SEMANTIC ASSOCIATIONS FOR CONTEXTUAL ADVERTISING

Massimiliano Ciaramita
Yahoo! Research Barcelona
massi@yahoo-inc.com

Vanessa Murdock
Yahoo! Research Barcelona
vmurdock@yahoo-inc.com

Vassilis Plachouras
Yahoo! Research Barcelona
vassilis@yahoo-inc.com

ABSTRACT

Contextual advertising systems place ads automatically in Web pages, based on the Web page content. In this paper we present a machine learning approach to contextual advertising using a novel set of features which aims to capture subtle semantic associations between the vocabularies of the ad and the Web page. We design a model for ranking ads with respect to a page which is learned using Support Vector Machines. We evaluate our model on a large set of manually evaluated ad placements. The proposed model significantly improves accuracy over a learned model using features from current work in contextual advertising.

Keywords: web advertising, contextual advertising, ranking, lexical associations

1. Introduction

The role of advertising in supporting and shaping the development of the Web has substantially increased over the past years. According to the Interactive Advertising Bureau (IAB, 2006), Internet advertising revenues in the U.S. totaled almost $8 billion in the first six months of 2006, a 36.7% increase over the same period in 2005, the last in a series of consecutive growths. Search, i.e., ads placed by Internet companies in Web pages or in response to specific queries, is the largest source of revenue, accounting for 40% of total revenue (IAB, 2006). The most important categories of Web advertising are *keyword match*, also known as *sponsored search,* or *paid listing,* which places ads in the search results for specific queries, and *content match*, also called *content-targeted advertising*, or *contextual advertising*, which places ads based on the Web page content.

Currently, most of the focus in Web advertising involves sponsored search. Content match has greater potential for content providers, publishers and advertisers, because users spend most of their time on the Web on content pages, as opposed to search engine result pages. However content match is a harder problem than sponsored search. Matching ads with query terms is to a certain degree straightforward, because advertisers themselves choose the keywords that describe their ads, which are matched against keywords chosen by users while searching. In contextual advertising, matching is determined automatically by the page content, which complicates the task considerably. Advertising touches challenging problems concerning how ads should be analyzed, and how systems accurately and efficiently select the best ads. This area of research is developing quickly in information retrieval. How best to model the structure and components of ads, and the interaction between the ads and the contexts in which they appear are open problems.

Information retrieval systems were designed to capture "relevance", and relevance is a basic concept in advertising as well. As with document retrieval, in the context of advertising we assume that an ad that is topically related to a Web page is relevant. Elements of an ad such as text and images tend to be mutually relevant, and often ads are placed in contexts which match the product at a topical level, such as an ad for sneakers placed on a sport news page. However, advertisements are not placed on the basis of topical relevance alone. For example, an ad for sneakers might be appropriate and effective on a page comparing MP3 players, because they share a target audience, for instance joggers. Still, they are different topics, and it is possible they share no common vocabulary. Conversely, there may be ads that are topically similar to a Web page, but cannot be placed there because they are inappropriate. An example might be placing ads for a product in the page of a competitor.

As advertisers attempt to capitalize on consumers' growing willingness to shop online, a number of studies have attempted to characterize Internet users who will become online consumers. Studies have focused on the effects of such factors as age, gender, and attitudes of trust toward online businesses [Levin et al. 2005; Zhou et al. 2007].

Mu and Galletta [2007] study the effects of pictures and words on Website recognition, to increase the likelihood of repeat visits to Websites. They conclude that salient pictures and text in a Web advertisement are more memorable if they are meaningful and represent the benefits of the product.

The language of advertising is rich and complex. For example, the phrase "I can't believe it's not butter!" implies at once that butter is the gold standard, and that this product is indistinguishable from butter. Furthermore, the imagery and layout of an ad contribute to the reader's interpretation of the text. A picture of a sunset in an ad for life insurance carries a different implication than a picture of a sunset in an ad for beer. The text may be dark on a light background, or light on a dark background, or placed in an image to carry a specific interpretation. The age, appearance and gender of people in an ad affect its meaning. Understanding advertisement involves inference processes which can be quite sophisticated [Vestergaard & Schroeder 1985], well beyond what traditional information retrieval systems are designed to cope with. In addition, the global context can be captured only partially by modeling text alone. These issues open new problems and opportunities for interdisciplinary research.

We investigate the problem of content match. The task is to choose ads from a pool to match the textual content of a particular Web page. Ads provide a limited amount of text: typically a few keywords, a title and brief description. The ad-placing system needs to identify relevant ads, from huge ad inventories, quickly and efficiently on the basis of this very limited amount of information. Recent work has proposed to improve content match by augmenting the representation of the page to increase the chance of a match [Ribeiro-Neto et al. 2005], or by using machine learning to find complex ranking functions [Lacerda et al. 2006], or by reducing the problem of content match to that of sponsored search by extracting keywords from the Web page [Yih et al. 2006]. All of these approaches are based on methods which quantify the similarity between the ad and the target page on the basis of traditional information retrieval notions such as cosine similarity and *tf-idf* features. The relevance of an ad for a page depends on the number of overlapping words, weighted individually and independently as a function of their individual distributional properties in the collection of documents or ads.

Based on the idea that successful advertising relies considerably on semantic inference, we propose an approach to content match which focuses on capturing subtler linguistic associations between the content of the page and the content of the ad. We implement these intuitions by means of simple and efficient distributional measures, which have been previously investigated in the context of natural language processing; e.g., in the area dealing with *lexical collocations*, that is, conventional multi-word expressions such as "big brother" or "strong tea", [Firth 1957]. We use these measures of semantic association to build features for a machine learning model based on ranking SVM [Joachims 2002a]. We evaluate our system on a dataset of real Web page-ad pairs, the largest evaluation presented to date, to the best of our knowledge. We compare our system with several baselines and learned models based on previous literature. The results show that our approach significantly outperforms other models and suggests promising new directions for future research. Our model uses pre-existing information in the form of simple word statistics which can be easily gathered in several ways. We propose several methods based on Web corpora, search engine indexes and query logs. The resulting model is essentially knowledge-free, as it does not require any language-specific resources beyond word counts. Furthermore, it can be applied to any language and any text or speech-based media.

2. Related Work

Web advertising presents peculiar engineering and modeling challenges and has motivated research in different areas. Systems need to be able to deal in real time with huge volumes of data and transactions involving billions of ads, pages, and queries. Hence several engineering constraints need to be taken into account; efficiency and computational costs are crucial factors in the choice of matching algorithms [The Yahoo! Research Team 2006]. Ad-placing systems might require new global architecture design; e.g., Attardi et al. [2004] proposed an architecture for information retrieval systems that need to handle large scale targeted advertising, based on an information filtering model. The ads that will appear on Web pages or search results pages will ultimately be determined taking into account expected revenues and the price of the ads. Modeling the microeconomics factors of such processes is a complex area of investigation in itself [Feng et al. 2007].

Another crucial issue is the evaluation of the effectiveness of the ad-placing systems. Studies have emphasized the impact of the quality of the matching on the success of the ad in terms of click-through rates [Gallagher et al. 2001]. Although click-through rates provide a traditional measure of effectiveness, it has been found that ads can be effective even when they do not solicit any conscious response and that the effectiveness of the ad is mainly determined by the level of congruency between the ad and the context in which it appears [Yoo 2006].

