Are Related Links Effective for Contextual Advertising?
A Preliminary Study

Giuliano Armano[1], Alessandro Giuliani[1] and Eloisa Vargiu [1,2]
[1]*Dept. of Electrical and Electronic Engineering, University of Cagliari, Piazza d'Armi, I09123, Cagliari, Italy*
[2]*Barcelona Digital Technological Center, Carrer Roc Boronat, 117, E08018, Barcelona, Spain*

Keywords: Collaborative Filtering, Contextual Advertising, Related Links.

Abstract: Classical contextual advertising systems suggest suitable ads to a given webpage just analyzing its content, without relying on further information. We claim that adding some information extracted by semantically related pages can improve the overall performances. To this end, this paper proposes an experimental study aimed at verifying to which extent the analysis of related links, i.e., inlinks and outlinks, can help contextual advertising. Experiments have been performed on about 15000 webpages extracted by DMoz. Results show that the adoption of related links significantly improves the performance of the adopted baseline system.

1 INTRODUCTION

Contextual Advertising (CA), also called Content Match, is aimed at suggesting to a webpage ads that are related to the content of it. The proposed state-of-the-art CA systems infer the context of a given webpage p by analyzing its content, without relying on any further information. 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. As a CA task can be also viewed as a recommendation task (Armano and Vargiu, 2010), we claim that CA systems can be improved by using collaborative filtering through the extraction of suitable information from semantically related links.

Marchiori (Marchiori, 1997) states that *'The power of the Web resides in its capability of redirecting the information flow via hyperlinks, so it should appear natural that in order to evaluate the information content of a Web object, the Web structure has to be carefully analyzed'*. The benefits of link information for information retrieval have been well researched (Koolen and Kamps, 2011). Link-based ranking algorithms use the implicit assumption that linked documents tend to be related each other and, therefore, that link information is potentially useful for retrieval and filtering (Cohen and Kjeldsen, 1987) (Kleinberg, 1999) (Lempel and Moran, 2001) (Shakery and Zhai, 2006).

In this paper, we present an experimental study aimed at investigating whether or not related links are effective for CA. To our best knowledge, this is the first attempt to assess the effectiveness of semantically related links in the field of CA. We consider as related links of a webpage p: its *inlinks* (also called *backlinks*), i.e., pages that link to p; and its *outlinks* (also called *inbound* and *outbound links* depending on the corresponding domain), i.e., pages that are linked by p. The motivation why we rely on related links is that if a page q links a page p, at least in principle, the topics of q are related to the topics of p (Koolen and Kamps, 2011). To assess whether related links are useful to improve CA systems, we developed a suitable CA system and performed several experiments on it. Results show that adopting related links significantly improves the performance of the adopted CA system.

The rest of the paper is organized as follows: Section 2 illustrates the approach adopted to assess whether related links are effective for CA. In Section 3, we report and discuss the experiments and their results. In Section 4, related work is recalled. Section 5 ends the paper with conclusions and future work.

2 METHODOLOGY

Given a webpage p, a CA system analyzes it in order to suggest suitable ads. As sketched in Figure 1, a CA task typically involves three phases: text summariza-

Armano G., Giuliani A. and Vargiu E..
Are Related Links Effective for Contextual Advertising? - A Preliminary Study.
DOI: 10.5220/0004135802210226
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2012), pages 221-226
ISBN: 978-989-8565-29-7

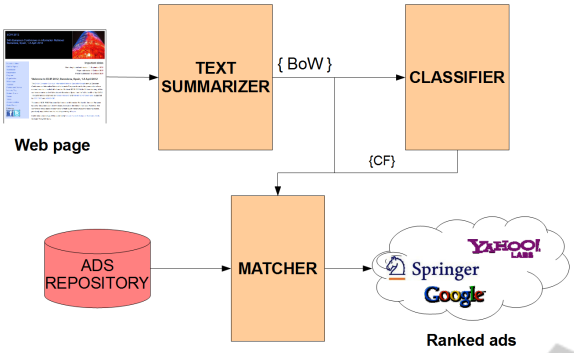

Figure 1: A generic approach to CA.

tion, classification, and matching. *Text summarization* is aimed at generating a short representation of p with a minor loss of information. This representation is typically given in terms of Bag of Words (*BoW*), in which each term is typically weighted by the TFIDF (Salton and McGill, 1984). *Classification* is devoted to alleviate possible harmful effects of summarization. To this end both page excerpts and ads are classified according to a relevant set of categories, usually organized in a taxonomy, giving as output the so-called Classification-based Features (*CF*). *Matching* is devoted to suggest ads to p according to a similarity score based on both *BoW* and *CF*. Let us note that this model is compliant with most of state-of-the-art systems, including those proposed in (Broder et al., 2007) (Anagnostopoulos et al., 2007) (Armano et al., 2011).

