Matching Ads in a Collaborative Advertising System

Giuliano Armano and Alessandro Giuliani

University of Cagliari
Department of Electrical and Electronic Engineering
{armano,alessandro.giuliani}@diee.unica.it

Abstract. Classical contextual advertising systems suggest suitable ads to a given webpage, without relying on further information – i.e. just analyzing its content. Although we agree that the target webpage is important for selecting ads, in this paper we concentrate on the importance of taking into account also information extracted from the webpages that link the target webpage (inlinks). According to this insight, contextual advertising can be viewed as a collaborative filtering process, in which selecting a suitable ad corresponds to estimate to which extent the ad matches the characteristics of the "current user" (the webpage), together with the characteristics of similar users (the inlinks). We claim that, in so doing, the envisioned collaborative approach is able to improve classical contextual advertising. Experiments have been performed comparing a collaborative system implemented in accordance with the proposed approach against (i) a classical content-based system and (ii) a system that relies only on the content of similar pages (disregarding the target webpage). Experimental results confirm the validity of the approach.

1 Introduction

Web 2.0 users need suggestions about online contents (e.g., news and photos), people (e.g., friends in social networks), goods for sale (e.g., books and CDs), and/or services and products (e.g., suitable ads), depending on their preferences and tastes. In this scenario, Information Filtering (IF) techniques, aimed at presenting only *relevant* information to users, need to be improved to make them more robust, intelligent, effective, and applicable to a broad range of real life applications. To this end, the corresponding research activities are focused on defining and implementing intelligent techniques rooted in several research fields, – including machine learning, text categorization, evolutionary computation, and semantic web [1].

IF is typically performed by using Recommender Systems (RS). Here, recommendations are typically provided by relying on Collaborative Filtering (CF), which consists of automatically making predictions (*filtering*) about the interests of a user by collecting preferences or tastes from similar users (*collaboration*). The underlying idea is that similar users have similar tastes.

C. Huemer and P. Lops (Eds.): EC-Web 2013, LNBIP 152, pp. 150–161, 2013.

Several CF systems have been developed to suggest items and goods [2]. Some proposals suggest to use CF also for Contextual Advertising (CA), i.e., for suggesting suitable ads to a webpage [3,4,5]. In fact, suggesting an ad to a webpage can be viewed as the task of recommending an item (the ad) to a user (the webpage) [6]. In classical CA systems an ad is typically suggested after matching the target webpage with the contents of candidate ads. Although we agree that the target webpage is important for selecting ads, in this paper we concentrate on the importance of taking into account also information extracted from the webpages that link the target webpage (inlinks). According to this insight, CA can be viewed as a collaborative filtering process, in which selecting a suitable ad corresponds to estimate to which extent the ad matches the characteristics of the "current user" (the webpage), together with the characteristics of similar users (the inlinks). We claim that, in so doing, the envisioned collaborative approach is able to improve classical contextual advertising.

Experiments have been performed comparing a collaborative system implemented in accordance with the proposed approach against (i) a classical content-based system and (ii) a system that relies only on the content of similar pages (disregarding the target webpage).

The remainder of the paper is organized as follows: Section 2 illustrates the proposed approach and the implemented system. Section 3 presents the adopted dataset, the evaluation metrics, and the baseline system. In Section 4, results are showed and discussed. Section 5 gives an overview of the related work. Section 6 concludes the paper sketching future work.

2 The Proposed Approach and the Corresponding System

It has been shown that the best performances in RS are obtained by adopting hybrid solutions, which make use of both collaborative and content-based techniques [7]. On the other hand, a preliminary study about the adoption of hybrid techniques (in particular those typically applied for RS) has been proposed in the work of Vargiu et al.[8]. This insight has also been investigated in the work

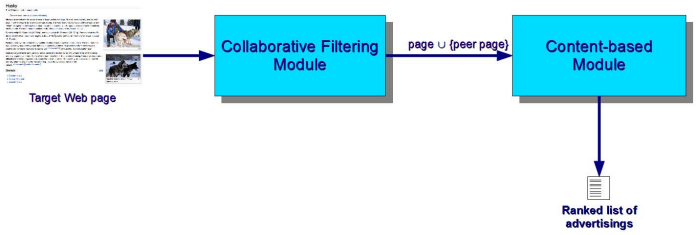

Fig. 1. The interaction between the CF module and the content-based module

of Armano et al. [9], focusing on the feasibility of devising a RS a la mode of CA, and vice versa.

