SEMANTIC ENRICHMENT OF CONTEXTUAL ADVERTISING BY USING CONCEPTS

Giuliano Armano, Alessandro Giuliani and Eloisa Vargiu
Dept. of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy

Keywords: Contextual advertising, Recommender systems, Information filtering.

Abstract: This paper focuses on Contextual Advertising, which is devoted to display commercial ads within the content of third-party Web pages. In the literature, several approaches estimate the relevance of an ad based only on syntactic approaches. However, these approaches may lead to the choice of a remarkable number of irrelevant ads. In order to solve these drawbacks, solutions that combine a semantic phase with a syntactic phase have been proposed. Framed within this approach, we propose an approach that uses to a semantic network able to supply commonsense knowledge. To this end, we developed and implemented a system that uses the ConceptNet 3 database. To our best knowledge this is the first attempt to use information provided by ConceptNet in the field of Contextual Advertising. Several experiments have been performed aimed at comparing the proposed system with a state-of-the-art system. Preliminary results show that the proposed system performs better.

1 INTRODUCTION

Nowadays, Web advertising is one of the major sources of income for a large number of websites. Its main goal is to suggest 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 Contextual Advertising (or Content Match). Sponsored Search advertising displays ads on the page returned from a Web search engine following a query; whereas Contextual Advertising (CA) 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 and improving user experience. The ads are selected and served by automated systems based on the content displayed to the user.

In the literature, several approaches to CA estimate the ad relevance based on co-occurrences of the same words or phrases within the ad and within the page. However, it is easy to note that targeting mechanisms based solely on phrases found within the text of the page can lead to problems. In order to solve them, matching mechanisms that combine a semantic phase with the traditional syntactic phase have been proposed (Broder et al., 2007).

In this paper, we propose a further semantic enrichment that exploits ConceptNet (Liu and Singh, 2004). To this end, we devised and implemented a system called ConCA (Concepts in Contextual Advertising). Experiments have been performed on the BankSearch Dataset (Sinka and Corne, 2002) and the baseline is the system proposed in (Armano et al., 2011b). Preliminary results show that the proposed system performs slightly better in terms of precision, while preserving the real-time constraint.

The rest of the paper is organized as follows. Section 2 recalls the CA problem. Section 3 presents a typical solution adopted in practice that is based on both a syntactic and a semantic approach. After recalling ConceptNet, Section 4 presents the proposed approach. In Section 5, we present our experimental results. In Section 6, we survey the most relevant work on CA and we compare them with the proposed solution. Section 7 ends the paper with conclusions and an overview of future work.

2 CONTEXTUAL ADVERTISING

CA is an interplay of four players (Broder et al., 2007): (i) the *advertiser*, who provides the supply of

Armano G., Giuliani A. and Vargiu E..
SEMANTIC ENRICHMENT OF CONTEXTUAL ADVERTISING BY USING CONCEPTS.
DOI: 10.5220/0003657502240229
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2011), pages 224-229
ISBN: 978-989-8425-79-9

ads and organizes her/his activity around campaigns which are defined by a set of ads with a particular temporal and thematic goal (e.g., sale of digital cameras during the holiday season); (ii) the *publisher*, i.e. the owner of the Web pages on which the advertising is displayed, who is aimed at maximizing advertising revenue while providing a good user experience; (iii) the *ad network*, which, as a mediator between the advertiser and the publisher, selects the ads to display on the Web pages; and (iv) the *users*, who visit the Web pages of the publisher and interact with the ads.

Upon a request initiated by the user through an HTTP get request, the Web server returns the requested page. As the page is being displayed, a JavaScript code embedded into the page (or loaded from a server) sends to the ad network a request for ads that contains the page URL and some additional data. The ad network model aligns the interests of publishers, advertisers and the network itself. In general, user's clicks bring benefits to the publisher and the ad network by providing revenue, and to the advertiser by bringing traffic to the target web site.