2.1. Keyword Based Models

Since the query-based ranking problem is better understood than contextual advertising, one way of approaching the latter would be to represent the content page as a set of keywords and then ranking the ads based on the keywords extracted from the content page. Carrasco et al. [2003] proposed clustering of bi-partite advertiser-keyword graphs for keyword suggestion and identifying groups of advertisers. Yih et al. [2006] proposed a system

for keyword extraction from content pages. The goal is to determine which keywords, or key phrases, best represent the topic of a Web page. Yih et al. develop a supervised approach to this task, from a corpus of pages where keywords have been manually identified. They show that a model learned with logistic regression outperforms traditional vector models based on fixed *tf-idf* weights. The most useful features to identify good keywords are term frequency and document frequency of the candidate keywords, and particularly the frequency of the candidate keyword in a search engine query log. Other useful features include the similarity of the candidate with the page's URL and the length, in number of words, of the candidate keyword. The accuracy of the best learned system is 30.06%, in terms of the top predicted keyword being in the set of manually generated keywords for a page, against 13.01% of the simpler *tf-idf* based model. While this approach is simple to apply and identifies potentially useful sources of information in automatically-generated keywords, it remains to be seen how accurate it is at identifying good ads for a page. We use a related keyword extraction method to improve content match.

2.2. Impedance Coupling

Ribeiro-Neto et al. [2005] introduce an approach to content match which focuses on the vocabulary mismatch problem. They notice that there is not enough overlap in the text of the ad and the target page to guarantee good accuracy; they call this the *vocabulary impedance problem*. To overcome this limitation they propose to generate an augmented representation of the target page by means of a Bayesian model previously applied to document retrieval [Ribeiro-Neto & Muntz 1996]. The expanded vector representation of the target page includes a significant number of additional words which potentially match some of the terms in the ad. They find that such a model improves over a baseline, evaluated by means of 11-point average precision on a test bed of 100 Web pages, from 0.168 to 0.253. One possible limitation is that this approach generates the augmented representation by crawling a significant number of additional related pages. It has also been argued [Yih et al. 2006] that this model complicates pricing of the ads because the keywords chosen by the advertisers might not be present in the content of the matching page.

2.3. Ranking Optimization with Genetic Programming

Lacerda et al. [2006] proposed to use machine learning to find good ranking functions for contextual advertising. They use the same dataset described in the paper by Ribeiro-Neto et al. [2005]. They use part of the data for training a model and part for evaluation purposes. They apply a genetic programming algorithm to select a ranking function which maximizes the average precision on the training data. The resulting ranking function is a non-linear combination of simple components based on the frequency of ad terms in the target page, document frequencies, document length and size of the collections. Lacerda et al. [2006] find that the ranking functions selected in this way are considerably more accurate than the baseline proposed in Ribeiro-Neto et al. [2005]; in particular, the best function selected by genetic programming achieves an average precision at position three of 0.508, against 0.314 of the baseline, on a test-bed of 20 Web pages.

2.4. Semantic Approaches to Contextual Advertising

Broder et al. [2007] notice that the standard string matching approach can be improved by adopting a matching model which additionally takes into account topical proximity. In their model the target page and the ad are classified with respect to a taxonomy of topics. The similarity of ad and target page estimated by means of the taxonomy provides an additional factor in the ads ranking function. The taxonomy, which has been manually built, contains approximately 6,000 nodes, where each node represents a set of queries. The concatenation of all queries at each node is used as a meta-document, ads and target pages are associated with a node in the taxonomy using a nearest neighbor classifier and *tf-idf* weighting. The ultimate score of an ad a_i for a page p is a weighted sum of the taxonomy similarity score and the similarity of a_i and p based on standard syntactic measures (vector cosine). On evaluation, Broder et al. [2007] report a 25% improvement for mid-range recalls of the syntactic-semantic model over the pure syntactic one. This approach is similar to ours in that it tries to capture semantic relations. The difference is that we do not rely on pre-existing language-dependent resources such as taxonomies.

2.5. Machine Translation Approaches to Contextual Advertising

Murdock et al. [2007] consider machine translation to overcome the vocabulary mismatch between target pages and ads. In more detail, the machine translation features they use correspond to the average translation probability of all words in the target page translated to the keywords or to the description of the ad, and the proportion of translations of the ad terms, or the ad keywords, that appear on the target page. Murdock et al. [2007] report that the machine translation probabilities produce statistically significant improvements in precision at rank one compared to a baseline, where the cosine similarity between the target page and each of the ad fields is weighted separately.

3. Formulation of the Ad-ranking Problem

Content match involves placing ads on a Web page, which we refer to as the *target page*. The typical elements of an advertisement are a set of *keywords*, a *title,* a *textual description,* and a hyperlink pointing to a page, the *landing page*, relative to a product or service (see *Illustration 1* for an example). In addition, an ad has an *advertiser id* and can be part of a *campaign*, i.e., a subset of all the ads with same advertiser id. This latter information can be used, for example, to impose constraints on the number of ads to display relative to a campaign or advertiser. While

this may be the most common layout, ads structure can vary significantly and include multimedia information.

3.1. The Ranking Problem for Ads-placing Systems

In general, the learning problem for an ad-placing system can be formalized as a ranking task. Let A be a set of ads and P the set of possible pages. A target page-ad pair (p,a), $p \in P$, $a \in A$, can be represented as a vector of real-valued features $\boldsymbol{x} = \Phi(p,a)$, where Φ is a feature map in a d-dimensional feature space $\boldsymbol{X} \subset \mathrm{R}^d$; i.e., Φ: $A \times P \rightarrow \boldsymbol{X}$. Useful features for ranking page-ad pairs include document similarity measures such as the vector cosine between the ad and the target page, possibly weighting each word's contribution with traditional *tf-idf* schemes [Baeza-Yates & Ribeiro-Neto 1999; Ribeiro-Neto et al. 2005]. The objective is to find a ranking function f: Φ $(p,a) \rightarrow \mathbf{R}$, which assigns scores to pairs (p,a), such that relevant ads are assigned a higher score than less relevant ads. In this paper we investigate several such functions. If Φ is a function that extracts one feature, the cosine between the ad, or one of its elements, and the target page, then f is a traditional information retrieval ranking function. We evaluate these types of features below in Section 0. However, we are mostly interested in ranking functions f_α which are parameterized by a real-valued vector $\alpha \in \mathrm{R}^d$, which weights the contribution of each feature individually. In particular, we focus on machine learning approaches to ads ranking in which the weight vector α is learned from a set of evaluated rankings.

