Studying the Impact of Text Summarization on Contextual Advertising

Giuliano Armano, Alessandro Giuliani and Eloisa Vargiu
Dept. of Electric and Electronic Engineering
University of Cagliari
Cagliari, Italy
{*armano, alessandro.giuliani, vargiu*}*@diee.unica.it*

***Abstract*—Web advertising, one of the major sources of income for a large number of Web sites, is aimed at suggesting products and services to the ever growing population of Internet users. A significant part of Web advertising consists of textual ads, the ubiquitous short text messages usually marked as *sponsored links*. There are two primary channels for distributing ads: Sponsored Search (or Paid Search Advertising) and Content Match (or Contextual Advertising). In this paper, we concentrate on the latter, which is devoted to display commercial ads within the content of third-party Web pages. In the literature, several approaches estimated the ad relevance based on co-occurrence of the same words or phrases within the ad and within the page. However, targeting mechanisms based solely on phrases found within the text of the page can lead to problems. In order to solve these problems, matching mechanisms that combine a semantic phase with the traditional syntactic phase have been proposed. We are mainly interested in studying the impact of the syntactic phase on contextual advertising. In particular, we perform a comparative study on text summarization in contextual advertising. Results show that implementing effective text summarization techniques may help to improve the corresponding contextual advertising system.**

***Keywords*-contextual advertising, text summarization, comparative assessment**

I. Introduction

Online Advertising is an emerging research field, at the intersection of Information Retrieval, Machine Learning, Optimization, and Microeconomics. Its main goal is to choose the right ads to present to a user engaged in a given task, such as sponsored search advertising or contextual advertising. Sponsored search advertising, also called paid search advertising, displays ads on the page returned from a Web search engine following a query. Contextual Advertising (CA), also called content match, displays ads within the content of a generic, third party, Web page. A commercial intermediary, namely ad network, is usually in charge of optimizing the selection of ads with the twofold goal of increasing revenue (shared between publisher and ad network) and improving user experience. In other words, CA is a form of targeted advertising for ads appearing on Web sites or other media, such as content displayed in mobile browsers. The ads themselves are selected and served by automated systems based on the content displayed to the user.

In our view, a CA system encompasses four main tasks: (i) pre-processing; (ii) text summarization (TS); (iii) classification; and (iv) matching. In this paper we focus on Text Summarization (TS) and we present a study on its impact in CA. First, we make a comparative experimental analysis among 10 TS techniques. Then, we study how their affect a CA system.

The rest of the paper is organized as follows. In section II, we survey the related work, giving a general overview of previous works on CA and then focusing on TS in CA. Section III outlines the implemented system. In Section IV, we present comparative experiments on TS and evaluate their impact on CA. We discuss our findings and draw conclusions in Section V.

II. Background

A. Contextual Advertising

As discussed in [1], CA is an interplay of four players:

- The *advertiser* provides the supply of ads. As in traditional advertising, the goal of the advertisers can be broadly defined as the promotion of products or services.
- The *publisher* is the owner of the Web pages on which the advertising is displayed. The publisher typically aims to maximize advertising revenue while providing a good user experience.
- The *ad network* is a mediator between the advertiser and the publisher; it selects the ads to display on the Web pages. The ad network shares the advertisement revenue with the publisher.
- The *Users* visit the Web pages of the publisher and interact with the ads.

CA is the economic engine behind a large number of non-transactional sites on the Web. A main factor for the success in CA is the relevance to the surrounding scenario. Each solution for CA evolved from search advertising, where a search query matches with a bid phrase of the ad.

A natural extension of search advertising is extracting phrases from the target page and matching them with the bid phrases of ads. Yih et al. [2] proposed a system for phrase extraction, which uses a variety of features to determine the importance of page phrases for advertising purposes. To this

end, authors proposed a supervised approach that relies on a training set built using a corpus of pages in which relevant phrases have been annotated by hand.

Ribeiro-Neto et al. [3] examined a number of strategies to match pages and ads based on extracted keywords. They represented both pages and ads in a vector space and proposed several strategies to improve the matching process. The authors explored the use of different sections of ads as a basis for the vector, mapping both page and ads in the same space. Since there is a discrepancy between the vocabulary used in the pages and in the ads (the so called impedance mismatch), the authors improved the matching precision by expanding the page vocabulary with terms from similar pages.