Another perspective consists on addressing a CA problem as a recommendation task. In fact, the task of suggesting an ad to a Web page can be also viewed as the task of recommending an item (the ad) to a user (the Web page) (Armano and Vargiu, 2010).

3 A TYPICAL SOLUTION TO CONTEXTUAL ADVERTISING

Our view of a generic solution to CA encompasses four main tasks: (i) *pre-processing*; (ii) *text summarization*; (iii) *classification*; and (iv) *matching*. Notably, most of the state-of-the-art solutions are compliant with this view.

Pre-processing is mainly aimed at transforming an HTML document (a Web page or an ad) into an easy-to-process document in plain-text format, while maintaining important information.

Text summarization is aimed at generating a short representation of the Web page with a negligible loss of information.

To alleviate possible harmful effects of summarization, both page excerpts and ads are classified according to a given set of categories (Broder et al., 2007). The corresponding Classification-based Features (CF) are then used in conjunction with the original Bag of Words (BoW) provided by the text summarization phase (Anagnostopoulos et al., 2007).

Matching is devoted to suggest ads (a) to the Web page (p) according to a similarity score based on both BoW and CF. In formula:

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

in which, α is a global parameter that permits to control the impact of the syntactic component with respect to the semantic one, whereas $sim_{BoW}(p,a)$ and $sim_{CF}(p,a)$ are cosine similarity scores between p and a using BoW and CF, respectively.

4 IMPROVING CONTEXTUAL ADVERTISING WITH CONCEPTNET

As shown by Broder et al. (Broder et al., 2007), CA systems can be improved using semantic information. In particular, as recalled in the previous section, they consider CF in conjunction with the BoW. Focusing on the importance of semantics, we studied a further semantic enrichment that uses concepts in conjunction with CF. To this end, we exploited the semantic information provided by ConceptNet 3 (Havasi et al., 2007). The proposed system, called ConCA (Concepts in Contextual Advertising), integrates the ConceptNet 3 database in the generic solution summarized in Section 3.

4.1 ConceptNet

The Open Mind Common Sense (OMCS) project is a distributed solution to the problem of commonsense acquisition by enabling users to enter commonsense into the system with no special training or knowledge of computer science. In 2000, the OMCS project began to collect statements from untrained volunteers on the Internet. These data have been used to automatically build a semantic network, called ConceptNet (Liu and Singh, 2004).

In ConceptNet, nodes are concepts and edges are predicates. Concepts are aspects of the world that people would talk about in natural language. They correspond to selected constituents of the commonsense statements that users have entered, and can represent *noun phrases*, *verb phrases*, *adjective phrases*, or *prepositional phrases*. Predicates express relationships between two concepts. They are extracted from natural language statements enter by contributors, and express relationships such as *IsA*, *PartOf*, *LocationOf*, and *UserFor*. In addition, there are also some underspecified relation types such as *ConceptuallyRelatedTo*, which means that a relationship exists between two concepts without any other semantic explanation.

Predicates in ConceptNet are created by a pattern-matching process. Each sentence is compared with an ordered list of patterns, which are regular expressions

that can also include additional constraints on phrase types based on the output of a natural language tagger and chunker. These patterns represent sentence structures that are commonly used to express different relationships. The phrases that fill the slots in a pattern will be turned into concepts. When a sentence is matched against a pattern, the result is a "raw predicate" that relates two strings of text. A normalization process determines which two concepts these strings correspond to, turning the raw predicate into an edge of ConceptNet.

ConceptNet has been adopted in several application fields. To our best knowledge this is the first attempt to use ConceptNet in CA.

