovich, V. Josifovski, S. Pandey, and B.Pang, 2010.

Automatic generation of bid phrases for online advertising  
*Proceedings of the third ACM international conference on Web search and data mining, ACM*



J. Liu, C. Wang, Z. Liu, and W. Yao, 2010.

Advertising Keywords Extraction from Web Pages  
*Web Information Systems and Mining*



N. Zhou, J. Wu, and S. Zhang, 2007.

A Keyword Extraction Based Model for Web Advertisement  
*Integration and Innovation Orient to E-Society Volume 2*



S. Kiritchenko and M. Jiline, 2008.

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*



B. Edelman, M. Ostrovsky, and M. Schwarz, 2005.

Internet advertising and the generalized second price auction: Selling billions of dollars worth of keywords  
*Ariel, vol. 02138, 2005*



S. Yang and A. Ghose, 2010.

Analyzing the relationship between organic and sponsored search advertising: Positive, negative, or zero interdependence?  
*Marketing Science, vol. 29, 2010*

# Thank you!

Data and Web Mining Group  
Athens University of Economics and Business  
<http://www.db-net.aueb.gr>

# Terminology

- **Keyword:** A word or phrase that matches a web-user's search query and at the same time describes the content advertised. Advertisers bid on keywords for ad auctions.
- **Ad-Creative:** The text that a web-user reads on an advertisement
- **Impression:** The appearance of an advertisement in a SERP after a web-users query
- **Click:** The action of a web-user clicking on an advertisement
- **Conversion:** Action (e.g. purchase, registration) after arriving to a website
- **Campaign:** Set of components and preferences for the advertising purpose
- **Ad Group:** Set of related ads, keywords, and placements within a campaign
- **Quality Score:** Measure of how relevant your ad, keyword, or webpage is. In our case QS is for keywords.



## Related Work

- One approach: 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.[S. Ravi et al.]
- Another common approach: Start with Keyword Extraction then expand terms...
  - Synonymous words  $\rightarrow$  TermsNet and Wordy authors exploit the power of *search engines* to generate a portfolio of terms
- New measure: Nonobviousness [A. Joshi and R. Motwatni]

<http://kriti.net/>

**Table:** Holiday destination, vacation packages in Crete

Google Keyword Suggestion	Alchemy	AAC	RelatedToURL
hotels in crete heraklion	Crete	kreta hotel directory	hotels in crete heraklion
hotels on crete	mysterious rejuvenating force	kriti travel guide	hotels in crete heraklion
hotel in hungary	family vacation crete	great family vacation	hotels in crete heraklion
fira in santorini	excellent fresh fruits	hotel directory friday	hotels on crete
hotels in crete chania	offer distinctive excursions	vital trading center	hotels on crete
lodging in greece	popular holiday destinations	popular holiday destinations	hotels on crete
boutique hotel hungary	vital trading center	quaint family taverns	hotel in hungary
apartments in heraklion	Nikos Kazantzakis	find extensive information	hotel in hungary
apartments in chania	pre-classical civilization	vacation packages	hotel in hungary
crete heraklion hotels	Agios Nikolaos	crete hotels	fira in santorini
zara boutique hotel	excellent souvlaki	kreta hotel	
vouliagmeni suites	five-star restaurants	kriti travel	
accommodation santorini fira	countless areas	travel guide	
santorini fira apartments	Greek novelist	family holidays	
royal myconian hotel	European civilization	hotel directory	
santorini fira accommodation	major ports	family vacation	
hotels in crete	tourism infrastructure	book crete	
suites in santorini	quaint family	sitia crete	
hotel santorini fira	Chania ierapetra	crete	
hotels in santorini fira	rare plant	agios nikolaos	