Impedance Coupling in Content-targeted Advertising

Berthier Ribeiro-Neto
Computer Science Department
Federal University of Minas Gerais
Belo Horizonte, Brazil
berthier@dcc.ufmg.br

Marco Cristo
Computer Science Department
Federal University of Minas Gerais
Belo Horizonte, Brazil
marco@dcc.ufmg.br

Paulo B. Golgher
Akwan Information Technologies
Av. Abraão Caram 430 - Pampulha
Belo Horizonte, Brazil
golgher@akwan.com.br

Edleno Silva de Moura
Computer Science Department
Federal University of Amazonas
Manaus, Brazil
edleno@dcc.ufam.edu.br

ABSTRACT

The current boom of the Web is associated with the revenues originated from on-line advertising. While search-based advertising is dominant, the association of ads with a Web page (during user navigation) is becoming increasingly important. In this work, we study the problem of associating ads with a Web page, referred to as *content-targeted advertising*, from a computer science perspective. We assume that we have access to the text of the Web page, the keywords declared by an advertiser, and a text associated with the advertiser's business. Using no other information and operating in fully automatic fashion, we propose ten strategies for solving the problem and evaluate their effectiveness. Our methods indicate that a matching strategy that takes into account the semantics of the problem (referred to as `AAK` for "ads and keywords") can yield gains in average precision figures of 60% compared to a trivial vector-based strategy. Further, a more sophisticated *impedance coupling strategy*, which expands the text of the Web page to reduce *vocabulary impedance* with regard to an advertisement, can yield extra gains in average precision of 50%. These are first results. They suggest that great accuracy in content-targeted advertising can be attained with appropriate algorithms.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.5.3 [**Pattern Recognition**]: Applications—*Text processing*

General Terms

Algorithms, Experimentation

SIGIR'05, August 15–19, 2005, Salvador, Brazil.

Keywords

Advertising, Web, Bayesian networks, kNN

1. INTRODUCTION

The emergence of the Internet has opened up new marketing opportunities. In fact, a company has now the possibility of showing its advertisements (ads) to millions of people at a low cost. During the 90's, many companies invested heavily on advertising in the Internet with apparently no concerns about their investment return [16]. This situation radically changed in the following decade when the failure of many Web companies led to a dropping in supply of cheap venture capital and a considerable reduction in on-line advertising investments [15, 16].

It was clear then that more effective strategies for on-line advertising were required. For that, it was necessary to take into account short-term and long-term interests of the users related to their information needs [9, 14]. As a consequence, many companies intensified the adoption of intrusive techniques for gathering information of users mostly without their consent [8]. This raised privacy issues which stimulated the research for less invasive measures [16].

More recently, Internet information gatekeepers as, for example, search engines, recommender systems, and comparison shopping services, have employed what is called paid placement strategies [3]. In such methods, an advertiser company is given prominent positioning in advertisement lists in return for a placement fee. Amongst these methods, the most popular one is a non-intrusive technique called *keyword targeted marketing* [16]. In this technique, keywords extracted from the user's search query are matched against keywords associated with ads provided by advertisers. A ranking of the ads, which also takes into consideration the amount that each advertiser is willing to pay, is computed. The top ranked ads are displayed in the search result page together with the answers for the user query.

The success of keyword targeted marketing has motivated information gatekeepers to offer their advertisement services in different contexts. For example, as shown in Figure 1, relevant ads could be shown to users directly in the pages of information portals. The motivation is to take advantage of

the users immediate information interests at browsing time. The problem of matching ads to a Web page that is browsed, which we also refer to as *content-targeted advertising* [1], is different from that of keyword marketing. In this case, instead of dealing with users' keywords, we have to use the contents of a Web page to decide which ads to display.

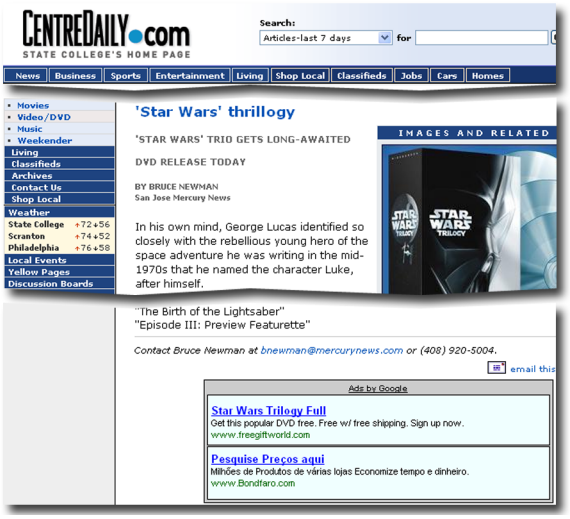

Figure 1: Example of content-based advertising in the page of a newspaper. The middle slice of the page shows the beginning of an article about the launch of a DVD movie. At the bottom slice, we can see advertisements picked for this page by Google's content-based advertising system, AdSense.

It is important to notice that paid placement advertising strategies imply some risks to information gatekeepers. For instance, there is the possibility of a negative impact on their credibility which, at long term, can demise their market share [3]. This makes investments in the quality of ad recommendation systems even more important to minimize the possibility of exhibiting ads unrelated to the user's interests. By investing in their ad systems, information gatekeepers are investing in the maintenance of their credibility and in the reinforcement of a positive user attitude towards the advertisers and their ads [14]. Further, that can translate into higher clickthrough rates that lead to an increase in revenues for information gatekeepers and advertisers, with gains to all parts [3].