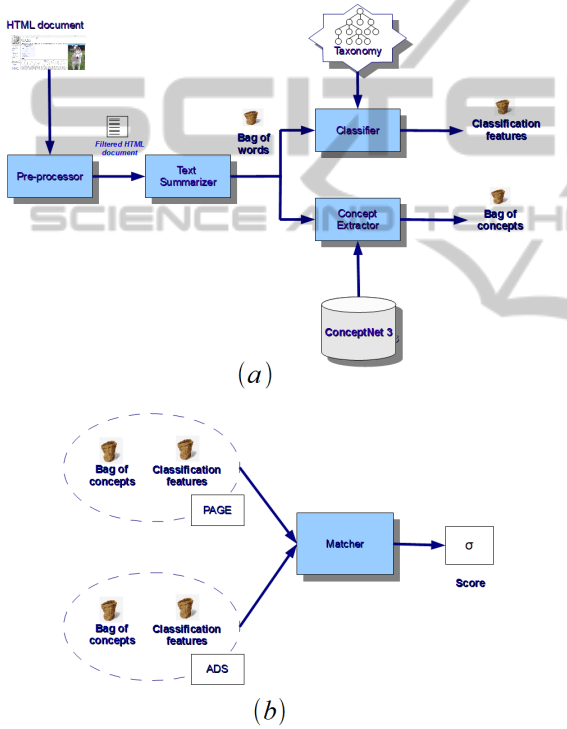

Figure 1: ConCA architecture.

4.2 ConCA

The system has been implemented in Java and, as shown in Figure 1, it encompasses five main modules: (i) *Pre-processor*; (ii) *Text Summarizer*; (iii) *Classifier*; (iv) *Concept Extractor*; (v) *Matcher*.

Pre-processor. To transform an HTML document into an easy-to-process document in plain-text format, the pre-processor removes HTML tags and stopwords[1]. First, any given HTML page is parsed to identify and remove noisy elements, such as tags, comments and other non-textual items. Then, stopwords are removed from each textual excerpt. Finally, the document is typically tokenized and each term stemmed.

Text Summarizer. It outputs a vector representation of the original HTML document as BoW, each word being weighted by TF-IDF. According to our previous comparative results (Armano et al., 2011b), we implemented the text summarization technique that has shown the best results, i.e., TFLP (Armano et al., 2011a). This technique takes into account information belonging to the Title, the First, and the Last Paragraph of the Web document (the page or the ad).

Classifier. To infer the topics of each page or ad, they are classified according to a given taxonomy. Each node of the taxonomy is first represented with its centroid. Each centroid component is defined as a sum of TF-IDF values of each component, normalized by the number of documents in the class. Then each document is classified by adopting the Rocchio classifier (Rocchio, 1971) with only positive examples and no relevance feedback.

Concept Extractor. It takes the BoW representation as input and, for each term, queries the ConceptNet 3 database to find the related meaningful concepts. Each element of the database is called *assertion*, and is uniquely defined by several attributes, for instance:

- a pair of concepts, i.e., two linked nodes of ConceptNet;
- the language of the concepts, e.g., English, Italian, or French;
- the type of relation that connects the two concepts, e.g., "IsA" or "PartOf";
- the score, i.e., a value given by the users that represents the reliability of the assertion;
- the frequency, i.e., a textual value (ranging from "never" to "always") that expresses how often a relation connects a given pair of concepts.

Conversely, each concept is also represented with several attributes, e.g., words, language, and number of assertions.

For each term, the Concept Extractor queries the ConceptNet 3 database in order to obtain the set of assertions that include the term, each being one of the two concepts in the assertion. The resulting assertion set is then filtered in order to reduce noise. In particular, we consider only the assertions that satisfy the following constraints:

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

- the language of the concepts is English;
- the type of the relation is "IsA" or "HasA";
- the score given by users is greater than 1;
- the frequency is "positive", e.g., assertions with "never" are not considered.

As for concepts, only those with a number of assertions greater than 3 are selected.

The output is a set of concepts with their occurrences, called Bag of Concepts (BoC). The BoC is represented in a vector space, in which each feature is weighted by TF-IDF.