In agreement with this view, we implement a CA system that uses hybrid techniques, taken from RS. We know that RS uses CF by relying on *peer users* (i.e., users with similar tastes). Hence, the proposed approach uses CF by relying on *peer pages* (i.e., at least in principle, pages related to the target webpage). The approach is sketched in Figure 1.

Given a target webpage p in which to display an ad, the proposed approach uses CF to extract information from webpages related to p and then classifies them according to a given taxonomy. In particular, the CF module uses the *collaboration* of p by retrieving a subset of its peer pages. The content of p and of its retrieved peer pages is then analyzed by a content-based module that is in charge of suggesting suitable ads to p. Suitable peer pages appear to be all the *inlinks* of p (also called *backlinks*). The underlying motivation is that most likely an inlink contains information strictly related to the topic of p [10]. However, we used the inlink's snippet[1], instead of taking into account the whole page. It is worth pointing out that the decision of using snippets is a trade-off between two conflicting issues: the need for retaining relevant information and the need for limiting the latency time (for more information on these issues see, for example, [8,11]).

Figure 2 depicts high-level architecture of the proposed system, composed of four modules: (i) *Inlink extractor*, (ii) *BoW builder*, (iii) *Classifier*, and (iv) *Matcher*.

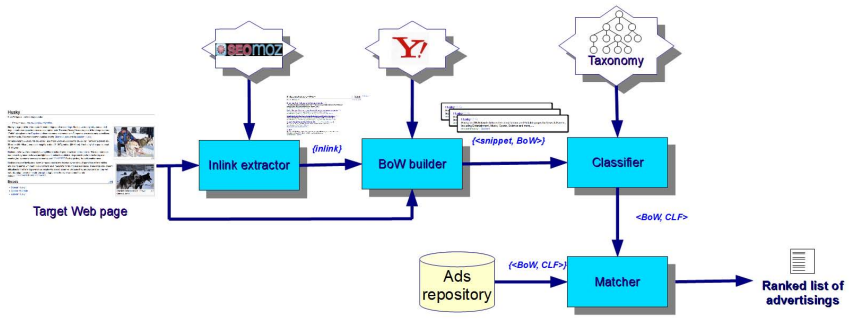

Fig. 2. The high-level architecture of the proposed system

Inlink Extractor. The inlink extractor represents the collaborative component of the system, aimed at finding peer pages[2]. This module collects the first 10 inlinks

[1] A snippet is the page excerpt provided by a search engines following the query of the user.

[2] In principle, also *outlinks* could be taken into account, as in fact, an outlink in p to q is an inlink to q from p. However, in the current version of the system we consider only inlinks, as they appear to be more informative than outlinks. This conjecture has been experimentally investigated in a preliminary study by Armano et al. [11].

of a given page by performing a query to SeoMoz[3], a special online service for Search Engine Optimization tasks.

BoW Builder. This module represents the content-based component of the system, aimed at extracting and analyzing snippets. The Bow builder first extracts the snippet of the target webpage and the snippets of its inlinks by querying Yahoo! (the content of each query is the URL of the link under analysis). This module outputs a vector representation of the original text as bag of words (BoW), each word being weighted by its TFIDF [12]. Let us recall that the effectiveness of snippets for text summarization has been experimentally proven in [13].

Classifier. This module is aimed at categorizing the target webpage and its inlinks according to a relevant taxonomy. The *Classifier* computes the so-called Classification Features (CLF), in accordance with the work of Broder et al. [14]. CLF are weighted and represented as a vector, whose generic feature w_j reports the averaged score given by the classifier to the target webpage (represented by the current BoW) for the category j. The module outputs a $\langle Bow, CLF \rangle$ pair – i.e., the bag of words and the classification features of the target webpage.