3.2. Optimization Approach

In the most general formulation of the ad ranking task, the ad-placing system is given a page p, and uses the ranking function to score all pairs (p,a_i), $\forall\ a_i \in A$. Ads are then ranked by the score $f_\alpha(p,a_i)$. The final ranking will take into account the bid on the ad and, in general, the microeconomic model adopted by the provider[1]. Here we limit our attention to the quality of the chosen ads, and ignore this final step. In our evaluation we use a large set of target pages, for which several human judges have assessed the relevance of the ads in each page, placed by a base system. Since the pool of ads can be very large, different systems might propose entirely different lists of ads, with little or no overlap[2]. In order to carry out evaluation, in this paper we make the assumption that an initial guess at the best N ads for a target page is given by a base system, where N can vary for different pages. Accordingly, we reformulate the original problem as a re-ranking, or optimization, problem. The goal is to find a good ranking for a target page from a subset of A, the ads proposed by the base system. This setting is similar to that of Ribeiro-Neto et al. [2005] and Lacerda et al. [2006]. However, all systems we propose can be applied to the full task of scoring all ads in A. Therefore in this paper we focus on the problem of ranking, given a page p, all pairs (p,a_i), $\forall\ a_i \in A_p \subset A$, where A_p is the subset of A selected for page p by the base system.

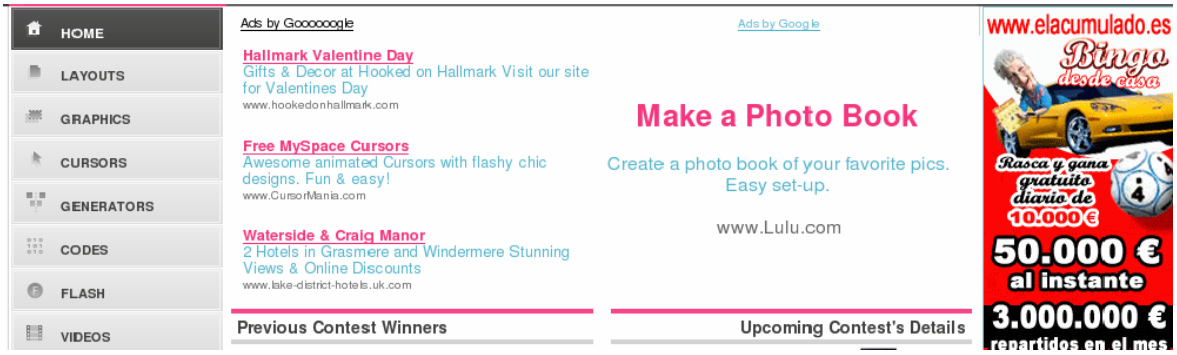

Illustration 2: An example of content targeted advertising. The ad on the far right is not part of the content targeted system.

4. Learning Semantic Associations for Contextual Advertisement

Previous work in content match has focused on traditional information retrieval notions of relevance. The relevance of an ad with respect to a target page is based on cosine similarity with *tf-idf* [Ribeiro-Neto et al. 2005]. More complex ranking functions are learned via genetic programming in [Lacerda et al. 2006], however the basic features which compose the selected ranking function are based on traditional measures such as term frequency, document frequency, document length and sizes of the collection of ads [Lacerda et al. 2006]. The limited context

[1] Constraints on the number of advertisers or campaigns can be easily implemented as post-ranking filters on the top of the ranked list of ads or included in the learning phase.

[2] This setting is problematic if the aim is to evaluate the quality of ads-placing systems by means of editorial human judgments, because the evaluation set is fixed.

provided by the ads, and the variance in type and composition of the target pages results in considerable vocabulary mismatch. We hypothesize that there may be many pairs of distinct words appearing in the ad and the target page which might be strongly related and provide useful features for ranking ads. As an example, the presence of pairs such as "exercise-diet", "usb-memory" or "lyrics-cd", might be useful in discriminating ads which may otherwise have the same overlapping keywords, and may appear similar based on simpler features. Modeling correlation at the lexical level could capture such semantic associations.

4.1. Semantic Association Features

We design an ad-placing system which exploits such lexical semantic associations by means of simple and efficient features. In the proposed system the feature map extracts properties of a target page-ad pair which include simple statistics about the degree of distributional correlation existing between words in the ad and words in the target page, in addition to more standard information retrieval features. We call this new class of features "semantic association features" because they capture distributional co-occurrence patterns between lexical items. Let (p,a) be a target page-ad pair and let $w_p \in p$, $w_a \in a$ be two words occurring in the target page and the ad respectively. To estimate the association between w_p and w_a we use several methods: point-wise mutual information (PMI), Pearson's χ^2 statistic [Manning & Schütze 1999], and clustering. PMI and Pearson's χ^2 are popular estimates of the degree of correlation between distributions. They have been used extensively in the natural language processing literature, e.g. to compare the similarity of corpora [Ciaramita & Baroni 2006], and to discover collocations, that is multiword expressions such as "real estate", which form idiomatic phrases [Dunning 1993]. All these measures are based on the joint and individual relative frequencies of the words considered; e.g., $P(w_p)$, $P(w_a)$ and $P(w_p,w_a)$. We computed word frequencies from different sources, namely, search engine indexes and query logs. As an example of the kind of associations picked up by such measures, *Table 1* lists the ten most strongly correlated words using Pearson's χ^2 statistic on the summary of the UK2006 collection [Castillo et al. 2006] for several word pairs found in our collection of ads and target pages.

Table 1: The 10 strongest correlated pairs of target page and ad words (w_p, w_a). The correlation corresponds to χ^2 with word counts from the UK2006 summary collection.

	w_p			
χ^2-ranked w_a	***basketball***	***hotel***	***cellphone***	***bank***
1	baseball	accommodation	ringtone	mortgage
2	hockey	airport	logos	secured
3	football	rooms	motorola	loan
4	nascar	inn	nokia	credit
5	nba	travel	cellular	equity
6	rugby	restaurant	cell	rate
7	nhl	destinations	samsung	refinance
8	sports	attractions	tone	accounts
9	mlb	reservation	ring	cash
10	lakers	flights	verizon	financial

Our goal is to use these association measures to build features which are useful for discriminating good matchings, based on the content of the target page and the ad. Section 5 below describes in detail the way these measures are computed and how they are aggregated as features. Overall we define a small set of features which can be computed efficiently. *Table 2* is a list of all features used in our experiments, including traditional and novel features. In the table, p stands for the target page, a stands for the ad, and T, D, K, L stand for the title, description, keywords and landing page of the ad. The features are described in detail in the corresponding sections. Similar to Yih et al. [2006], for the PMI and CSQ features, we use only a subset of the words in the target page and the ad (see Section 0).

4.2. Learning Ranking Function with SVM

Lacerda et al. [2006] use genetic programming to learn a ranking function, which maximizes the Average Precision [Baeza-Yates & Ribeiro-Neto 1999] of an ad-placing system. Following Joachims [2002a] we depart from the binary relevance provided by average precision and adopt Kendall's τ [Kendall 1938] as the objective function.