Matcher. It is devoted to choose the relevant ads according to a score based on the similarity between the target page and each ad. Let us recall that in this work we are interested in semantically enriching information belonging to both pages and ads. To this end, we calculate the score similarity taking into account CF and BoC, without considering BoW. In formula:

$$\sigma(p,a) = \alpha \cdot sim_{BoC}(p,a) + (1-\alpha) \cdot sim_{CF}(p,a) \quad (2)$$

in which α is a global parameter that permits to control the impact of BoC with respect to CF, whereas $sim_{BoC}(p,a)$ and $sim_{CF}(p,a)$ are cosine similarity scores between p and a using BoC and CF, respectively.

5 EXPERIMENTS AND RESULTS

To perform experiments we used the BankSearch Dataset (Sinka and Corne, 2002), built using the Open Directory Project and Yahoo! categories[2], consisting of about 11000 Web pages classified by hand according to 11 different classes that belong to a given taxonomy. In (Sinka and Corne, 2002), the authors shown that this structure provides a good benchmark not only for generic classification/clustering methods, but also for hierarchical techniques.

To perform experiments, 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.

To evaluate the performances of ConCA, our baseline is the system adopted in (Armano et al., 2011b). Let us note that this baseline system is an implementation of the model described in Section 3, in which TFLP (Title, First and Last Paragraph summarization), is adopted as text summarization technique.

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

Five different experiments have been performed for each system augmenting the suggested ads from 1 to 5. Results are then calculated in terms of the precision $p@k$, with $k \in [1,5]$. For each experiment, Table 1 reports the results obtained by varying α. According to equation (2) a value of 0.0 means that only CF are considered, whereas a value of 1.0 considers only BoC in ConCA and BoW in the baseline system.

Results show that ConCA performs slightly better than the baseline system. In particular, for both systems, the best performance is obtained with a low value of α (i.e., 0.1 for ConCA). It means that CF have more impact than BoC in ConCA. Similarly, CF have more impact than BoW in the baseline system. Nevertheless, concepts positively affect results when suggesting 1 or 5 ads. Since these preliminary results are encouraging, we are currently performing experiments that consider BoW, BoC, and CF in conjunction. In this way, we adopt both the syntactic and the semantic contributions.

As a final remark, running times for ConCA and the baseline system are comparable, so that the real-time constraint is preserved.

6 RELATED WORK

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. (Yih et al., 2006) 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, the 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. Since the repository of ads adopted in our work is composed by Web pages of companies, we do not take into account the phrase extraction but resort only on extraction-based text summarization.

Ribeiro-Neto et al. (Ribeiro-Neto et al., 2005) 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

Table 1: Results of CA systems comparison.

	Baseline System					**ConCA**				
α	**p@1**	**p@2**	**p@3**	**p@4**	**p@5**	**p@1**	**p@2**	**p@3**	**p@4**	**p@5**
0.0	0.765	0.746	0.719	0.696	**0.663**	0.765	0.746	0.719	0.696	0.663
0.1	0.767	0.749	0.724	0.698	**0.663**	**0.773**	**0.752**	**0.728**	**0.701**	**0.668**
0.2	**0.768**	**0.750**	**0.729**	0.699	0.662	0.761	0.747	0.724	0.696	0.662
0.3	0.766	0.749	**0.729**	**0.701**	0.661	0.736	0.730	0.709	0.685	0.650
0.4	0.756	0.747	0.728	0.698	0.658	0.701	0.704	0.686	0.668	0.636
0.5	0.744	0.734	0.720	0.692	0.651	0.661	0.664	0.660	0.643	0.614
0.6	0.722	0.717	0.703	0.681	0.640	0.609	0.624	0.623	0.614	0.588
0.7	0.684	0.687	0.680	0.658	0.625	0.561	0.573	0.578	0.568	0.551
0.8	0.632	0.637	0.635	0.614	0.586	0.512	0.518	0.517	0.513	0.501
0.9	0.557	0.552	0.548	0.534	0.512	0.481	0.471	0.462	0.455	0.440
1.0	0.439	0.421	0.408	0.388	0.372	0.427	0.407	0.394	0.379	0.360

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. Nevertheless, since our ads are actually Web pages, we do not take into account the impedance mismatch and we measure the direct match of the page p and the ad a by calculating the corresponding cosine similarity.