In this work, we focus on the problem of content-targeted advertising. We propose new strategies for associating ads with a Web page. Five of these strategies are referred to as *matching strategies*. They are based on the idea of matching the text of the Web page directly to the text of the ads and its associated keywords. Five other strategies, which we here introduce, are referred to as *impedance coupling strategies*. They are based on the idea of expanding the Web page with new terms to facilitate the task of matching ads and Web pages. This is motivated by the observation that there is frequently a mismatch between the vocabulary of a Web page and the vocabulary of an advertisement. We say that there is a *vocabulary impedance problem* and that our technique provides a positive effect of *impedance coupling* by reducing the vocabulary impedance. Further, all our strategies rely on information that is already available to information gatekeepers that operate keyword targeted advertising systems. Thus, no other data from the advertiser is required.

Using a sample of a real case database with over 93,000 ads and 100 Web pages selected for testing, we evaluate our ad recommendation strategies. First, we evaluate the five matching strategies. They match ads to a Web page using a standard vector model and provide what we may call trivial solutions. Our results indicate that a strategy that matches the ad plus its keywords to a Web page, requiring the keywords to appear in the Web page, provides improvements in average precision figures of roughly 60% relative to a strategy that simply matches the ads to the Web page. Such strategy, which we call `AAK` (for "ads and keywords"), is then taken as our baseline.

Following we evaluate the five impedance coupling strategies. They are based on the idea of expanding the ad and the Web page with new terms to reduce the vocabulary impedance between their texts. Our results indicate that it is possible to generate extra improvements in average precision figures of roughly 50% relative to the `AAK` strategy.

The paper is organized as follows. In section 2, we introduce five matching strategies to solve content-targeted advertising. In section 3, we present our impedance coupling strategies. In section 4, we describe our experimental methodology and datasets and discuss our results. In section 5 we discuss related work. In section 6 we present our conclusions.

2. MATCHING STRATEGIES

Keyword advertising relies on matching search queries to ads and its associated keywords. Context-based advertising, which we address here, relies on matching ads and its associated keywords to the text of a Web page.

Given a certain Web page p, which we call *triggering* page, our task is to select advertisements related to the contents of p. Without loss of generality, we consider that an advertisement a_i is composed of a title, a textual description, and a hyperlink. To illustrate, for the first ad by Google shown in Figure 1, the title is "Star Wars Trilogy Full", the description is "Get this popular DVD free. Free w/ free shopping. Sign up now", and the hyperlink points to the site "www.freegiftworld.com". Advertisements can be grouped by advertisers in groups called *campaigns*, such that a campaign can have one or more advertisements.

Given our triggering page p and a set $\mathcal{A}$ of ads, a simple way of ranking $a_i \in \mathcal{A}$ with regard to p is by matching the contents of p to the contents of a_i. For this, we use the vector space model [2], as discussed in the immediately following.

In the vector space model, queries and documents are represented as weighted vectors in an n-dimensional space. Let w_{iq} be the weight associated with term t_i in the query q and w_{ij} be the weight associated with term t_i in the document d_j. Then, $\vec{q} = (w_{1q}, w_{2q}, ..., w_{iq}, ..., w_{nq})$ and $\vec{d_j} = (w_{1j}, w_{2j}, ..., w_{ij}, ..., w_{nj})$ are the weighted vectors used to represent the query q and the document d_j. These weights can be computed using classic *tf-idf* schemes. In such schemes, weights are taken as the product between factors that quantify the importance of a term in a document (given by the term frequency, or *tf*, factor) and its rarity in the whole collection (given by the inverse document factor, or *idf*, factor), see [2] for details. The ranking of the query q with regard to the document d_j is computed by the cosine similarity

formula, that is, the cosine of the angle between the two corresponding vectors:

$$sim(q, d_j) = \frac{\vec{q} \bullet \vec{d_j}}{|\vec{q}| \times |\vec{d_j}|} = \frac{\sum_{i=1}^{n} w_{iq} \cdot w_{ij}}{\sqrt{\sum_{i=1}^{n} w_{iq}^2} \sqrt{\sum_{i=1}^{n} w_{ij}^2}} \quad (1)$$

By considering p as the query and a_i as the document, we can rank the ads with regard to the Web page p. This is our first matching strategy. It is represented by the function AD given by:

$$\texttt{AD}(p, a_i) = sim(p, a_i)$$

where AD stands for "direct match of the ad, composed by title and description" and $sim(p, a_i)$ is computed according to Eq. (1).

In our second method, we use other source of evidence provided by the advertisers: the keywords. With each advertisement a_i an advertiser associates a keyword k_i, which may be composed of one or more terms. We denote the association between an advertisement a_i and a keyword k_i as the pair $(a_i, k_i) \in \mathcal{K}$, where $\mathcal{K}$ is the set of associations made by the advertisers. In the case of keyword targeted advertising, such keywords are used to match the ads to the user queries. In here, we use them to match ads to the Web page p. This provides our second method for ad matching given by:

$$\texttt{KW}(p, a_i) = sim(p, k_i)$$

where $(a_i, k_i) \in \mathcal{K}$ and KW stands for "match the ad keywords".