Kendall's τ is defined as follows:

$$\tau = \frac{C - D}{\frac{1}{2}N(N-1)} \quad (1)$$

Table 2: List of the features used for the learned models

Φ_i	Range	Description	Section
$x \in \{ a, a_T, a_D, a_K, a_L \}$	Real	$sim(p,x)$ where sim is cosine similarity	0
K	Binary	$\|[\forall w \in a_K \quad w \in p]\|$ and $\|[\exists w \in a_K \quad w \notin p]\|$, where $\|[\ \cdot\ ]\|$ denotes the indicator function	0
NIST	Real	Functional of overlapping n-grams between p_T and a_T	0
PMI	Real	max PMI(w_p, w_a) and avg PMI(w_p, w_a) where PMI is the point-wise mutual information between w_p and w_a	0
CSQ_z	Real	Number of pairs (w_p, w_a) in top z% ranked pairs according to χ^2	0
Clustering	Binary	Cluster identifier of the ad, page, and both ad and page	0

Kendall's τ measures the degree of correlation between two rankings and assesses the degree of significance of the correlation. Given two rankings R_1 and R_2 of the same set of N objects, C counts the number of concordant pairs of rankings in R_1 and R_2, while D counts the discordant pairs. The denominator is equal to the number of possible pairs for the N objects. Kendall's τ yields values between -1 and 1. A value of -1 means negative correlation, a value of 1 denotes complete agreement and 0 indicates that the rankings are independent. Kendall's statistic provides a more sensitive measure of correlation than average precision, and it has been used to optimize and improve the original ranking produced by search engines [Joachims 2002a][3]. Joachims [2002a] presents a formulation of Support Vector Machines learning [Vapnik 1995] based on Kendall's τ which minimizes the number of discordant pairs. We adopt a similar approach and use SVM to learn and evaluate several ranking functions. Other methods can be used to learn similar or related models such as perceptrons [Crammer & Singer 2003] and boosting [Schapire & Singer 2000]. The choice of SVM is motivated by the fact that it currently provides state of the art accuracy in several machine learning problems. In our experiments we used the implementation in SVM-light [Joachims 2002b][4].

4.3. Ranking Functions for Ad-placing

In summary, our method focuses on learning a ranking function f_α, which assigns a score to target page-ad pairs (p,a). We define a feature map $\Phi(p,a)$ which extracts traditional information retrieval features based on term and document frequencies, and also semantic association features based on statistical similarity measures. The score of a pair is a linear combination of the weights associated with each feature which defines the ranking function:

$$f_\alpha(p,a) = \langle \alpha, \boldsymbol{\Phi}(p,a) \rangle$$

where $\langle x,y \rangle$ is the inner product between vectors x and y, and α is learned with a ranking SVM.

5. Evaluation

A search engine has a database of millions of ads, which need to be matched to each of the Web pages in a stream of incoming contextual-match requests. While a retrieval algorithm is able to find ads that are topically related to the Web page, the task of contextual advertising is a high precision task. As stated earlier we propose that the ad ranking takes several steps: the first step finds a small subset of ads, the second re-ranks the ad subset to put the more relevant ads at the top of the list (see Section 3.2). In this section, we describe the data and the relevance assessment study, as well as the implementation of the features.

5.1. Data

Our data had 13,789 target page-ad pairs. Pairs for which no ad landing page was available were excluded from the data. We also excluded target pages for which there was only one candidate ad, and target pages for which all ad candidates were assigned the same relevance score by the assessors, because they are not useful for learning a ranking function. After filtering these examples from the data we were left with 11,231 pairs, corresponding to 980 target pages, where the average number of ads per target page is 11. All 980 target pages were used for evaluation.

5.2. Relevance Assessment

Each pair was evaluated by assessors on a three-point scale. The assessors were experts in content match

[3] Kendall's τ and Average Precision are related, since the number of discordant pairs D is a lower bound on Average Precision (Joachims, 2002a).

[4] Available from http://svmlight.joachims.org/.

evaluation and assigned a score of one for ads that were relevant to the target page, two for ads that were somewhat relevant, and three for ads that were nonrelevant. The assessor scores were then averaged to produce a composite score, and converted to binary relevance scores by assuming the target page-ad pairs that had a composite score of 2.34 or higher were nonrelevant, and all others were relevant. We chose a threshold of 2.34 because it corresponds to a sum of 7 for three assessors' scores. A sum of 7 can only be obtained with a combination of (2, 2, 3) or (3, 3, 1), which intuitively correspond to a collective vote of "nonrelevant". For pairs judged by only two editors, the combinations resulting in an average higher than 2.34 are (2,3) and (3,3). Furthermore, our assessment study (described below) supports the choice of 2.34 as a threshold, because that is the point of maximum agreement between our assessments and the composite binary relevance score. In our setting, only topical relevance was considered. Issues such as the appropriateness of content (for example, placing ads for a product in the target page of a competitor) or specificity (for example, placing ads for Christian pop music, as opposed to general pop music, in target pages about Christian music) were not considered. We compared the ranking learned with SVM to the ranking according to the composite scores.

We did not have access to the original relevance judgments, so we could not estimate the inter-assessor agreement between the original assessors. To estimate this, we judged 90 target pages (almost 10% of the evaluation data), sampled at random from the complete corpus, and then assessed the agreement between our judgments and the assessors' judgments. *Table 3* shows the results. In all, we assessed 997 target page-ad pairs. The Cohen's Kappa [Cohen 1960] between our assessments and the composite assessment score was 0.63. Cohen's Kappa is a measure of inter-assessor agreement. If two assessors agree completely, Cohen's kappa is one. If they disagree completely, Cohen's kappa is zero. A score of 0.63 indicates a high level of agreement with the composite assessment scores.

Table 3: The agreement between the composite scores of the original assessors and our own scores, for a sample of 997 target page-ad pairs for 90 target pages and all of their associated ads. Cohen's Kappa is 0.63.

		Our Assessments		
Original Assessments		Relevant	Nonrelevant	Total
	Relevant	424	92	516
	Nonrelevant	91	390	481
	Total	515	482	997

5.3. Experimental Setting

We implemented a retrieval baseline, which follows the approaches described in the literature [Ribeiro-Neto et al. 2005]. In these experiments, the ads were stemmed using the Krovetz stemmer [Krovetz 1993], and stop words were removed. The stop words were from a list of 733 words from the Terrier Retrieval Platform [Ounis et al. 2006]. The ads were indexed such that the ad description, title, and keywords were a "bag of words". The pairs of target page p and advertisement a were ranked according to their cosine-similarity, which employed *tf-idf* term weights, as follows:

$$sim(p,a) = \frac{\sum_{t \in p \cap a} w_{pt} \cdot w_{at}}{\sqrt{\sum_{t \in p} (w_{pt})^2} \cdot \sqrt{\sum_{t \in a} (w_{at})^2}} \quad (1)$$

In the above equation, the weight w_{pt} of term t in the target page p corresponds to its *tf-idf* score:

$$w_{pt} = tf \cdot log\left(\frac{|P|+1}{n_t + 0.5}\right) \quad (2)$$

where n_t is the target page frequency of term t, and $|P|$ is the number of target pages.

We also performed retrieval experiments using Okapi BM25 [Robertson et al. 1994], which has three parameters: $k1$, b and $k3$. The parameter b which adjusts the document length normalization in BM25 was fixed to 0.5, because the variance in the length of the ads was found to be small, and optimizing b is not expected to enhance retrieval performance. The parameter $k3$, which adjusts the saturation of the frequency of terms in the query, was set to 1000, as suggested by Robertson et al. [1994]. The parameter $k1$ was set after performing a 10-fold cross validation. All remaining experiments were performed with SVM, as described in Section 4.3.