The classifier is trained according to the Rocchio algorithm [15]. In particular, for each category of the taxonomy, all relevant snippets are used to evaluate its centroid, with only positive examples and no relevance feedback. In formula:

$$\vec{c_j} = \frac{1}{|C_j|} \sum_{\vec{d} \in C_j} \frac{\vec{d}}{\left\| \vec{d} \right\|} \tag{1}$$

where $\vec{c_j}$ is the centroid for class C_j and d (i.e., the BoW representation of snippet in terms of its TFIDF encoding) ranges over the documents of a particular class. The classification of a snippet is based on the cosine of the angle between the snippet s and the centroid of the class C_j. In formula:

$$C^* = \underset{c_j \in C}{argmax} \left(\frac{\vec{c_j}}{\left\| \vec{c_j} \right\|} \cdot \frac{\vec{s}}{\left\| \vec{s} \right\|} \right) = \underset{c_j \in C}{argmax} \frac{\sum_{i \in F} c_j^i \cdot s^i}{\sqrt{\sum_{i \in F} \left(c_j^i \right)^2} \sqrt{\sum_{i \in F} \left(s^i \right)^2}} \tag{2}$$

where F is the set of features. To produce comparable scores, each score is normalized with the snippet and the class length. The terms c_j^i and s^i represent the weight of the ith feature, based on the standard TFIDF formula, in the jth class centroid and in the s snippet, respectively.

Matcher. This module assigns a score s to each ad according to its similarity with the given target webpage, according to the following formula:

$$s(p, a) = \alpha \cdot sim_{BoW}(p, a) + (1 - \alpha) \cdot sim_{CLF}(p, a) \tag{3}$$

[3] `http://www.seomoz.org`

in which α is a global parameter that permits to control the impact of BoW with respect to CLF, whereas $sim_{BoW}(p, a)$ and $sim_{CLF}(p, a)$ are cosine similarity scores between the target page (p) and the ad (a) using BoW and CLF, respectively. For $\alpha = 0$ only semantic analysis is considered, whereas for $\alpha = 1$ only syntactic analysis is considered.

After ranking categories, the matcher selects the first k categories (k is a fixed parameter that depends on the agreement between publisher and advertiser).

Each ad, which in our work is represented by the description (the so-called *creative*) of a product or service company's webpage (*landing page*), is processed in a similar way and it is represented by suitable BoW and CLF, where the CLF are computed only for the landing page's snippet. To choose the ads relevant to the target page, the :

3 Experiments Set Up

To assess the effectiveness of the proposed approach, we performed comparative experiments with a content-based system that does not implement any collaborative approach and with a system that analyzes only the content of similar pages.

The Adopted Dataset

We first trained the classifier with a subset of DMOZ[4]. Let us recall that DMOZ is the collection of HTML documents referenced in a Web directory developed in the Open Directory Project (ODP). Experiments have been performed on 21 selected categories, arranged in a hierarchy with depth 3.

As for the ads, we built a suitable repository in which they are manually classified according to the given taxonomy. We created the repository focusing on the DMOZ subtree rooted by the "Shopping" category. In this repository each ad is represented by the creative and the title of the landing webpage.

Evaluation Measures

Given a page p and an ad a, the corresponding $\langle p,\ a\rangle$ pair has been scored on a 1 to 3 scale, defined as follows (see Figure 3):

1 **Relevant (Figure 3-a)**. Occurs when a is directly related to the main subject of p. This case holds when both p and a belong to the same class (F).
2 **Somewhat relevant**. Three cases may occur: (i) a is related to a similar subject of p (*sibling*, Figure 3-b.1); (ii) a is related to the main topic of p in a more general way (*generalization*,Figure 3-b.2); or (iii) a is related to the main topic of p in a very specific way (*specialization*, Figure 3-b.3).
3 **Irrelevant (Figure 3-c)**. When the ad is unrelated to the page, i.e., they are in different branches of the taxonomy.

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

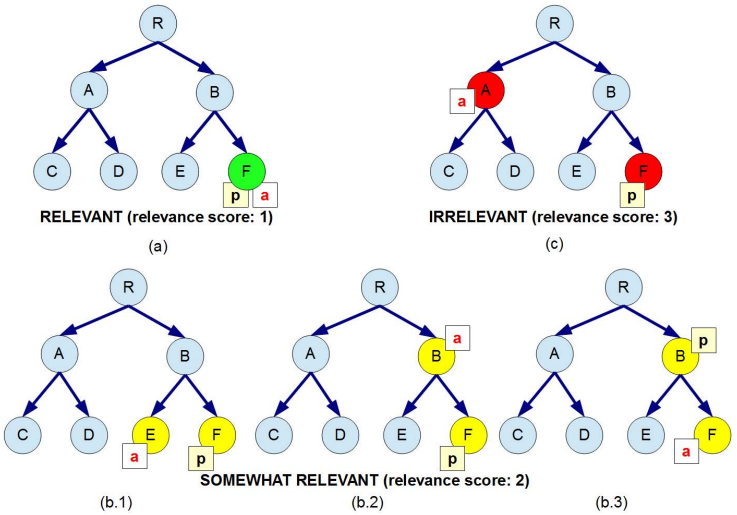

Fig. 3. The adopted policy for calculating relevance scores

According to state-of-the-art (e.g., [14]), we considered as True Positives (TP) pairs scored 1 or 2, and as False Positives (FP) pairs scored 3. In so doing, we are able to calculate the precision p of a system in the classical way.