We notice that most of the keywords selected by advertisers are also present in the ads associated with those keywords. For instance, in our advertisement test collection, this is true for 90% of the ads. Thus, instead of using the keywords as matching devices, we can use them to emphasize the main concepts in an ad, in an attempt to improve our AD strategy. This leads to our third method of ad matching given by:

$$\texttt{AD\_KW}(p, a_i) = sim(p, a_i \cup k_i)$$

where $(a_i, k_i) \in \mathcal{K}$ and AD_KW stands for "match the ad and its keywords".

Finally, it is important to notice that the keyword k_i associated with a_i could not appear at all in the triggering page p, even when a_i is highly ranked. However, if we assume that k_i summarizes the main topic of a_i according to an advertiser viewpoint, it can be interesting to assure its presence in p. This reasoning suggests that requiring the occurrence of the keyword k_i in the triggering page p as a condition to associate a_i with p might lead to improved results. This leads to two extra matching strategies as follows:

$$\texttt{ANDKW}(p, a_i) = \begin{cases} sim(p, a_i) & \text{if } k_i \subseteqq p \\ 0 & \text{if otherwise} \end{cases}$$

$$\texttt{AD\_ANDKW}(p, a_i) = \texttt{AAK}(p, a_i) = \begin{cases} sim(p, a_i \cup k_i) & \text{if } k_i \subseteqq p \\ 0 & \text{if otherwise} \end{cases}$$

where $(a_i, k_i) \in \mathcal{K}$, ANDKW stands for "match the ad keywords and force their appearance", and AD_ANDKW (or AAK for "ads and keywords") stands for "match the ad, its keywords, and force their appearance".

As we will see in our results, the best among these simple methods is AAK. Thus, it will be used as baseline for our impedance coupling strategies which we now discuss.

3. IMPEDANCE COUPLING STRATEGIES

Two key issues become clear as one plays with the content-targeted advertising problem. First, the triggering page normally belongs to a broader contextual scope than that of the advertisements. Second, the association between a good advertisement and the triggering page might depend on a topic that is not mentioned explicitly in the triggering page.

The first issue is due to the fact that Web pages can be about any subject and that advertisements are concise in nature. That is, ads tend to be more topic restricted than Web pages. The second issue is related to the fact that, as we later discuss, most advertisers place a small number of advertisements. As a result, we have few terms describing their interest areas. Consequently, these terms tend to be of a more general nature. For instance, a car shop probably would prefer to use "car" instead of "super sport" to describe its core business topic. As a consequence, many specific terms that appear in the triggering page find no match in the advertisements. To make matters worst, a page might refer to an entity or subject of the world through a label that is distinct from the label selected by an advertiser to refer to the same entity.

A consequence of these two issues is that vocabularies of pages and ads have low intersection, even when an ad is related to a page. We cite this problem from now on as the *vocabulary impedance problem*. In our experiments, we realized that this problem limits the final quality of direct matching strategies. Therefore, we studied alternatives to reduce the referred vocabulary impedance.

For this, we propose to expand the triggering pages with new terms. Figure 2 illustrates our intuition. We already know that the addition of keywords (selected by the advertiser) to the ads leads to improved results. We say that a keyword reduces the vocabulary impedance by providing an alternative matching path. Our idea is to add new terms (words) to the Web page p to also reduce the vocabulary impedance by providing a second alternative matching path. We refer to our expansion technique as impedance coupling. For this, we proceed as follows.

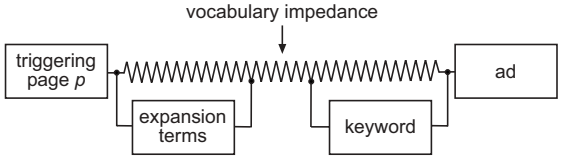

Figure 2: Addition of new terms to a Web page to reduce the vocabulary impedance.

An advertiser trying to describe a certain topic in a concise way probably will choose general terms to characterize that topic. To facilitate the matching between this ad and our triggering page p, we need to associate new general terms with p. For this, we assume that Web documents similar to the triggering page p share common topics. Therefore,

by inspecting the vocabulary of these similar documents we might find good terms for better characterizing the main topics in the page p. We now describe this idea using a Bayesian network model [10, 11, 13] depicted in Figure 3.

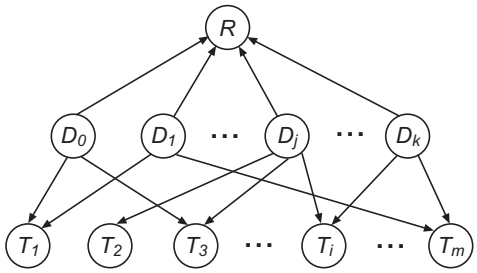

Figure 3: Bayesian network model for our impedance coupling technique.

In our model, which is based on the belief network in [11], the nodes represent pieces of information in the domain. With each node is associated a binary random variable, which takes the value 1 to mean that the corresponding entity (a page or terms) is *observed* and, thus, relevant in our computations. In this case, we say that the information was *observed*. Node R represents the page r, a new representation for the triggering page p. Let $\mathcal{N}$ be the set of the k most similar documents to the triggering page, including the triggering page p itself, in a large enough Web collection $\mathcal{C}$. Root nodes D_0 through D_k represent the documents in $\mathcal{N}$, that is, the triggering page D_0 and its k nearest neighbors, D_1 through D_k, among all pages in $\mathcal{C}$. There is an edge from node D_j to node R if document d_j is in $\mathcal{N}$. Nodes T_1 through T_m represent the terms in the vocabulary of $\mathcal{C}$. There is an edge from node D_j to a node T_i if term t_i occurs in document d_j. In our model, the observation of the pages in $\mathcal{N}$ leads to the observation of a new representation of the triggering page p and to a set of terms describing the main topics associated with p and its neighbors.