We evaluated all experiments with precision at K, which is the number of relevant ads in the top K ads for K={1, 3, 5}. We also evaluated Kendall's τ, which is a measure of the degree to which two ranked lists agree, as defined in *Equation* ***Error! Reference source not found.***. As the composite score is the average of the assessor's scores, there may be ties. In this setting, we must account for the ties using a modified version of Kendall's τ [Adler 1957]:

$$\tau_b = \frac{C - D}{\sqrt{\left(\frac{1}{2}N(N-1) - T_1\right)\left(\frac{1}{2}N(N-1) - T_2\right)}} \quad (3)$$

where T_1 and T_2 correspond to the ties found in the first and the second ranking, respectively.

The set of target pages was partitioned so that no page appeared both in training and evaluation. The learned models used ten-fold cross validation so that a predicted ranking for one partition was given by a model trained on the remaining partitions. Statistical significance is reported for precision at K, using a two-tailed T-test.

5.4. Selecting Keywords from Target Pages

The semantic association features described below are based on correlations between pairs of words. To bound the number of comparisons we select a subset of terms in the target page and a subset of terms in the ad. From the ad, we use the keywords and the title. The subset of keywords extracted from a target page corresponds to the most informative keywords of the target page. We obtain the 50 most informative keywords using the term weighting model Bo1 from the Divergence From Randomness (DFR) framework [Amati 2003]. The model Bo1, which has been used effectively for automatic query expansion, assigns a high score to terms whose distribution is different in the target document p and in the set of all target pages. The weight $w(t)$ of a term t is computed as follows:

$$w(t) = tf_x \log_2 \frac{1 + P_n}{P_n} + \log_2(1 + P_n) \quad (4)$$

where tf_x is the frequency of a term in the target document p, and $P_n = F / |P|$ is the probability that the term t occurs in the set of target documents. F is the frequency of t in the set of $|P|$ target documents.

5.5. Features Implementation

We focus on four broad types of features: textual similarity (Section 5.5.1), keyword overlap (Section 5.5.2), semantic association (Section 5.5.3), and document-level similarity (Section 5.5.4).

5.5.1 Text Similarity Features

The first type of feature is the text similarity between a target page and the ad, or particular fields of the ad, such as the title of the advertisement (a_T). We also consider the textual similarity between the target page and the landing page a_L. We use cosine similarity with *tf-idf* term weights (see *Equations (1) and (2)* in Section 5.3).

5.5.2 Exact Match Features

A different type of feature, which Ribeiro-Neto et al. [2005] showed to be effective in content match of a target page and an advertisement corresponds to the overlap of keywords between the target page and the ad. Ribeiro-Neto et al. described their approach in a retrieval setting. They exclude the retrieved pairs of target page and ads, in which the target page did not contain all the ad keywords. In our data, out of 11,231 pairs, there were only 2000 pairs in which all ad keywords were present in the target page, corresponding to 700 pages out of 980. To capture that constraint, we consider two complementary binary features. For a given pair, the first feature is 1 if all the keywords of the ad appear in the target page, otherwise it is 0. The second feature is the complement of the first feature, (it is 0 when all the keywords of the advertisement appear in the target page, and otherwise it is 1). We denote this pair of features by "K" in the result tables.

Another way to measure overlap between the ads and the target pages is to identify n-grams they have in common. Modeling n-grams is also motivated by the observation that longer keywords, about four words long, lead to increased click-through rates [OneUpWeb 2005]. To provide a score that summarizes the level of overlap in n-grams between the ad and the target page, we computed the BLEU score. BLEU is a metric commonly used to evaluate machine translations. It was first proposed by Papineni et al. [2002]. In our data, the BLEU score between the ad title and the target page title was zero for nearly every pair. Instead we used a variant of BLEU, referred to as the NIST score [NIST Report 2002][5]:

$$NIST = \sum_{n=1}^{N} \left\{ \sum_{w_{1..n} co-occuring} Info(w_{1..n}) / \sum_{w_{1..n} output} (1) \right\} \bullet exp\left\{ \beta \log^2 \left[min\left(\frac{L_{sys}}{L_{ref}}, 1 \right) \right] \right\} \quad (6)$$

[5] http://www.nist.gov/speech/tests/mt/resources/scoring.htm (March 2007)

where $w_{1..k}$ is an n-gram of length k, β is a constant that regulates the penalty for short "translations", $N = 5$, L_{ref} is the average number of words in the target page title, and L_{sys} is the number of words in the ad title. In addition,

$$Info(w_{1..n}) = log_2\left(\frac{\text{count}(w_{1..n-1})}{\text{count}(w_{1..n})}\right) \tag{7}$$

where the counts of the n-grams are computed over the target page title. The idea is to give less weight to very common n-grams (like "of the") and more weight to infrequent, potentially very informative n-grams.

5.5.3 Semantic Association Features

Both text similarity features and exact match features, presented in Sections 5.4.1 and 5.4.2, are based on the matching of keywords between a target page and an ad. As Ribeiro-Neto et al. [2005] have pointed out, however, the number of matching keywords between the target and the ad can be low. We propose that the vocabulary mismatch between a target page and an ad can be overcome if we consider the semantic association between terms. We estimate the association of pairs of terms which do not necessarily occur in both the target page and the ad, using two statistical association estimates: point-wise mutual information (PMI) and Pearson's χ^2 [Manning & Schütze 1999]. We estimate PMI and Pearson's χ^2 with counts from three different corpora: i) the Web, ii) the summary of the UK2006 collection, consisting of 2.8 million Web pages, and iii) a query log from the Yahoo! search engine. In the case of the Web and the UK2006 collection, we count the number of documents in which terms occur, while in the case of the query log, we count the number of distinct queries in which terms occur.

Point-wise Mutual Information

The point-wise mutual information (PMI) between two keywords t_1 and t_2 is given as follows:

$$PMI(t_1, t_2) = log\frac{P(t_1, t_2)}{P(t_1)P(t_2)} \tag{5}$$

where $P(t)$ is the probability that keyword t appears in a document of the reference corpus and $P(t_1,t_2)$ is the probability that keywords t_1 and t_2 co-occur in a document. We use PMI to compute the association between a target document and an advertisement in the following way. For a subset of keywords from p and a subset of keywords from a, we compute the PMI of all the possible pairs of keywords. Then we use both the average $PMI_{AVG}(p,a)$ and the maximum $PMI_{MAX}(p,a)$ as two features.

Pearson's χ^2

Given a pair of terms t_1 and t_2, we count the number of documents in a reference corpus of M documents, in which the terms appear and we generate the following 2×2 table:

	t_1	$\neg t_1$
t_2	o_{11}	o_{12}
$\neg t_2$	o_{21}	o_{22}

where o_{11} is the number of documents that contain terms t_1 and t_2, o_{12} is the number of documents that contain term t_2 but not term t_1. Similarly, o_{22} is the number of documents that do not contain t_1 or t_2. We compute χ^2 by using the closed form equation:

$$\chi^2 = \frac{M(o_{11}o_{22} - o_{12}o_{21})^2}{(o_{11}+o_{12})(o_{11}+o_{21})(o_{12}+o_{22})(o_{21}+o_{22})} \tag{6}$$

We compute the χ^2 statistic for the pairs of keywords we extract from the target pages and the advertisements. Normally, the χ^2 statistic is compared to the χ^2 distribution to assess significance. In our case, due to the magnitude of counts, such comparison was not reliable. For this reason, we opted for considering a given percentage of the keyword pairs with the highest value of the χ^2 statistic. We sort the pairs in decreasing order of the χ^2 statistic, then for each pair we use the number of keyword pairs that have a χ^2 statistic in the top z% of all the pairs. We use a different feature for 0.1%, 0.5%, 1%, and 5%. We denote these features by CSQ_z where z represents the percentage of the most strongly related keyword pairs. For example, CSQ_1 for a given pair of target document and advertisement is the number of keyword pairs with a χ^2 statistic in the top 1% of the χ^2 statistic values.