Broder et al. (Broder et al., 2007) classified both pages and ads according to 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. This was the first attempt to semantically enrich CA by introducing the contribution given by CF. Results shown a better accuracy than those performed by systems based only on syntactic matching. According to their results and taking into account the effectiveness of adopting CF, ConCA implements the same classifier.

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. Anagnostopoulos et al. (Anagnostopoulos et al., 2007) presented a methodology for Web advertising in real time, focusing on the contributions of different fragments of a Web page. This methodology allows to identify short but informative excerpts of the Web page by using several text summarization techniques, in conjunction with the model developed in (Broder et al., 2007). According to their work, in order to analyze the entire body of Web pages on-the-fly, also ConCA performs a text summarization task.

In a previous work (Armano et al., 2011b), we studied the impact of text summarization on CA and proved that the best performances in terms of precision are obtained by using the TFLP technique (i.e., a simple but effective extraction technique that considers the title, the first and the last paragraph of a Web page). In accordance with these results, the text summarization module implemented in ConCA adopts the TFLP technique.

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. (Murdock et al., 2007) considered a statistical machine translation model to overcome the problem of the impedance mismatch between page 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, Ciaramita et al. (Ciaramita et al., 2008) used a machine learning approach, based on the model described in (Murdock et al., 2007), to define an innovative set of features able to extract the semantic correlations between the page and ad vocabularies. According on this view, we calculate the performance of ConCA resorting to $p@k$, a classical measure adopted in query search, which in this case represents the capability of the system to suggest k relevant ads to a Web page.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a semantic enrichment of Contextual Advertising, taking into account the concepts extracted by ConceptNet 3. To this end, we devised a system called ConCA. The system has been evaluated in terms of precision by performing comparative experiments with a state-of-the-art system. Results shown that the adoption of concepts positively

affects the choice of ads.

As for future work, we are studying the impact of different combinations of features (BoW, BoC, and CF) on the precision of ConCA. Furthermore, we are investigating further semantic solutions aimed at improving ConCA. In particular, we are investigating novel solutions to extract concepts, by adopting WordNet (Miller, 1995) and/or Yago (Suchanek et al., 2007). Moreover, we are planning to modify the classifier adopting the hierarchical text categorization solution proposed in (Addis et al., 2010), instead of the Rocchio classifier. We deem that taking into account the taxonomic relationship among classes would improve the overall performance of the classifier. We are also about to calculate its performances with further datasets, such as a larger dataset extracted by DMOZ.

ACKNOWLEDGEMENTS

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

Addis, A., Armano, G., and Vargiu, E. (2010). Assessing progressive filtering to perform hierarchical text categorization in presence of input imbalance. In *Proceedings of International Conference on Knowledge Discovery and Information Retrieval (KDIR 2010)*.

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. (2011a). Experimenting text summarization techniques for contextual advertising. In *IIR'11: Proceedings of the 2nd Italian Information Retrieval (IIR) Workshop*.

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

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.

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.

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.

Havasi, C., Speer, R., and Alonso, J. (2007). Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent Advances in Natural Language Processing*, Borovets, Bulgaria.

Liu, H. and Singh, P. (2004). Conceptnet: A practical commonsense reasoning tool-kit. *BT Technology Journal*, 22:211–226.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

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.

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.

Sinka, M. and Corne, D. (2002). 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*, pages 881–890. Press.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: A Core of Semantic Knowledge. In *16th international World Wide Web conference (WWW 2007)*, New York, NY, USA. ACM Press.

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.