In a subsequent work, Lacerda et al. [4] proposed a method to learn the impact of individual features using genetic programming. The results shown that genetic programming is able to find improved matching functions.

Broder et al. [1] classified both pages and ads into a given taxonomy and matched ads to the page falling into the same node of the taxonomy. Each node of the taxonomy is built as a set of bid phrases or queries corresponding to a certain topic. Results shown a better accuracy than that corresponding to the classic systems (i.e., systems based only on syntactic match). Let us also note that, to improve performances, this system could be used in conjunction with more general approaches.

Nowadays, ad networks need to deal in real time with a large amount of data, involving billions of pages and ads. Therefore, several constraints have to be taken into account for building CA systems. In particular, efficiency and computational costs are crucial factors in the choice of methods and algorithms. To this end, Anagnostopoulos et al. [5] presented a methodology for Web advertising in real time, focusing on the contributions of the different fragments of a Web page. This methodology allows to identify short but informative excerpts of the Web page by using several TS techniques, in conjunction with the model developed in [1].

Since bid phrases are basically search queries, another relevant approach is to view CA as a problem of query expansion and rewriting. Murdock et al. [6] considered a statistical machine translation model to overcome the problem of the impedance mismatch between pages and ads. To this end, they proposed and developed a system able to re-rank the ad candidates based on a noisy-channel model. In a subsequent work, Ciramita et al. [7] used a machine learning approach, based on the model described in [1], to define an innovative set of features able to extract the semantic correlations between the page and ad vocabularies.

B. Text Summarization in Contextual Advertising

Summarization techniques can be divided into two groups [8]: (i) those that extract information from the source documents (extraction-based approaches) and (ii) those that abstract from the source documents (abstraction-based approaches). The former impose the constraint that a summary uses only components extracted from the source document, whereas the latter relax the constraints on how the summary is created. Extraction-based approaches are mainly concerned with what the summary content should be, usually relying solely on extraction of sentences. On the other hand, abstraction-based approaches put strong emphasis on the form, aiming to produce a grammatical summary, which usually requires advanced language generation techniques. Although potentially more powerful, abstraction-based approaches have been far less popular than their extraction-based counterparts, mainly because it is easier to generate the latter.

As the input of a contextual advertiser is an HTML document, CA systems typically rely on extraction-based approaches, which are applied to the relevant blocks of a Web page.

In the work of Kolcz et al. [8] seven straightforward (but effective) extraction-based TS techniques have been proposed and compared. In all cases, a word occurring at least three times in the body of a document is a keyword, while a word occurring at least once in the title of a document is a title-word. For the sake of completeness, let us recall the proposed techniques:

- *Title* (T), the title of a document;
- *First Paragraph* (FP), the first paragraph of a document;
- *First Two Paragraphs* (F2P), the first two paragraphs of a document;
- *First and Last Paragraphs* (FLP), the first and the last paragraphs of a document;
- *Paragraph with most keywords* (MK), the paragraph that has the highest number of keywords;
- *Paragraph with most title-words* (MT), the paragraph that has the highest number of title-words;
- *Best Sentence* (BS), sentences in the document that contain at least 3 title-words and at least 4 keywords.

One may argue that the above methods are too simple. However, as shown in [9], extraction-based summaries of news articles can be more informative than those resulting from more complex approaches. Also, headline-based article descriptors proved to be effective in determining user's interests [10].

Furthermore, in [11] we proposed to enrich some of the techniques introduced by Kolcz et al. with information extracted from the title, as follows:

- *Title and First Paragraph* (TFP), the title of a document and its first paragraph:
- *Title and First Two Paragraphs* (TF2P), the title of a document and its first two paragraphs;
- *Title, First and Last Paragraphs* (TFLP), the title of a document and its first and last paragraphs;

- *Most Title-words and Keywords* (MTK), the paragraph with the highest number of title-words and that with the highest number of keywords.

We also defined a further technique, called *NKeywords* (NK), which selects the N most frequent keywords.[1].