5.5.4 Clustering

The features described above model the association between target pages and advertisements at the lexical level. A natural extension of our method could include features which estimate the similarity between ads and Web pages

at the document level. We carried out a few preliminary experiments in which we included document similarity features compiled by means of clustering. The intuition is that knowing what cluster an ad or web page belongs to might provide useful discriminative information. We used K-Means clustering [Duda et al. 2000], with *tf-idf* cosine similarity, computed separately on the collection of ads and on the collection of content pages. We selected three fixed sizes for the number *k* of clusters: 5, 10 and 15. The clustering features are categorical features consisting in the cluster id of the ad, the cluster id of the Web page, and the pair of ids for both, for all three values of *k*. An advantage of using clustering features is that, as with the lexical semantic features, they can be computed efficiently from the raw data without any additional knowledge or language specific tools.

6. Results

In this section we describe the results of the empirical evaluation. We compare our method with several information retrieval baselines, as well as machine learned baseline methods.

6.1. Retrieval Baselines

The problem of content match can be cast as an information retrieval task, as in the baseline experiments of Ribeiro-Neto *et al*. [2005]. We match target documents with ads by performing retrieval and rank pairs according to the cosine similarity between the target document and the advertisement, or specific fields of the advertisement. Treating content match as a retrieval task may result in retrieving fewer pairs because of a lack of matching keywords. For these cases, we randomly rank the pairs that have not been retrieved. We perform this process five times and report average evaluation measures. The standard deviation in all cases was equal to or less than 0.003, suggesting there is not a high degree of variability in the results due to the random re-ranking of the pairs that have not been retrieved.

Table 4 summarizes the results of experiments with the retrieval baselines. We report Kendall´s τ_b, and precision at 5, 3 and 1. The table shows that cosine similarity performs as well as Okapi BM25, where b=0.5 and $k1$ is optimized with ten-fold cross validation. We use cosine when computing text similarity for the rest of the experiments, because it has no associated free parameters. When considering the different fields of the ads, we see that the title is the most effective field for computing the similarity with respect to all evaluation measures.

Table 4: The mean of five retrieval runs, where pairs of target documents and advertisements that have not been ranked by the retrieval system are randomly re-ranked.

Cosine similarity	Kendall's τ_b	P@5	P@3	P@1
p-a	0.233	0.623	0.663	0.685
p-a$_T$	**0.251**	**0.632**	**0.664**	**0.690**
p-a$_D$	0.216	0.610	0.642	0.659
p-a$_K$	0.206	0.616	0.646	0.681
p-a$_L$	0.157	0.604	0.646	0.680
BM25	Kendall's τ_b	P@5	P@3	P@1
p-a	0.237	0.627	0.655	0.676

In *Table 4*, as in all tables in this paper, the precision at rank one is higher than precision at ranks three and five. This is in part due to the fact that not all of the Web pages in our data have five relevant ads. In fact, some of the Web pages have fewer than five ads total. Because of this, if the system is ranking the relevant ads near the top of the list, precision at five would always be lower than precision at one. For example, consider a Web page which has one relevant ad. If it is placed at rank one, precision at one will be 1.0 and precision at five will be 0.20. The fact that our results are uniformly better for precision at rank one suggests that the system is placing most of the relevant ads at the top of the ranked list.

6.2. Learned Models

In this section we evaluate the effectiveness of the learning approach based on SVM. In this setting the cosine similarity between the target page and the ad, or a particular ad field, is used as a feature and weighted individually by SVM. In addition, we can combine arbitrary features such as those described above. We first evaluate the performance of the textual similarity features, described in Section 5.5.1, the keyword overlap features described in Section 5.5.2, the semantic association features discussed in Section 5.5.3, and finally, the clustering features described in Section 5.5.4.

6.2.1 Cosine Similarity Features

The evaluation of the cosine similarity features is shown in *Table 5*. As an example, *p-aa*$_{TDK}$ identifies four features: the cosine similarity between *p* and *a* (the entire ad), *p* and a_T (the ad title), *p* and a_D (the ad description), and *p* and a_K (the ad keywords). As expected, the cosine similarity between the target page and the advertisement as

a feature performs as well as the corresponding retrieval experiment (see *Table 4*). The SVM-weighted combination of features improves Kendall's τ_b but the changes in precision between *p-a* or *p-a*$_T$ and *p-a*$_{TDK}$, respectively, are not significant. The best performing combination of features (the row denoted *p-a*$_{TDKL}$) serves as the baseline for comparisons and significance tests throughout the rest of the paper.

Table 5: Evaluation of cosine similarity features between the target pages and the advertisements or fields of the advertisements

Features	Kendall's τ_b	P@5	P@3	P@1
p-a	0.243	0.625	0.663	0.684
p-a$_T$	0.266	0.632	0.665	**0.688**
p-a$_D$	0.221	0.611	0.641	0.657
p-a$_K$	0.217	0.617	0.648	0.681
p-a$_L$	0.157	0.603	0.640	0.665
p-a$_{TDK}$	0.276	0.635	0.668	0.686
p-a$_{TDKL}$	**0.279**	**0.637**	**0.676**	0.687
p-aa$_L$	0.255	0.630	0.663	0.685
p-aa $_{TDK}$	0.275	0.634	0.668	0.685
p-aa$_{TDKL}$	0.275	0.636	0.671	0.687

6.2.2 Keyword Overlap Features

As noted by Ribeiro-Neto et al. [2005], ads whose keywords are all contained in a target page are a good match for that page. In *Table 6*, *p-aa*$_{TDKL}$K performs better than the baseline system, although the result is not statistically significant. We carry this system forward in future experiments because it represents the state-of-the art, and is the best performing combination of features in *Table 6*.

Table 6: Evaluation of keyword overlap features

Features	Kendall's τ_b	P@5	P@3	P@1
p-aa$_L$	0.255	0.630	0.663	0.685
p-a$_{TDKL}$ (baseline)	0.279	0.637	0.676	0.687
p-aa$_{TDKL}$	0.275	0.636	0.671	0.687
p-aa$_L$K	0.261	0.635	0.673	0.707
p-a$_{TDKL}$K	0.269	0.638	0.673	0.696
p-aa$_{TDKL}$K	**0.286**	**0.643**	**0.681**	**0.716**

To enforce a stricter match between the ad and the target page, we look for shared n-grams summarized by the NIST score between the titles of the ad and the target. We also tried the NIST and BLEU scores between the ad landing page and the target page, but found that these did not perform as well. *Table 7* compares the baseline system and the best performing system from *Table 6* with the NIST score included. The improvement in precision at rank one is statistically significant, and we carry this model forward in the following experiments.