As we rely on a graded relevance scale of evaluation, to measure the effectiveness of the approach we also made comparisons by relying on two further evaluation metrics: Normalized Discounted Cumulative Gain (nDCG)[16] and Expected Reciprocal Rank (ERR)[17].

Let us note that we do not have any information regarding the Click-through Rate (CTR) and no further comparison measures can be provided (in fact, this information is not given by companies that develop advertising systems, e.g., Yahoo!, Google, or Microsoft).

The Systems Adopted for Comparisons

As we are interested in studying the impact of CF on CA tasks, to perform experiments we devised a content-based system in which the target webpage is classified with the same *Classifier* adopted in the proposed system, but without resorting to any collaborative approach. The corresponding system, depicted in Figure 4, is compliant with the system proposed by Anagnostopoulos et al.[18], in which only classification features are considered in the matching phase (let us remark that creating a snippet is an extraction-based text summarization technique). The system takes the target webpage as input. The *BoW builder*

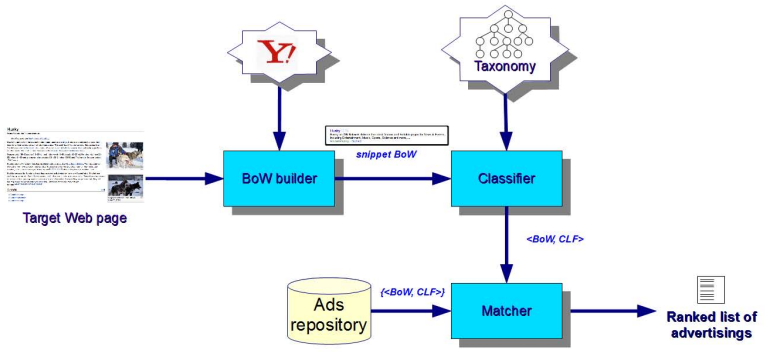

Fig. 4. The baseline system

first retrieves the snippet of the page and then removes stop-words and performs stemming. Starting from the *BoW* provided by the *BoW builder*, the *Classifier* classifies the page's snippet according to the given taxonomy by adopting a centroid-based approach. Finally, the *Matcher* ranks the ads contained in the *Ads repository* according to equation 3.

To study the impact of the peer pages alone (i.e., without taking into account the target webpage), we modified the proposed hybrid system removing the content of the target webpage from the input of the *BoW builder* module. In other words, in the system depicted in Figure 2 the target webpage is not given as input to the *BoW builder*, meaning that only inlinks have been processed.

4 Results

As pointed out, all systems embed the same *Classifier*, which has been trained by using the same training set. A total of about 2100 webpages, each belonging to one category, has been adopted to train and test the systems and to make experimental comparisons. Experiments have been performed by running 10-fold cross-validation.

We evaluated the performance of each system by running five different experiments, in which from 1 to 5 ads are selected for the target page, respectively. In Figure 5, each chart reports the precision of systems while varying α. According to equation 3, a value of 0.0 means that only semantic analysis is considered, whereas a value of 1.0 considers only syntactic analysis. The three systems are, respectively, our CF proposal (***Page + Inlinks***), the proposal in which only peer pages are taken into account (***Inlinks***), and the content-based baseline system (***Page***). As expected, the best performance in each chart is provided by the adoption of page and inlinks snippets. Furthermore, in each chart, the peak of precision is obtained with low values of α (ranging from 0 to 0.5). Observing the charts, we can state that, in this kind of problem, the CLF have the discriminant