Given these definitions, we can now use the network to determine the probability that a term t_i is a good term for representing a topic of the triggering page p. In other words, we are interested in the probability of observing the final evidence regarding a term t_i, given that the new representation of the page p has been observed, $P(T_i = 1|R = 1)$. This translates into the following equation[1]:

$$P(T_i|R) = \frac{1}{P(R)} \sum_{\mathbf{d}} P(T_i|\mathbf{d}) P(R|\mathbf{d}) P(\mathbf{d}) \qquad (2)$$

where $\mathbf{d}$ represents the set of states of the document nodes. Since we are interested just in the states in which *only* a single document d_j is observed and $P(\mathbf{d})$ can be regarded as a constant, we can rewrite Eq. (2) as:

$$P(T_i|R) = \frac{\nu}{P(R)} \sum_{j=0}^{k} P(T_i|\mathbf{d_j}) P(R|\mathbf{d_j}) \qquad (3)$$

where $\mathbf{d_j}$ represents the state of the document nodes in which *only* document d_j is observed and ν is a constant associated with $P(\mathbf{d_j})$. Eq. (3) is the general equation to compute the probability that a term t_i is related to the triggering page. We now define the probabilities $P(T_i|\mathbf{d_j})$ and $P(R|\mathbf{d_j})$ as follows:

$$P(T_i|\mathbf{d_j}) = \eta \; w_{ij} \qquad (4)$$

$$P(R|\mathbf{d_j}) \;\; = \;\; \begin{cases} (1-\alpha) & j = 0 \\ \alpha \; sim(r, d_j) & 1 \leq j \leq k \end{cases} \qquad (5)$$

where η is a normalizing constant, w_{ij} is the weight associated with term t_i in the document d_j, and $sim(p, d_j)$ is given by Eq. (1), i.e., is the cosine similarity between p and d_j. The weight w_{ij} is computed using a classic tf-idf scheme and is zero if term t_i does not occur in document d_j. Notice that $P(\overline{T_i}|\mathbf{d_j}) = 1 - P(T_i|\mathbf{d_j})$ and $P(\overline{R}|\mathbf{d_j}) = 1 - P(R|\mathbf{d_j})$. By defining the constant α, it is possible to determine how important should be the influence of the triggering page p to its new representation r. By substituting Eq. (4) and Eq. (5) into Eq. (3), we obtain:

$$P(T_i|R) = \rho \; ((1-\alpha) \; w_{i0} + \alpha \sum_{j=1}^{k} w_{ij} \; sim(r, d_j)) \qquad (6)$$

where $\rho = \eta \; \nu$ is a normalizing constant.

We use Eq. (6) to determine the set of terms that will compose r, as illustrated in Figure 2. Let t_{top} be the top ranked term according to Eq. (6). The set r is composed of the terms t_i such that $\frac{P(T_i|R)}{P(T_{top}|R)} \geq \beta$, where β is a given threshold. In our experiments, we have used $\beta = 0.05$. Notice that the set r might contain terms that already occur in p. That is, while we will refer to the set r as expansion terms, it should be clear that $p \cap r \neq \varnothing$.

By using $\alpha = 0$, we simply consider the terms originally in page p. By increasing α, we relax the context of the page p, adding terms from neighbor pages, turning page p into its new representation r. This is important because, sometimes, a topic apparently not important in the triggering page offers a good opportunity for advertising. For example, consider a triggering page that describes a congress in London about digital photography. Although London is probably not an important topic in this page, advertisements about hotels in London would be appropriate. Thus, adding "hotels" to page p is important. This suggests using $\alpha > 0$, that is, preserving the contents of p and using the terms in r to expand p.

In this paper, we examine both approaches. Thus, in our sixth method we match r, the set of new expansion terms, directly to the ads, as follows:

$$\texttt{AAK\_T}(p, a_i) = \texttt{AAK}(r, a_i)$$

where `AAK_T` stands for "match the ad and keywords to the set r of expansion terms".

In our seventh method, we match an expanded page p to the ads as follows:

$$\texttt{AAK\_EXP}(p, a_i) = \texttt{AAK}(p \cup r, a_i)$$

where `AAK_EXP` stands for "match the ad and keywords to the expanded triggering page".

[1]To simplify our notation we represent the probabilities $P(X = 1)$ as $P(X)$ and $P(X = 0)$ as $P(\overline{X})$.

To improve our ad placement methods, other external source that we can use is the content of the page h pointed to by the advertisement's hyperlink, that is, its *landing* page. After all, this page comprises the real target of the ad and perhaps could present a more detailed description of the product or service being advertised. Given that the advertisement a_i points to the landing page h_i, we denote this association as the pair $(a_i, h_i) \in \mathcal{H}$, where $\mathcal{H}$ is the set of associations between the ads and the pages they point to. Our eighth method consists of matching the triggering page p to the landing pages pointed to by the advertisements, as follows:

$$\texttt{H}(p, a_i) = sim(p, h_i)$$

where $(a_i, h_i) \in \mathcal{H}$ and `H` stands for "match the hyperlink pointed to by the ad".