To study the role of related links in CA, we devised the model depicted in Figure 2. The proposed model embodies four modules: *Related Link Extractor*, *Text Summarizer*, *Classifier*, and *Matcher*. Notably, the model is compliant with that reported in Figure 1 in which only CF are considered in the matching phase. In particular, they coincide on everything, but the *related link extractor*.

Let us recall that with "related links" of a webpage p we denote both inlinks and outlinks. Figure 3 gives a view of two different kinds of related links: those

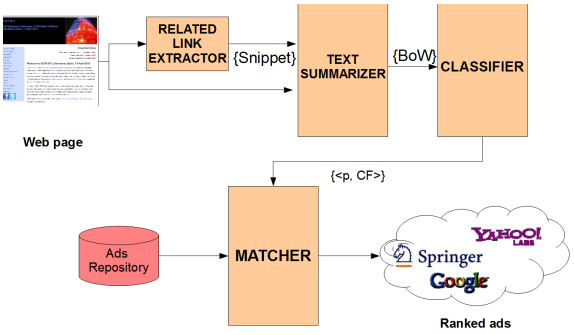

Figure 2: The model of the adopted approach.

that link to an external domain (i.e., from A and to B in the Figure) and those that link to the same domain of the target webpage (i.e., from $T1$ and to $T2$ in the Figure). Without loss of generality, in this work, we consider only links belonging to different domains; in other words, we intentionally disregard inlinks that come from the same Web domain and inbounds (e.g., in the example reported in Figure, we do not consider $T1$ and $T2$).

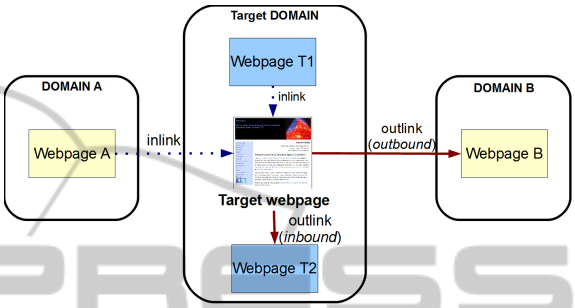

Figure 3: A graphical view of the adopted related links.

Related Link Extractor. This module extracts the related links of a given webpage p. It collects the set of inlinks and the set of outlinks of the webpage. The former is built by performing a query to a special service of Yahoo! Search (http://siteexplorer.search.yahoo.com)[1], the performed query being the URL of the page, in order to ask for the inlinks. The latter is built by parsing p ()looking for the anchor tag $< a >$)[2]. The module selects the URL of the first 10 inlinks and the first 10 outlinks, if available. For the sake of experimental reproducibility we decided to consider the first 10 outlinks rather than a random selection.

Text Summarizer. Instead of considering the whole content of pages, we first extract the snippets of, respectively, p, its inlinks, and its outlinks by asking to the Yahoo! search engine (http://www.yahoo.com). The content of each query is the URL of the link under analysis. Summarizing, the main purpose of this module is to remove stopwords and to stem each term through the Porter's algorithm (Porter, 1980).The output is a vector representation of the original text as *BoW*, each word being weighted by TFIDF (TFIDF score is computed over a training collection of webpage snippets).

Classifier. To infer the topics related to each inlink, outlink, or page in hand, snippets are classified according to a given taxonomy. First, for each node

[1]Currently it is merged into Bing Webmaster Tools (http://www.bing.com/toolbox/webmaster).

[2]To minimize the impact of "general-purpose" websites, we filter links such as Facebook, Twitter, Google+, and so on.

of the taxonomy, we merge all its training documents into a single compound document. We then use it as a centroid for the Rocchio classifier (Rocchio, 1971) with only positive examples and no relevance feedback. Each centroid is defined as a sum of the TFIDF values of each term, normalized by the number of the training documents in the class. Snippet classification is based on the cosine of the angle between the snippet and the centroid of the class. The classifier outputs the snippet-category matrix, whose generic element w_{ij} reports the score given by the classifier for the similarity between the category j and the snippet i.