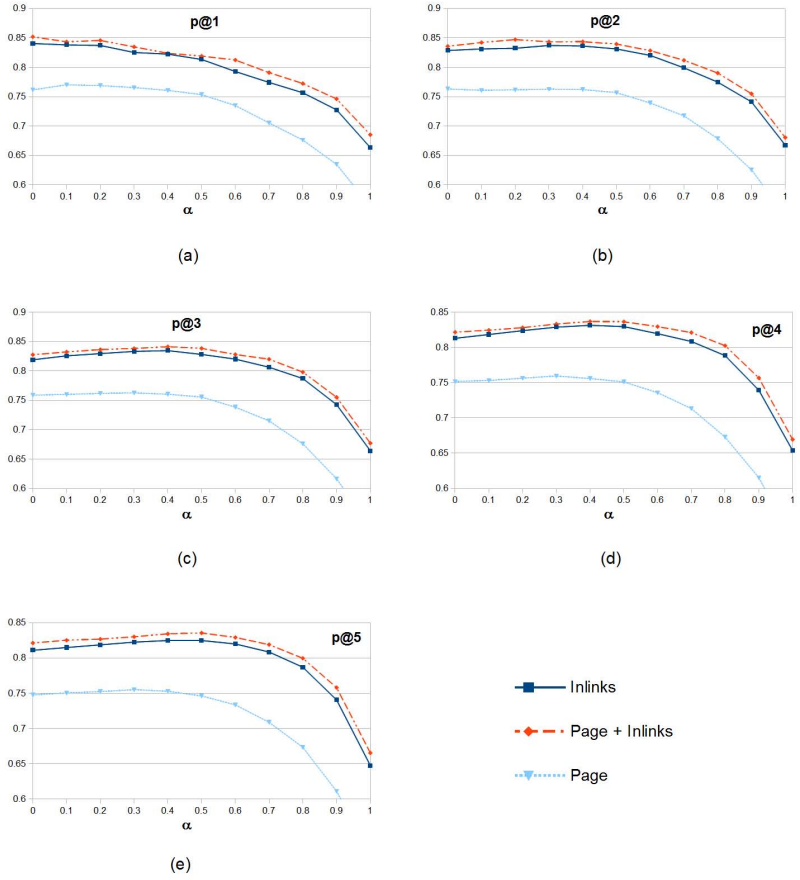

Fig. 5. The results in terms of p@k while varying α

impact at low values of k, whereas the *BoW* could help to improve the performance for higher value of k. This behavior is in accordance with previous works, meaning that the semantic information has more impact than the syntactic one.

For the sake of clarity we report in Figure 6 the precision of the systems, for each value k, according to the best value of α. Figure 6 highlights that the adoption of inlinks improves the performance of the baseline system.

Finally, Table 1 shows the performance of each system in suggesting 5 ads, in terms of precision, nDCG and ERR, confirming the assumption that linking documents can be helpful.

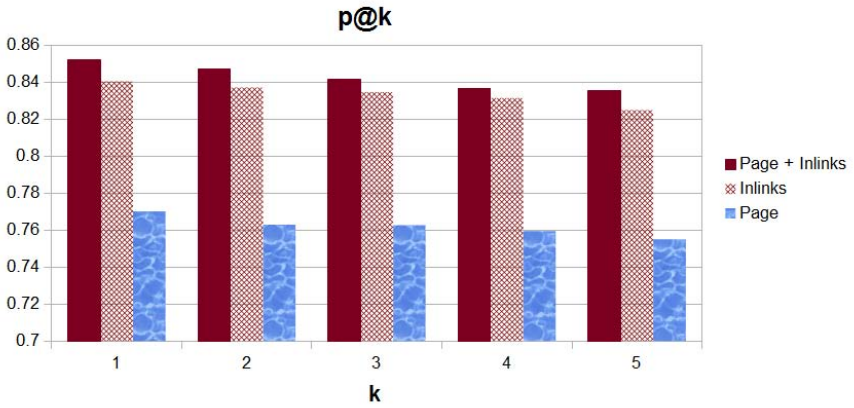

Fig. 6. The results in terms of p@k

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

	p@5	**nDCG@5**	**ERR@5**
Page	0.755	0.540	0.539
Inlinks	0.825	0.605	0.593
Page + Inlinks	0.836	0.617	0.601