Table 7: Adding the NIST scores as features to the best performing keyword overlap features gives a statistically significant improvement in precision at 1 over the baseline system, using a two-tailed T-test, $p < 0.05$.

Features	Kendall's τ_b	P@5	P@3	P@1
baseline	**0.279**	0.637	0.676	0.687
p-aa$_{TDKL}$K-NIST	0.278	**0.638**	**0.681**	**0.732***

6.2.3 Semantic Association Features

Table 8 summarizes the results of the model which includes the semantic association features. Rows labeled with PMI show point-wise mutual information features. CSQ$_z$ indicates the χ^2 features with corresponding threshold *z* on the percentage of significant terms. As these features use frequencies from external corpora we indicate with "Web" the search engine index, with "UK" the UK2006 summary collection, and with "Qlog" the query logs.

The inclusion of this class of features improves performance compared to the baseline. The best performing combination of features is the χ^2 statistic where the feature is estimated from a search engine query log. The performance of this model is slightly better than the performance of the model using point-wise mutual information, but the differences between the two are not significant. The results indicated with an asterisk or dagger are statistically significant with respect to the baseline.

Table 8: Evaluation of semantic features.

Features	Kendall's τ_b	P@5	P@3	P@1
baseline	0.279	0.637	0.676	0.687
$p\text{-}aa_{\text{TDKL}}$K-NIST-$\text{PMI}_{\text{Web}}$	0.321	0.654	**0.698**	0.745†
$p\text{-}aa_{\text{TDKL}}$K-NIST-$\text{PMI}_{\text{UK}}$	**0.322**	**0.655**	0.696	0.741†
$p\text{-}aa_{\text{TDKL}}$K-NIST-$\text{PMI}_{\text{Qlog}}$	0.290	0.641	0.684	0.716
$p\text{-}aa_{\text{TDKL}}$K-NIST-$\text{CSQ}_{0.1,\text{Web}}$	0.290	0.644	0.688	0.733*
$p\text{-}aa_{\text{TDKL}}$K-NIST-$\text{CSQ}_{0.1,\text{UK}}$	0.295	0.643	0.688	0.735*
$p\text{-}aa_{\text{TDKL}}$K-NIST-$\text{CSQ}_{1,\text{Qlog}}$	0.313	0.652	0.697	**0.753†**

** Results indicated with an asterisk are statistically significant at the p<0.05 level.*
† Results indicated with a dagger are statistically significant at the p<0.01 level.
Significance results are with respect to the baseline system, using a two-tailed T-test.

6.2.4 Association at the Document Level

The semantic association features attempt to solve the vocabulary mismatch problem by finding pairs of words in the target page and ad that are correlated. This approach can be extended to capture semantic associations at the document level, for example, by means of clustering. We performed a preliminary investigation of the impact of clustering, and present the results in *Table 9*. The table shows the results of adding clustering to the baseline system, to the baseline with the NIST features, and to the χ^2 and mutual information features. The precision at rank one results for all clustering systems were statistically significantly better than the baseline system. The clustering improves results for each individual model. In particular, adding clustering to the best model produces the best results for all evaluation metrics. We did not carry out an exhaustive investigation of clustering, however these results suggest this is a promising area for future research.

Table 9: A preliminary investigation of cluster-based features suggests this may be an area for future work.

Features	τ_b	P@5	P@3	P@1
baseline	0.279	0.637	0.676	0.687
$p\text{-}aa_{\text{TDKL}}$K-Clustering	0.299	0.648	0.695	0.738*
$p\text{-}aa_{\text{TDKL}}$K-NIST-Clustering	0.301	0.645	0.697	0.742†
$p\text{-}aa_{\text{TDKL}}$K-NIST-$\text{PMI}_{\text{Web}}$-Clustering	0.317	0.658	0.703	0.747†
$p\text{-}aa_{\text{TDKL}}$K-NIST-$\text{CSQ}_{1,\text{Qlog}}$–Clustering	**0.326**	**0.660**	**0.716***	**0.757††**

** Results indicated with an asterisk are statistically significant at the p<0.05 level.*
† Results indicated with a dagger are statistically significant at the p<0.01 level.
†† Results indicated with a double dagger are statistically significant at the p < 0.001 level.
Significance results are with respect to the baseline system, using a two-tailed T-test.

7. Discussion

Treating content match as a retrieval problem is a natural formulation of the ad-placing task. In this task, the ad title proved to be the most effective representative of the ad. One drawback of this approach is that it is not clear how to include other information about the ads and target pages. In a retrieval system, we are limited by the representation of the ad and the representation of the target page. It is not possible to include relationships between terms found in other corpora (such as the point-wise mutual information or χ^2 statistics), or relationships between documents, as represented by the clustering features. To incorporate these types of information, a different framework is necessary. As expected, the performance of the SVM on the cosine similarity features was indistinguishable from the retrieval results for Precision at *K* (as shown in *Table 4* and *Table 5*). We would expect this to be true, because the learning model uses only one feature, which is the same as the retrieval model.

Table 10: Evaluation of cosine similarity versus PMI alone

Features	Kendall's τ_b	P@5	P@3	P@1
$p\text{-}a_{\text{T}}$	0.266	0.632	0.665	0.688
PMI	0.219	0.620	0.652	0.680

In a learning framework, it is possible to deconstruct the ad into its constituent parts, and weight each part's contribution separately. In doing so we are able to put a soft constraint that the keywords from the ad must be present in the target page, and we found that this improves performance (*Table 6*), in agreement with findings by

Ribeiro-Neto et al. [2005]. Cosine similarity between the target page and the ad, represented as a vector of *tf-idf* weights, is a good feature.

Figure 1 shows the distribution of cosine similarity scores for the relevant and nonrelevant classes. We can see that the two classes have different distributions of scores; in fact they are statistically significantly different according to a T-test, $p<0.05$. *Figure 2* shows a similar plot for the point-wise mutual information and the χ^2 features.

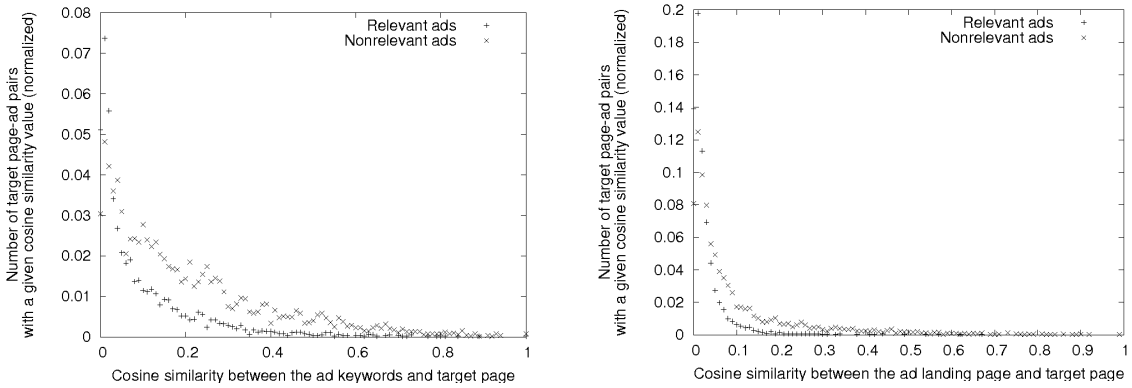

Figure 1: The frequency of cosine similarity scores, where cosine similarity is computed between the ad keywords and the target page (left) and the ad landing pages and target pages (right).