We can also combine this information with the more promising methods previously described, `AAK` and `AAK_EXP` as follows. Given that $(a_i, h_i) \in \mathcal{H}$ and $(a_i, k_i) \in \mathcal{K}$, we have our last two methods:

$$\texttt{AAK\_H}(p, a_i) = \begin{cases} sim(p, a_i \cup h_i \cup k_i) & \text{if } k_i \subseteqq p \\ 0 & \text{if otherwise} \end{cases}$$

$$\texttt{AAK\_EXP\_H}(p, a_i) = \begin{cases} sim(p \cup r, a_i \cup h_i \cup k_i) & \text{if } k_i \subseteqq (p \cup r) \\ 0 & \text{if otherwise} \end{cases}$$

where `AAK_H` stands for "match ads and keywords also considering the page pointed by the ad" and `AAH_EXP_H` stands for "match ads and keywords with expanded triggering page, also considering the page pointed by the ad".

Notice that other combinations were not considered in this study due to space restrictions. These other combinations led to poor results in our experimentation and for this reason were discarded.

4. EXPERIMENTS

4.1 Methodology

To evaluate our ad placement strategies, we performed a series of experiments using a sample of a real case ad collection with 93,972 advertisements, 1,744 advertisers, and 68,238 keywords[2]. The advertisements are grouped in 2,029 campaigns with an average of 1.16 campaigns per advertiser.

For the strategies `AAK_T` and `AAK_EXP`, we had to generate a set of expansion terms. For that, we used a database of Web pages crawled by the TodoBR search engine [12] (`http://www.todobr.com.br/`). This database is composed of 5,939,061 pages of the Brazilian Web, under the domain ".br". For the strategies `H`, `AAK_H`, and `AAK_EXP_H`, we also crawled the pages pointed to by the advertisers. No other filtering method was applied to these pages besides the removal of HTML tags.

Since we are initially interested in the placement of advertisements in the pages of information portals, our test collection was composed of 100 pages extracted from a Brazilian newspaper. These are our triggering pages. They were crawled in such a way that only the contents of their articles was preserved. As we have no preferences for particular topics, the crawled pages cover topics as diverse as politics, economy, sports, and culture.

For each of our 100 triggering pages, we selected the top three ranked ads provided by each of our 10 ad placement strategies. Thus, for each triggering page we select no more than 30 ads. These top ads were then inserted in a pool for that triggering page. Each pool contained an average of 15.81 advertisements. All advertisements in each pool were submitted to a manual evaluation by a group of 15 users. The average number of relevant advertisements per page pool was 5.15. Notice that we adopted the same pooling method used to evaluate the TREC Web-based collection [6].

To quantify the precision of our results, we used 11-point average figures [2]. Since we are not able to evaluate the entire ad collection, recall values are relative to the set of evaluated advertisements.

4.2 Tuning Idf factors

We start by analyzing the impact of different idf factors in our advertisement collection. Idf factors are important because they quantify how discriminative is a term in the collection. In our ad collection, idf factors can be computed by taking ads, advertisers or campaigns as documents. To exemplify, consider the computation of "ad idf" for a term t_i that occurs 9 times in a collection of 100 ads. Then, the inverse document frequency of t_i is given by:

$$idf_i = log \frac{100}{9}$$

Hence, we can compute ad, advertiser or campaign idf factors. As we observe in Figure 4, for the `AD` strategy, the best ranking is obtained by the use of campaign idf, that is, by calculating our idf factor so that it discriminates campaigns. Similar results were obtained for all the other methods.

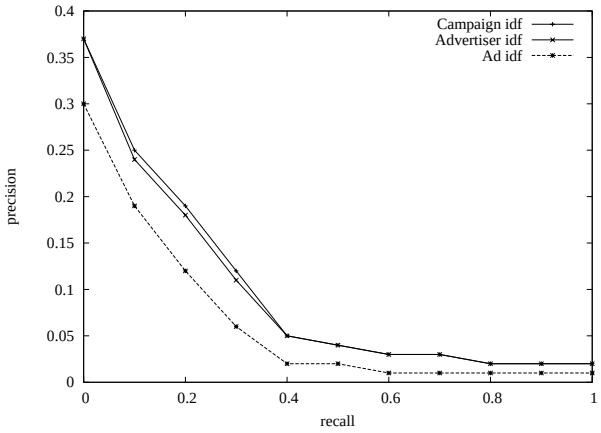

Figure 4: Precision-recall curves obtained for the `AD` strategy using ad, advertiser, and campaign idf factors.

This reflects the fact that terms might be better discriminators for a business topic than for an specific ad. This effect can be accomplished by calculating the factor relative to idf advertisers or campaigns instead of ads. In fact, campaign idf factors yielded the best results. Thus, they will be used in all the experiments reported from now on.

[2]Data in portuguese provided by an on-line advertisement company that operates in Brazil.