Matcher. This module is devoted to suggest ads to the webpage, according to the given taxonomy. First, for each column of the page-category matrix, the matcher calculates the sum of scores as follows:

$$\sigma_j = \sum_{i=1}^{N} w_{ij} \tag{1}$$

where N is the total number of extracted pages. Then, the *Matcher* selects k categories, i.e., those with the highest values of σ, k being dependent on the agreement between publisher and advertiser. Finally, for each selected category, an ad is randomly extracted from the *Ads repository*[3].

3 EXPERIMENTAL RESULTS

We propose an automatic evaluation algorithm, based on the associated categories for both page p and ad a (categories should belong to the adopted taxonomy). Given a page p and an ad a, the $\langle p, a \rangle$ pair has been scored on a 1 to 3 scale, defined as follows:

1 - Relevant: a is semantically and directly related to the main subject of p (i.e.,p and a belong to the same node of the taxonomy);

2 - Somewhat Relevant: (i) a is related to a similar topic of p (*sibling* nodes), (ii) a is related to the main topic of p in a more general way (*generalization*, p category is child of a category), or (iii) a is related to the main topic of p in a too specific way (*specification*, a category is child of p category);

3 - Irrelevant. a is unrelated to p.

According to the existing literature(e.g., (Broder et al., 2007)), we considered as True Positives (TP) ads scored as 1 or 2, and as False Positives (FP) ads scored as 3. Performances have been calculated in terms of *precision at k* ($\pi@k$) with $k \in [1,5]$ (i.e., the precision in suggesting k ads), as follows:

$$\pi@k = \frac{\sum_{i=1}^{N}\sum_{j=1}^{k} TP_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{k}(TP_{ij}+FP_{ij})} \tag{2}$$

Since we rely on a graded relevance scale of evaluation, to measure the effectiveness of the approach we adopt two further evaluation metrics: the *Normalized Discounted Cumulative Gain (nDCG)* and the *Expected Reciprocal Rank (ERR)*. The former measures the usefulness, or gain, of the suggested ad categories based on its position in the result list (Järvelin and Kekäläinen, 2000). The latter is defined as the expected (inverse) rank at which the user will stop and click the associated ad. In our case it considers the inverse ranking of the first relevant category (Chapelle et al., 2009).

Experiments have been performed on a total of about 15000 webpages extracted by a subset of DMoz[4]. In particular, we selected 18 categories, all belonging to the root *Recreation*. The taxonomy depth equals to 3. All the systems embed the same *Classifier* that has been firstly trained by using about 100 webpages per class.

As for the ads to be suggested, we manually built a suitable repository in which ads are classified according to the given taxonomy. In that repository each ad is represented by the webpage of a product or service company (landing page). The ad repository contains 135 ads.

We made experiments aimed at evaluating to which extent related links affect the performance of a CA system. Due to the two-tiered nature of a related link (inlink or outlink), we also separately evaluated the contributions of inlinks and outlinks. Summarizing, experiments have been performed by taking into account:

- the webpage alone (**P**)[5];
- the set of inlinks alone (**I**);
- the set of outlinks alone (**O**);
- the webpage in conjunction with its inlinks (**P+I**);
- the webpage in conjunction with its outlinks (**P+O**);
- the set of related links (inlinks and outlinks) alone (**RL**);

[3]We assume that a repository of ads is available, in which company or service webpages are classified according to the given taxonomy.

[4]http://www.dmoz.org

[5]baseline system, compliant with most state-of-the-art systems

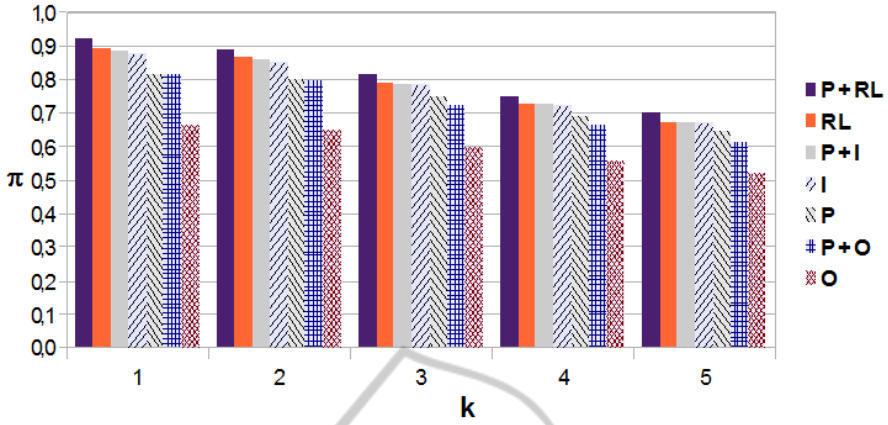

Figure 4: Precision at k.