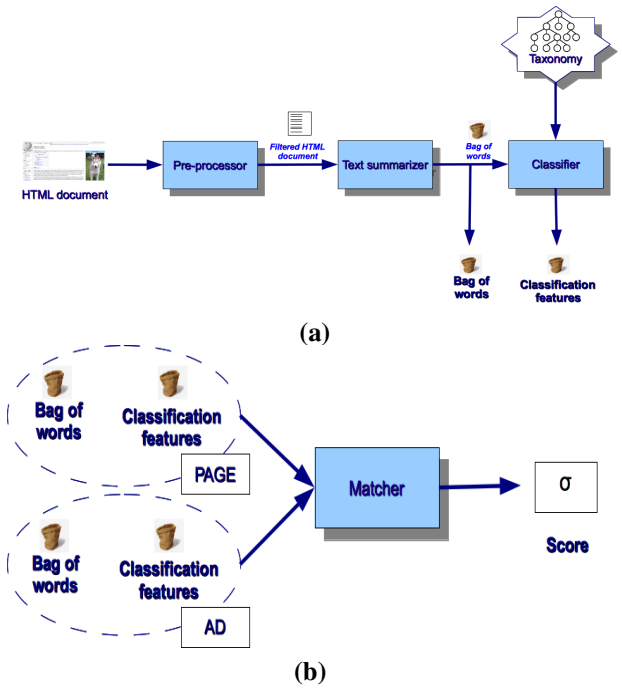

Figure 1. The architecture of the implemented system.

III. The Adopted System

Our view of CA is sketched in Figure 1, which illustrates a generic architecture that can give rise to specific systems depending on the choices made on each specific module. Notably, most of the state-of-the-art solutions are compliant with this view.

We implemented that architecture in Java and we use the corresponding system to perform our comparative experiments (see Section IV-C).

Pre-processor. Its main purpose is to transform an HTML document (a Web page or an ad) into an easy-to-process document in plain-text format, while maintaining important information. This is obtained by preserving the blocks of the original HTML document, while removing HTML tags and stop-words[2]. First, any given HTML page is parsed to identify and remove noisy elements, such as tags, comments and other non-textual items. Then, stop-words are removed from each textual excerpt. Finally, the document is tokenized and each term stemmed using the Porter's algorithm [12].

Text summarizer. It outputs a vector representation of the original HTML document as bag of words (BoW), each word being weighted by TF-IDF [13].

Classifier. TS is a purely syntactic analysis and the corresponding Web-page classification is usually inaccurate. To alleviate possible harmful effects of summarization, both page excerpts and advertisings are classified according to a given set of categories [5]. The corresponding classification-based features (CF) are then used in conjunction with the original BoW. In the current implementation, we adopt a centroid-based classification technique [14], which represents each class with its centroid calculated starting from the training set. A page is classified measuring the distance between its vector and the centroid vector of each class by adopting the cosine similarity.

Matcher. It is devoted to suggest ads (a) to the Web page (p) according to a similarity score based on both BoW and CF [5]. In formula (α is a global parameter that permits to control the emphasis of the syntactic component with respect to the semantic one):

$$score(p,a) = \alpha \cdot sim_{BoW}(p,a) + (1-\alpha) \cdot sim_{CF}(p,a) \quad (1)$$

where, $sim_{BoW}(p,a)$ and $sim_{CF}(p,a)$ are cosine similarity scores between p and a using BoW and CF, respectively.

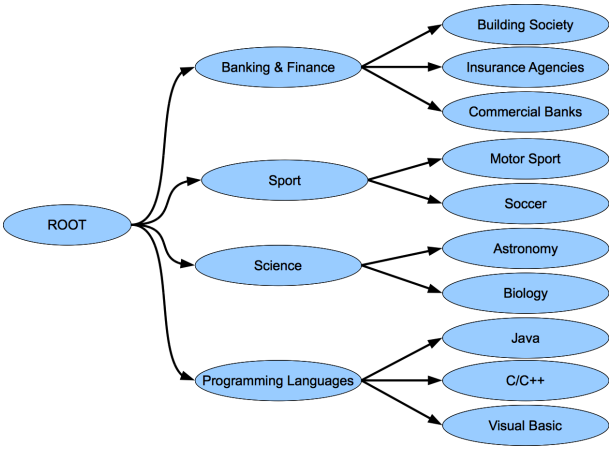

Figure 2. The taxonomy of BankSearch Dataset.

IV. Experimental Results

A. The Adopted Dataset

To perform experiments we used the BankSearch Dataset [15], built using the Open Directory Project and Yahoo! Categories[3], consisting of about 11000 Web pages classified

[1]N is a global parameter that can be set starting from some relevant characteristics of the input (e.g., from the average document length).