4.3 Results

Matching Strategies

Figure 5 displays the results for the matching strategies presented in Section 2. As shown, directly matching the contents of the ad to the triggering page (AD strategy) is not so effective. The reason is that the ad contents are very noisy. It may contain messages that do not properly describe the ad topics such as requisitions for user actions (e.g, "visit our site") and general sentences that could be applied to any product or service (e.g, "we delivery for the whole country"). On the other hand, an advertiser provided keyword summarizes well the topic of the ad. As a consequence, the KW strategy is superior to the AD and AD_KW strategies. This situation changes when we require the keywords to appear in the target Web page. By filtering out ads whose keywords do not occur in the triggering page, much noise is discarded. This makes ANDKW a better alternative than KW. Further, in this new situation, the contents of the ad becomes useful to rank the most relevant ads making AD_ANDKW (or AAK for "ads and keywords") the best among all described methods. For this reason, we adopt AAK as our baseline in the next set of experiments.

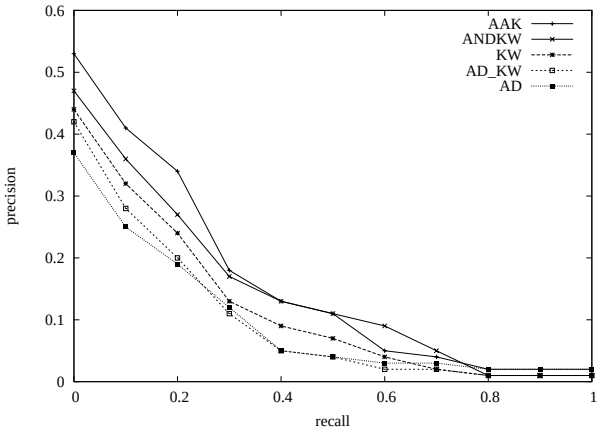

Figure 5: Comparison among our five matching strategies. AAK ("ads and keywords") is superior.

Table 1 illustrates average precision figures for Figure 5. We also present actual hits per advertisement slot. We call "hit" an assignment of an ad (to the triggering page) that was considered relevant by the evaluators. We notice that our AAK strategy provides a gain in average precision of 60% relative to the trivial AD strategy. This shows that careful consideration of the evidence related to the problem does pay off.

Impedance Coupling Strategies

Table 2 shows top ranked terms that occur in a page covering Argentinean wines produced using grapes derived from the Bordeaux region of France. The p column includes the top terms for this page ranked according to our tf-idf weighting scheme. The r column includes the top ranked expansion terms generated according to Eq. (6). Notice that the expansion terms not only emphasize important terms of the target page (by increasing their weights) such as "wines" and "whites", but also reveal new terms related to the main topic of the page such as "aroma" and "red". Further, they avoid some uninteresting terms such as "obtained" and "country".

Methods	Hits				11-pt average	
	#1	#2	#3	total	score	gain(%)
AD	41	32	13	86	0.104	
AD_KW	51	28	17	96	0.106	+1.9
KW	46	34	28	108	0.125	+20.2
ANDKW	49	37	35	121	0.153	+47.1
AD_ANDKW (AAK)	51	48	39	138	0.168	+61.5

Table 1: Average precision figures, corresponding to Figure 5, for our five matching strategies. Columns labelled #1, #2, and #3 indicate total of hits in first, second, and third advertisement slots, respectively. The AAK strategy provides improvements of 60% relative to the AD strategy.

Rank	p		r	
	term	score	term	score
1	argentina	0.090	wines	0.251
2	obtained*	0.047	wine*	0.140
3	class*	0.036	whites	0.091
4	whites	0.035	red*	0.057
5	french*	0.031	grape	0.051
6	origin*	0.029	bordeaux	0.045
7	france*	0.029	acideness*	0.038
8	grape	0.017	argentina	0.037
9	sweet*	0.016	aroma*	0.037
10	country*	0.013	*blanc**	0.036
...				
35	wines	0.010	-	-
...				

Table 2: Top ranked terms for the triggering page p according to our tf-idf weighting scheme and top ranked terms for r, the expansion terms for p, generated according to Eq. (6). Ranking scores were normalized in order to sum up to 1. Terms marked with '*' are not shared by the sets p and r.

Figure 6 illustrates our results when the set r of expansion terms is used. They show that matching the ads to the terms in the set r instead of to the triggering page p (AAK_T strategy) leads to a considerable improvement over our baseline, AAK. The gain is even larger when we use the terms in r to expand the triggering page (AAK_EXP method). This confirms our hypothesis that the triggering page could have some interesting terms that should not be completely discarded.

Finally, we analyze the impact on the ranking of using the contents of pages pointed by the ads. Figure 7 displays our results. It is clear that using only the contents of the pages pointed by the ads (H strategy) yields very poor results. However, combining evidence from the pages pointed by the ads with our baseline yields improved results. Most important, combining our best strategy so far (AAK_EXP) with pages pointed by ads (AAK_EXP_H strategy) leads to superior results. This happens because the two additional sources of evidence, expansion terms and pages pointed by the ads, are distinct and complementary, providing extra and valuable information for matching ads to a Web page.

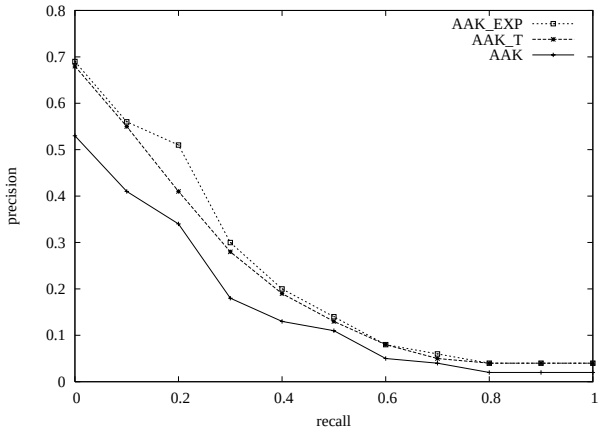

Figure 6: Impact of using a new representation for the triggering page, one that includes expansion terms.