- the webpage in conjunction with its related links (**P+RL**).

We report in the Figure 4 the precisions of each system.

The first step was concerned with evaluating the performance of the baselin system (P). The second step was concerned with evaluating the contribution of inlinks. Figure 4 shows that ***P+I*** leads to an improvement with respect to ***I*** and ***P***. Notably, the introduction of inlinks leads to an increment of about 7.3%, whereas the introduction of the page with respect to the sole inlinks leads to an increment of about 0.6%.

We then evaluated the impact of outlinks. Figure 4 shows that ***P*** performs better than ***O*** and than ***P+O***. Hence, the adoption of outlinks leads to a remarkable decrease of performance.

Finally, we evaluated the conjuncted adoption of inlinks and outlinks (the related links). Figure 4 shows that the best performances are always obtained by ***P+RL***. Notably, the introduction of related links with respect ***P*** leads to an increment of about 9.2%, whereas the introduction of the page with respect ***RL*** leads to an increment of about 1.4%.

Table 1 reports the performances of each system, in suggesting 5 ads, in terms of nDCG and ERR. It confirms the behavior of each system, giving rise to the contribution of related links, confirming the assumption that linked documents have related content.

Table 1: nDCG and ERR of each approach in suggesting 5 ads.

	P	I	O	RL	P+I	P+O	P+RL
nDCG	0.839	0.843	0.748	0.861	0.844	0.834	**0.875**
ERR	0.552	0.567	0.409	0.584	0.574	0.539	**0.597**

Summarizing, it is clear, from the results, how the adoption of related links increases the performances. They also put into evidence that the main contribution is given by the introduction of inlinks. The differences between the adoption of inlinks and the adoption of outlinks could be due to the different amount of examined links. In particular, for a given webpage, it is easy to find at least 10 inlinks, while it is more difficult that the webpage contains at least the same number of outlinks. In fact, we analyzed the dataset in order to compute the number of outlinks per page, and we found that a webpage contains an average number of 3.6 inlinks.

4 RELATED WORK

4.1 Semantically Related Links

Many researchers investigated the role of links in information retrieval, see, for example, (Marchiori, 1997). In particular, links have been used to (i) enhance document representation (Picard and Savoy, 2003), (ii) improve document ranking by propagating document score (Frei and Stieger, 1995), (iii) provide an indicator of popularity (Brin and Page, 1998), and (iv) find hubs and authorities for a given topic (Chakrabarti et al., 1999).

It is worth noting that a key problem in this research field is how to measure the semantic relatedness of documents, see (Budanitsky and Hirst, 2006) for a survey. In this work, we just propose a preliminary study on the impact of related links in CA without taking into account this problem.

4.2 Contextual Advertising

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. (Yih et al., 2006) proposed a system for phrase extraction to determine the importance of page phrases for advertising purposes. Since the repository of ads adopted in our work is composed by webpages of companies, we do not take into account the phrase extraction but rely only on extraction-based text summarization by using the snippets provided by Yahoo!.

Ribeiro-Neto et al. (Ribeiro-Neto et al., 2005) 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. According to their work, we represent both pages and ads in a vector space and we consider the contribution of similar pages. Nevertheless, since our ads are actually webpages, we do not take into account the impedance mismatch.

Broder et al. (Broder et al., 2007) improved the performance of CA by classifying both pages and ads according to a given taxonomy and matching 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. We adopted the same Rocchio classifier, according to their results and taking into account the effectiveness of adopting CF.

Nowadays, ad networks need to deal in real time with a large amount of data, involving billions of pages and ads. Therefore, several constraints must be taken into account for building CA systems. In particular, efficiency and computational costs are crucial factors in the choice of methods and algorithms. In order to analyze the entire body of webpages on-the-fly, state-of-the-art systems use text summarization techniques (Anagnostopoulos et al., 2007) (Armano et al., 2011). Our choice for text summarization exploits the snippet provided by Yahoo!. The effectiveness of using snippets as text summarization techniques has been proved in (Armano et al., 2012).

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., 2007) (Ciaramita et al., 2008). According to this view, we assess the performance of our approach by adopting $\pi@k$, a classical measure adopted in query search, which in this case represents the capability of the system to suggest k relevant ads to a webpage.