5 Related Work

Based on the observation that relevant ads have higher probabilities of being clicked by users than generic or irrelevant ads, much research work has attempted to improve the relevance of retrieved ads. Several works focused on the extraction of meaningful keywords [19] [20] [21]. Broder et al. [14] classified both pages and ads according to a given taxonomy and matched ads to the page falling into the same node of the taxonomy, giving rise to a semantic analysis. Another approach that combines syntax and semantics has been proposed in [22]. The corresponding system, called ConCA (Concepts on Contextual Advertising), relies on ConceptNet, a semantic network able to provide commonsense knowledge [23]. The choice of the classifier is in accordance with these works. Nowadays, ad networks need to deal in real time with a large amount of data, at least in principle, involving billions of pages and ads. Hence, efficiency and computational costs are crucial factors in the choice of methods and algorithms. Anagnostopoulos et al. [18] presented a methodology for Web advertising in real time, focusing on the contributions of the different fragments of a webpage. This methodology allows to identify short but informative excerpts of the webpage by means of several text summarization techniques, used in conjunction with the model developed in [14]. According to this view, Armano et al. [24] [25] studied the impact of text summarization in CA, showing that effective text summarization techniques may help to improve the behavior of a CA system. The adoption of snippets is compliant with these works; in fact, a snippet is usually

a summary of the webpage's content. In RS, different ways of combining collaborative and content-based methods have been adopted [2]: (i) implementing collaborative and content-based methods separately and combining their predictions; (ii) embedding some content-based characteristics into a collaborative approach; (iii) embedding some collaborative characteristics into a content-based approach; and (iv) devising a unifying model able to incorporate both content-based and collaborative characteristics. As the best performances in RS are achieved by adopting CF in conjunction with content-based approaches [7], we propose the hybrid CA system that uses CF in a content-based setting according to the third approach. Many researchers investigated the role of links in information retrieval. In particular, links have been used to (i) enhance document representation [26], (ii) improve document ranking by propagating document score [27], (iii) provide an indicator of popularity [28], and (iv) find hubs and authorities for a given topic [29]. Our choice to rely on inlinks is consistent with link-based ranking algorithms, which are based on the assumption that linking documents have related content [10].

6 Conclusions and Future Work

It is well known that the best performances in recommender systems are obtained by adopting collaborative filtering, in conjunction with content-based approaches. In this paper, we proposed a collaborative advertising system which makes use oh hybrid techniques, in particular, collaborative filtering. To suggest an ad to a target webpage, we perform a direct matching between the webpage and each ad. We adopt the collaboration of suitable pages related thereto, i.e., pages similar to the target webpage, at least in the topics. We performed comparative experiments with a content-based and with a system that takes into account only the peer pages (while disregarding the target one). Experiments have been performed on about 2100 webpages from the Open Directory Project. Results indicate the effectiveness of the proposed approach and show that the proposed hybrid contextual advertising system performs better than the baseline system.

As for future work, we are currently studying how to improve the collaborative and/or the content-based module. As for the former, further techniques for selecting similar pages are under study, for instance link prediction methods [30] and the adoption of clustering techniques. As for the latter, we are planning to modify the classifier by adopting a hierarchical approach, such as the one proposed in [31]. In fact, in our view, taking into account the taxonomic relationship among categories should improve overall classifier performance.

References

1. Armano, G., de Gemmis, M., Semeraro, G., Vargiu, E.: Intelligent Information Access. SCI, vol. 301. Springer, Heidelberg (2010)
2. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17(6), 734–749 (2005)