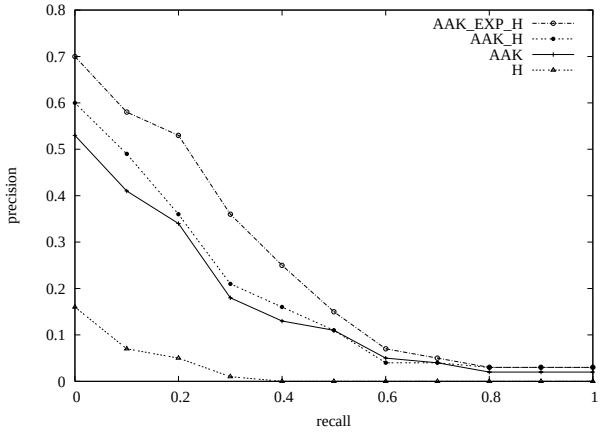

Figure 7: Impact of using the contents of the page pointed by the ad (the hyperlink).

Figure 8 and Table 3 summarize all results described in this section. In Figure 8 we show precision-recall curves and in Table 3 we show 11-point average figures. We also present actual hits per advertisement slot and gains in average precision relative to our baseline, `AAK`. We notice that the highest number of hits in the first slot was generated by the method `AAK_EXP`. However, the method with best overall retrieval performance was `AAK_EXP_H`, yielding a gain in average precision figures of roughly 50% over the baseline (`AAK`).

4.4 Performance Issues

In a keyword targeted advertising system, ads are assigned at query time, thus the performance of the system is a very important issue. In content-targeted advertising systems, we can associate ads with a page at publishing (or updating) time. Also, if a new ad comes in we might consider assigning this ad to already published pages in offline mode. That is, we might design the system such that its performance depends fundamentally on the rate that new pages

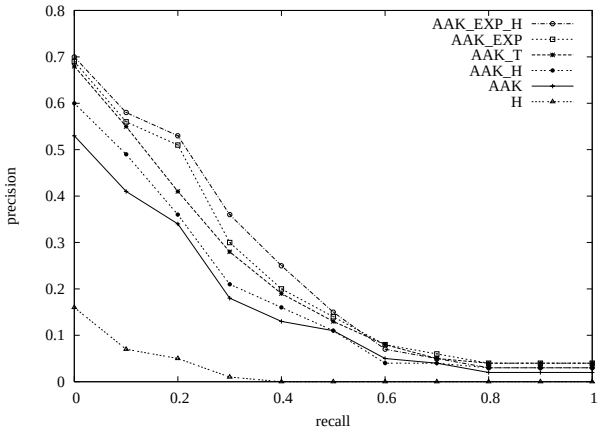

Figure 8: Comparison among our ad placement strategies.

Methods	Hits				11-pt average	
	#1	#2	#3	total	score	gain(%)
`H`	28	5	6	39	0.026	-84.3
`AAK`	51	48	39	138	0.168	
`AAK_H`	52	50	46	148	0.191	+13.5
`AAK_T`	65	49	43	157	0.226	+34.6
`AAK_EXP`	70	52	53	175	0.242	+43.8
`AAK_EXP_H`	64	61	51	176	0.253	+50.3

Table 3: Results for our impedance coupling strategies.

are published and the rate that ads are added or modified. Further, the data needed by our strategies (page crawling, page expansion, and ad link crawling) can be gathered and processed offline, not affecting the user experience. Thus, from this point of view, the performance is not critical and will not be addressed in this work.

5. RELATED WORK

Several works have stressed the importance of relevance in advertising. For example, in [14] it was shown that advertisements that are presented to users when they are not interested on them are viewed just as annoyance. Thus, in order to be effective, the authors conclude that advertisements should be relevant to consumer concerns at the time of exposure. The results in [9] enforce this conclusion by pointing out that the more targeted the advertising, the more effective it is.

Therefore it is not surprising that other works have addressed the relevance issue. For instance, in [8] it is proposed a system called ADWIZ that is able to adapt online advertisement to a user's short-term interests in a non-intrusive way. Contrary to our work, ADWIZ does not directly use the content of the page viewed by the user. It relies on search keywords supplied by the user to search engines and on the URL of the page requested by the user. On the other hand, in [7] the authors presented an intrusive approach in which an agent sits between advertisers and the user's browser allowing a banner to be placed into the currently viewed page. In spite of having the opportunity to use the page's content,

the agent infers relevance based on category information and user's private information collected along the time.

In [5] the authors provide a comparison between the ranking strategies used by Google and Overture for their keyword advertising systems. Both systems select advertisements by matching them to the keywords provided by the user in a search query and rank the resulting advertisement list according to the advertisers' willingness to pay. In particular, Google approach also considers the clickthrough rate of each advertisement as an additional evidence for its relevance. The authors conclude that Google's strategy is better than that used by Overture. As mentioned before, the ranking problem in keyword advertising is different from that of content-targeted advertising. Instead of dealing with keywords provided by users in search queries, we have to deal with the contents of a page which can be very diffuse.