Since the task of suggesting an ad to a webpage can be also viewed as the task of recommending an item (the ad) to a user (the webpage), another perspective consists on addressing a CA problem as a recommendation task (Armano and Vargiu, 2010). According to this view, the system developed to perform the experiments can be though as a hybrid recommender system in which collaborative filtering is used in a content-based setting.

5 CONCLUSIONS AND FUTURE WORK

This paper was aimed at assessing whether or not *related links can be effective for contextual advertising*. To this end, we conducted a preliminary experimental study aimed at investigating the impact of related links in a generic contextual advertising system. Experiments have been performed considering the impact of related links. Results clearly show that the adoption of related links increases the performances. They also put into evidence that the main contribution is given by the introduction of inlinks. Hence, we can sentence that related links are effective for contextual advertising.

As for future work, we are implementing a system in which the matching is performed by taking into account the information of each ad rather than the ad category. Moreover, we are studying how to improve the proposed approach by considering also the title and the url of each page, in conjunction with its snippet.

ACKNOWLEDGEMENTS

We wish to thank Massimo Puddu, Ugo Sirca and Alessandro Zou for their support in developing the system.

REFERENCES

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

Armano, G., Giuliani, A., and Vargiu, E. (2011). Studying the impact of text summarization on contextual advertising. In *8th International Workshop on Text-based Information Retrieval*.

Armano, G., Giuliani, A., and Vargiu, E. (2012). Using snippets in text summarization: a comparative study

and an application. In *IIR'12: 3rd Italian Information Retrieval (IIR) Workshop*.

Armano, G. and Vargiu, E. (2010). A unifying view of contextual advertising and recommender systems. In *Proceedings of International Conference on Knowledge Discovery and Information Retrieval (KDIR 2010)*, pages 463–466.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30:107–117.

Broder, A., Fontoura, M., Josifovski, V., and Riedel, L. (2007). 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*, pages 559–566, New York, NY, USA. ACM.

Budanitsky, A. and Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Comput. Linguist.*, 32:13–47.

Chakrabarti, S., van den Berg, M., and Dom, B. (1999). Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11–16):1623–1640.

Chapelle, O., Metlzer, D., Zhang, Y., and Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 621–630, New York, NY, USA. ACM.

Ciaramita, M., Murdock, V., and Plachouras, V. (2008). Online learning from click data for sponsored search. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 227–236, New York, NY, USA. ACM.

Cohen, P. R. and Kjeldsen, R. (1987). Information retrieval by constrained spreading activation in semantic networks. *Information Processing and Management*, 23(4):255–268.

Frei, H. P. and Stieger, D. (1995). The use of semantic links in hypertext information retrieval. *Inf. Process. Manage.*, 31:1–13.

Järvelin, K. and Kekäläinen, J. (2000). Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 41–48, New York, NY, USA. ACM.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of ACM*, 46:604–632.

Koolen, M. and Kamps, J. (2011). Are semantically related links more effective for retrieval? In *Proceedings of the 33rd European conference on Advances in information retrieval*, ECIR'11, pages 92–103, Berlin, Heidelberg. Springer-Verlag.

Lempel, R. and Moran, S. (2001). SALSA: the stochastic approach for link-structure analysis. *ACM Transactions on Information Systems*, 19:131–160.

Marchiori, M. (1997). The quest for correct information on the web: hyper search engines. *Comput. Netw. ISDN Syst.*, 29:1225–1235.

Murdock, V., Ciaramita, M., and Plachouras, V. (2007). A noisy-channel approach to contextual advertising. In *Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising*, ADKDD '07, pages 21–27, New York, NY, USA. ACM.

Picard, J. and Savoy, J. (2003). Enhancing retrieval with hyperlinks: a general model based on propositional argumentation systems. *J. Am. Soc. Inf. Sci. Technol.*, 54:347–355.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Ribeiro-Neto, B., Cristo, M., Golgher, P. B., and Silva de Moura, E. (2005). 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*, pages 496–503, New York, NY, USA. ACM.

Rocchio, J. (1971). *The SMART Retrieval System: Experiments in Automatic Document Processing*, chapter Relevance feedback in information retrieval, pages 313–323. PrenticeHall.

Salton, G. and McGill, M. (1984). *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company.

Shakery, A. and Zhai, C. X. (2006). A probabilistic relevance propagation model for hypertext retrieval. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 550–558, New York, NY, USA. ACM.

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