3. Broder, A.Z., Ciccolo, P., Fontoura, M., Gabrilovich, E., Josifovski, V., Riedel, L.: Search advertising using web relevance feedback. In: Proc. of 17th. Int. Conference on Information and Knowledge Management, pp. 1013–1022 (2008)
4. Anastasakos, T., Hillard, D., Kshetramade, S., Raghavan, H.: A collaborative ltering approach to ad recommendation using the query-ad click graph. In: Proc. of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, pp. 1927–1930. ACM, New York (2009)
5. Vargiu, E., Urru, M.: Exploiting web scraping in a collaborative ltering-based approach to web advertising. Artificial Intelligence Research 2(1), 44–54 (2013)
6. Armano, G., Vargiu, E.: A unifying view of contextual advertising and recommender systems. In: Proc. of Int. Conference on Knowledge Discovery and Information Retrieval (KDIR 2010), pp. 463–466 (2010)
7. Burke, R.: Hybrid recommender systems: Survey and experiments. User Modeling and User-Adapted Interaction 12(4), 331–370 (2002)
8. Vargiu, E., Giuliani, A., Armano, G.: Improving contextual advertising by adopting collaborative filtering. ACM Transaction on the Web (in press, 2013)
9. Armano, G., Giuliani, A., Vargiu, E.: Intelligent Techniques in Recommender Systems and Contextual Advertising: Novel Approaches and Case Studies. In: Intelligent Techniques in Recommendation Systems: Contextual Advancements and New Methods, pp. 105–128. IGI Global (2012)
10. Koolen, M., Kamps, J.: Are semantically related links more effective for retrieval? In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 92–103. Springer, Heidelberg (2011)
11. Armano, G., Giuliani, A., Vargiu, E.: Are related links effective for contextual advertising? a preliminary study. In: Int. Conference on Knowledge Discovery and Information Retrieval (2012)
12. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill Book Company (1984)
13. Armano, G., Giuliani, A., Vargiu, E.: Using snippets in text summarization: A comparative study and an application. In: IIR 2012: 3rd Italian Information Retrieval (IIR) Workshop (2012)
14. Broder, A., Fontoura, M., Josifovski, V., Riedel, L.: A semantic approach to contextual advertising. In: SIGIR 2007: Proc. of the 30th annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 559–566. ACM, New York (2007)
15. Rocchio, J.: Relevance feedback in information retrieval. In: The SMART Retrieval System: Experiments in Automatic Document Processing, pp. 313–323. Prentice-Hall (1971)
16. Järvelin, K., Kekäläinen, J.: IR evaluation methods for retrieving highly relevant documents. In: Proc. of the 23rd Int. ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2000, pp. 41–48. ACM, New York (2000)
17. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: Proc. of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, pp. 621–630. ACM, New York (2009)
18. Anagnostopoulos, A., Broder, A.Z., Gabrilovich, E., Josifovski, V., Riedel, L.: Just-in-time contextual advertising. In: CIKM 2007: Proc. of the Sixteenth ACM Conference on Information and Knowledge Management, pp. 331–340. ACM, New York (2007)

19. Ribeiro-Neto, B., Cristo, M., Golgher, P.B., Silva de Moura, E.: Impedance coupling in content-targeted advertising. In: SIGIR 2005: Proc. of the 28th Int. ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 496–503. ACM, New York (2005)
20. Lacerda, A., Cristo, M., Gonçalves, M.A., Fan, W., Ziviani, N., Ribeiro-Neto, B.: Learning to advertise. In: SIGIR 2006: Proc. of the 29th Int. ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 549–556. ACM, New York (2006)
21. Yih, W.T., Goodman, J., Carvalho, V.R.: Finding advertising keywords on web pages. In: WWW 2006: Proc. of the 15th Int. Conference on World Wide Web, pp. 213–222. ACM, New York (2006)
22. Armano, G., Giuliani, A., Vargiu, E.: Semantic enrichment of contextual advertising by using concepts. In: Int. Conference on Knowledge Discovery and Information Retrieval (2011)
23. Liu, H., Singh, P.: Conceptnet: A practical commonsense reasoning tool-kit. BT Technology Journal 22, 211–226 (2004)
24. Armano, G., Giuliani, A., Vargiu, E.: Experimenting text summarization techniques for contextual advertising. In: IIR 2011: Proc. of the 2nd Italian Information Retrieval (IIR) Workshop (2011)
25. Armano, G., Giuliani, A., Vargiu, E.: Studying the impact of text summarization on contextual advertising. In: 8th Int. Workshop on Text-based Information Retrieval (2011)
26. Picard, J., Savoy, J.: Enhancing retrieval with hyperlinks: a general model based on propositional argumentation systems. Journal of the American Society for Information Science and Technology 54, 347–355 (2003)
27. Frei, H.P., Stieger, D.: The use of semantic links in hypertext information retrieval. Information Processing and Management 31, 1–13 (1995)
28. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. 30, 107–117 (1998)
29. Chakrabarti, S., van den Berg, M., Dom, B.: Focused crawling: a new approach to topic-specific Web resource discovery. Computer Networks 31(11-16), 1623–1640 (1999)
30. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology 58, 1019–1031 (2007)
31. Addis, A., Armano, G., Vargiu, E.: Assessing progressive filtering to perform hierarchical text categorization in presence of input imbalance. In: Proc. of Int. Conference on Knowledge Discovery and Information Retrieval, KDIR 2010 (2010)