Finally, the work in [4] focuses on improving search engine results in a TREC collection by means of an automatic query expansion method based on kNN [17]. Such method resembles our expansion approach presented in section 3. Our method is different from that presented by [4]. They expand user queries applied to a document collection with terms extracted from the top k documents returned as answer to the query in the same collection. In our case, we use two collections: an advertisement and a Web collection. We expand triggering pages with terms extracted from the Web collection and then we match these expanded pages to the ads from the advertisement collection. By doing this, we emphasize the main topics of the triggering pages, increasing the possibility of associating relevant ads with them.

6. CONCLUSIONS

In this work we investigated ten distinct strategies for associating ads with a Web page that is browsed (content-targeted advertising). Five of our strategies attempt to match the ads directly to the Web page. Because of that, they are called *matching strategies*. The other five strategies recognize that there is a vocabulary impedance problem among ads and Web pages and attempt to solve the problem by expanding the Web pages and the ads with new terms. Because of that they are called *impedance coupling strategies*.

Using a sample of a real case database with over 93 thousand ads, we evaluated our strategies. For the five matching strategies, our results indicated that planned consideration of additional evidence (such as the keywords provided by the advertisers) yielded gains in average precision figures (for our test collection) of 60%. This was obtained by a strategy called `AAK` (for "ads and keywords"), which is taken as the baseline for evaluating our more advanced impedance coupling strategies.

For our five impedance coupling strategies, the results indicate that additional gains in average precision of 50% (now relative to the `AAK` strategy) are possible. These were generated by expanding the Web page with new terms (obtained using a sample Web collection containing over five million pages) and the ads with the contents of the page they point to (a hyperlink provided by the advertisers).

These are first time results that indicate that high quality content-targeted advertising is feasible and practical.

7. ACKNOWLEDGEMENTS

This work was supported in part by the GERINDO project, grant MCT/CNPq/CT-INFO 552.087/02-5, by CNPq grant 300.188/95-1 (Berthier Ribeiro-Neto), and by CNPq grant 303.576/04-9 (Edleno Silva de Moura). Marco Cristo is supported by Fucapi, Manaus, AM, Brazil.

8. REFERENCES

[1] The Google adwords. Google content-targeted advertising. `http://adwords.google.com/select/ct_faq.html`, November 2004.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley-Longman, 1st edition, 1999.

[3] H. K. Bhargava and J. Feng. Paid placement strategies for internet search engines. In *Proceedings of the eleventh international conference on World Wide Web*, pages 117–123. ACM Press, 2002.

[4] E. P. Chan, S. Garcia, and S. Roukos. Trec-5 ad hoc retrieval using k nearest-neighbors re-scoring. In *The Fifth Text REtrieval Conference (TREC-5)*. National Institute of Standards and Technology (NIST), November 1996.

[5] J. Feng, H. K. Bhargava, and D. Pennock. Comparison of allocation rules for paid placement advertising in search engines. In *Proceedings of the 5th international conference on Electronic commerce*, pages 294–299. ACM Press, 2003.

[6] D. Hawking, N. Craswell, and P. B. Thistlewaite. Overview of TREC-7 very large collection track. In *The Seventh Text REtrieval Conference (TREC-7)*, pages 91–104, Gaithersburg, Maryland, USA, November 1998.

[7] Y. Kohda and S. Endo. Ubiquitous advertising on the www: merging advertisement on the browser. *Comput. Netw. ISDN Syst.*, 28(7-11):1493–1499, 1996.

[8] M. Langheinrich, A. Nakamura, N. Abe, T. Kamba, and Y. Koseki. Unintrusive customization techniques for web advertising. *Comput. Networks*, 31(11-16):1259–1272, 1999.

[9] T. P. Novak and D. L. Hoffman. New metrics for new media: toward the development of web measurement standards. *World Wide Web J.*, 2(1):213–246, 1997.

[10] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of plausible inference*. Morgan Kaufmann Publishers, 2nd edition, 1988.

[11] B. Ribeiro-Neto and R. Muntz. A belief network model for IR. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 253–260, Zurich, Switzerland, August 1996.

[12] A. Silva, E. Veloso, P. Golgher, B. Ribeiro-Neto, A. Laender, and N. Ziviani. CobWeb - a crawler for the brazilian web. In *Proceedings of the String Processing and Information Retrieval Symposium (SPIRE'99)*, pages 184–191, Cancun, Mexico, September 1999.

[13] H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, July 1991.

[14] C. Wang, P. Zhang, R. Choi, and M. Daeredita. Understanding consumers attitude toward advertising. In *Eighth Americas Conference on Information Systems*, pages 1143–1148, August 2002.

[15] M. Weideman. Ethical issues on content distribution to digital consumers via paid placement as opposed to website visibility in search engine results. In *The Seventh ETHICOMP International Conference on the Social and Ethical Impacts of Information and Communication Technologies*, pages 904–915. Troubador Publishing Ltd, April 2004.

[16] M. Weideman and T. Haig-Smith. An investigation into search engines as a form of targeted advert delivery. In *Proceedings of the 2002 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology*, pages 258–258. South African Institute for Computer Scientists and Information Technologists, 2002.

[17] Y. Yang. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In W. B. Croft and e. C. J. van Rijsbergen, editors, *Proceedings of the 17rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 13–22. Springer-Verlag, 1994.