[2]To this end, the Jericho API for Java has been adopted, described at the Web page: http://jericho.htmlparser.net/docs/index.html

[3]http://www.dmoz.org and http://www.yahoo.com, respectively.

by hand in 11 different classes. Figure 2 shows the overall taxonomy. The 11 selected classes are the leaves of the taxonomy, together with the class *Sport*, which contains Web documents from all the sites that were classified as *Sport*, except for the sites that were classified as *Soccer* or *Motor Sport*. In [15], the authors show that this structure provides a good benchmark not only for generic classification/clustering methods, but also for hierarchical techniques.

To perform comparative experiments on CA, we also built a repository of ads, composed of 5 relevant company Web pages for each class of the adopted taxonomy. In so doing, there are 55 different ads in the repository.

B. Comparative Experiments on Text Summarization

To evaluate the TS techniques we used the same classifier described in Section III, in which each class is represented by its centroid. The pages are classified by considering the highest score(s) obtained by the cosine similarity method. First, in order to evaluate the effectiveness of the classifier, we performed a preliminary experiment in which pages are classified without resorting to TS. The classifier shown a precision of 0.862 and a recall of 0.858.

Then, we performed comparative experiments among the methods of Kolcz et al. (see Section II-B), except "Best Sentence" [4] and the corresponding enriched TS techniques proposed in [11].

Table I
COMPARATIVE RESULTS ON TS TECHNIQUES.

	P	R	F1	T
T	**0.798**	0.692	0.729	3
FP	0.606	0.581	0.593	13
F2P	0.699	0.673	0.686	24
FLP	0.745	**0.719**	**0.732**	24
MK	0.702	0.587	0.639	25
MT	0.717	0.568	0.634	15
TFP	0.802	0.772	0.787	16
TF2P	0.822	0.789	0.805	27
TFLP	**0.832**	**0.801**	**0.816**	26
MTK	0.766	0.699	0.731	34

Table I shows the performances in terms of macro-precision (P), macro-recall (R), and F-measure (F1). For each technique, the average number of unique extracted terms (T) is shown.

As already noted in [11], just adding information about the title improves the performances of TS.

C. The Impact of Text Summarization in Contextual Advertising

To study the impact of TS in CA, we used the system described in Section III, comparing results while varying the adopted TS technique. Let us note that, even if modern advertising networks should work in real time, for the sake of completeness we calculated performances without exploiting TS (the first row in Table II).

CA systems choose the relevant ads contained in the repository according to the scores obtained by the Matcher. The ads with the highest scores are displayed on the target page.

Three different experiments have been performed for each system, in which 1, 3, and 5 ads are selected for the target page, respectively. Table II reports, for each TS technique, the precision at k ($k = 1, 3, 5$) in correspondence of the best value of α (α_b).

Table II
COMPARATIVE RESULTS ON CA.

	p@1	α_b	p@3	α_b	p@5	α_b
no TS	**0.785**	0.0	**0.775**	0.0	**0.753**	0.2
T	**0.680**	0.1	**0.652**	0.1	**0.595**	0.3
FP	0.488	0.2	0.448	0.3	0.391	0.1
F2P	0.613	0.2	0.588	0.1	0.514	0.3
FLP	0.674	0.0	0.617	0.2	0.546	0.1
MK	0.631	0.2	0.581	0.1	0.500	0.3
MT	0.640	0.4	0.610	0.2	0.547	0.3
TFP	0.744	0.0	0.691	0.1	0.637	0.1
TF2P	0.740	0.0	0.721	0.1	**0.678**	0.0
TFLP	**0.768**	0.2	**0.729**	0.3	0.663	0.0
MTK	0.711	0.2	0.685	0.2	0.608	0.2

As expected, the best results are obtained without adopting any TS technique. Among the selected techniques, TFLP has the best performances in terms of $p@1$ and $p@3$, whereas TF2P in terms of $p@5$.

To further highlight the impact of TS, Table III reports the results obtained by using TFLP while varying α. According to equation (1), a value of 0.0 means that only semantic analysis is considered, whereas a value of 1.0 considers only the syntactic analysis.

Table III
RESULTS WITH TFLP BY VARYING α.

α	p@1	p@3	p@5
0.0	0.765	0.719	**0.663**
0.1	0.767	0.724	**0.663**
0.2	**0.768**	**0.729**	0.662
0.3	0.766	**0.729**	0.661
0.4	0.756	**0.729**	0.658
0.5	0.744	0.721	0.651
0.6	0.721	0.703	0.640
0.7	0.685	0.680	0.625
0.8	0.632	0.634	0.586
0.9	0.557	0.548	0.512
1.0	0.408	0.372	0.640

Results show that the more α increases, the more the precision decreases. This means that the impact of classification (i.e., of CF) is stronger than that of TS (i.e., of BoW). Nevertheless, except for $p@5$, the adoption of a good TS technique improves the precision of the corresponding system.

[4]This method was defined to extract summaries from textual documents such as articles, scientific papers and books. In fact, we are interested in summarizing HTML documents, which are often too short to find meaningful sentences composed by at least 3 title-words and 4 keywords in the same sentence.

V. Conclusions and Future Work

In this paper, we presented a comparative study on TS techniques applied to CA. In particular, we considered some straightforward extraction-based techniques that improve those proposed in the literature, and we evaluated the corresponding CA systems. Experimental results confirm the intuition that adopting TS techniques allows to improve performances in term of precision.

As for future directions, further experiments are also under way. In particular, we are setting up the system to calculate its performances with a larger dataset extracted by DMOZ.

Acknowledgment

This work has been partially supported by Hoplo srl. We wish to thank, in particular, Ferdinando Licheri and Roberto Murgia for their help and useful suggestions.

References

[1] A. Broder, M. Fontoura, V. Josifovski, and L. Riedel, "A semantic approach to contextual advertising," in *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2007, pp. 559–566.

[2] W.-t. Yih, J. Goodman, and V. R. Carvalho, "Finding advertising keywords on web pages," in *WWW '06: Proceedings of the 15th international conference on World Wide Web*. New York, NY, USA: ACM, 2006, pp. 213–222.

[3] B. Ribeiro-Neto, M. Cristo, P. B. Golgher, and E. Silva de Moura, "Impedance coupling in content-targeted advertising," in *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2005, pp. 496–503.

[4] A. Lacerda, M. Cristo, M. A. Gonçalves, W. Fan, N. Ziviani, and B. Ribeiro-Neto, "Learning to advertise," in *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2006, pp. 549–556.

[5] A. Anagnostopoulos, A. Z. Broder, E. Gabrilovich, V. Josifovski, and L. Riedel, "Just-in-time contextual advertising," in *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. New York, NY, USA: ACM, 2007, pp. 331–340.

[6] V. Murdock, M. Ciaramita, and V. Plachouras, "A noisy-channel approach to contextual advertising," in *Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising*, ser. ADKDD '07. New York, NY, USA: ACM, 2007, pp. 21–27.

[7] M. Ciaramita, V. Murdock, and V. Plachouras, "Online learning from click data for sponsored search," in *Proceeding of the 17th international conference on World Wide Web*, ser. WWW '08. New York, NY, USA: ACM, 2008, pp. 227–236.

[8] A. Kolcz, V. Prabakarmurthi, and J. Kalita, "Summarization as feature selection for text categorization," in *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*. New York, NY, USA: ACM, 2001, pp. 365–370.

[9] R. Brandow, K. Mitze, and L. F. Rau, "Automatic condensation of electronic publications by sentence selection," *Inf. Process. Manage.*, vol. 31, pp. 675–685, September 1995.

[10] A. Kołcz and J. Alspector, "Asymmetric missing-data problems: Overcoming the lack of negative data in preference ranking," *Inf. Retr.*, vol. 5, pp. 5–40, January 2002.

[11] G. Armano, A. Giuliani, and E. Vargiu, "Experimenting text summarization techniques for contextual advertising," in *IIR'11: Proceedings of the 2nd Italian Information Retrieval (IIR) Workshop*, 2011.

[12] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[13] G. Salton and M. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, 1984.

[14] E.-H. Han and G. Karypis, "Centroid-based document classification: Analysis and experimental results," in *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, ser. PKDD '00. London, UK: Springer-Verlag, 2000, pp. 424–431.

[15] M. Sinka and D. Corne, "A large benchmark dataset for web document clustering," in *Soft Computing Systems: Design, Management and Applications, Volume 87 of Frontiers in Artificial Intelligence and Applications*. Press, 2002, pp. 881–890.