Cosine similarity only allows matching at the term level. The models based on NIST and BLEU capture a small amount of structure in the form of n-grams. The fact that they improve performance implies that language structure is an important aspect in finding relevant ads. N-grams represent a relatively unsophisticated structure and the application of more complex structures merits further investigation.

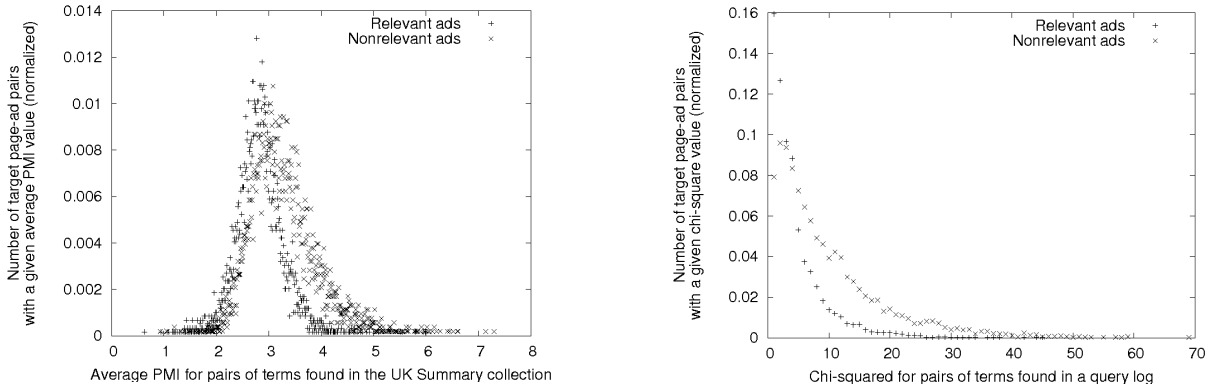

Figure 2: The frequency distribution of point-wise mutual information computed from the UK2006 summary collection (left) and Pearson's χ^2 statistic computed from the query log (right) for relevant and nonrelevant ads.

None of the cosine similarity features, or NIST or BLEU captures semantic relations between a target page and an ad. We introduced the semantic association features, based on point-wise mutual information and χ^2 statistics. The features built on PMI and χ^2 summarize the relatedness between an ad and a target page, beyond textual overlap. With these features, we can exploit relationships between terms that do not appear in both the target page and the ad.

The features based on clustering show that similarity at the document-level provides useful discriminative information. The topical relatedness of a set of pages is more reliably assessed because the distribution of terms in the set is not as sparse as in the individual target pages or ads. If we know that documents are related, we can exploit this fact to better place ads.

Our evaluation is not directly comparable to the systems described by Ribeiro-Neto et al. [2005] and Lacerda et

al. [2006]. However, our findings concerning the retrieval and learned baselines are consistent with their results. Our evaluation is considerably larger than theirs and we obtain statistically significant improvements over the information retrieval and learned baselines. In addition, our features can be applied to any learning framework, including genetic programming.

8. Conclusions

The role of advertising in supporting and shaping the development of the Web has substantially increased over the past years. The task of contextual advertising is complicated by the necessity of determining matches automatically based on the page content. The information retrieval notion of relevance and traditional search concepts are insufficient for content match. The language of advertising involves inferential processes which can be quite sophisticated. We propose a first step towards addressing such issues by means of simple distributional features and a machine learning approach. Based on the idea that successful advertising relies considerably on semantic inference, our approach focuses on more subtle linguistic associations between the content of the page and the ad.

Our method is language independent and does not require any external resources. The features range from simple word overlap to semantic associations using point-wise mutual information and χ^2 between pairs of terms. Cosine similarity is a robust feature both in retrieval and learning settings, and PMI on its own achieves slightly lower precision than cosine similarity. The semantic association features capture similarity along different dimensions than cosine similarity, and they are present in all the best performing models we experimented with in this article. Clustering seems another promising feature of semantic association at the document-level, and warrants further investigation. For example, it may be useful for avoiding inappropriate matches.

Other areas of future work include applying these techniques to multimedia advertising and extending them to include light-weight language-aware features. In addition, features of the microeconomic model can be incorporated into the same learning framework to optimize the revenue from contextual advertising.

Acknowledgment

This research was funded by Yahoo!, Inc.

REFERENCES

Adler, M. L., "A Modification of Kendall's Tau for the Case of Arbitrary Ties in Both Rankings," *Journal of the American Statistical Association*, Vol. 52, No. 277:33-35, 1957.

Amati, G., "Probabilistic Models for Information Retrieval Based on Divergence from Randomness," *PhD thesis*, Department of Computing Science, University of Glasgow, 2003.

Attardi, G., A. Esuli, and M. Simi, "Best Bets, Thousands of Queries in Search of a Client," *Proceedings of the 13th International Conference on World Wide Web*, Alternate Track Papers and Posters, ACM Press, 2004.

Baeza-Yates, R. and B. A. Ribeiro-Neto, *Modern Information Retrieval*, ACM Press/Addison-Wesley, 1999.

Broder, A., M. Fontoura, V. Josifovski and L. Riedel, "A Semantic Approach to Contextual Advertising," *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, 2007.

Carrasco, J.J., D. Fain, K. Lang, and L. Zhukov, "Clustering of Bipartite Advertiser-Keyword Graph," *Proceedings of the Workshop on Clustering Large Datasets*, IEEE Conference on Data Mining, IEEE Computer Society Press, 2003.

Castillo, C., D. Donato, L. Becchetti, P. Boldi, S. Leonardo, M. Santini, and S. Vigna, "A Reference Collection for Web Spam," *ACM SIGIR Forum*, Vol. 40, No. 2:11-24, 2006.

Ciaramita, M. and M. Baroni, "A Figure of Merit for the Evaluation of Web-Corpus Randomness," *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, 2006.

Cohen, J., "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, Vol. 20:37-46, 1960.

Crammer, K. and Y. Singer, "A New Family of Online Algorithms for Category Ranking," *Journal of Machine Learning Research*, Vol. 3:1025-1058, 2003.

Duda, R.O. and P. E. Hart and D. G. Stork, "Pattern Classification (2nd edition)", Wiley Interscience, 2002.

Dunning, T., "Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics*, Vol. 19, No. 1:61-74, 1993.

Feng, J., H. Bhargava, and D. Pennock, "Implementing Sponsored Search in Web Search Engines: Computational Evaluation of Alternative Mechanisms," *Informs Journal on Computing*, Vol. 19, No 1:134-148, 2007.

Firth, J.R, A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, pp. 1-32. Oxford: Philological Society, 1957. Reprinted in F.R. Palmer (ed.), Selected Papers of J. R. Firth 1952-1959, London:

Longman, 1968.
Gallagher, K., D. Foster, and J. Parsons, "The Medium is not the Message: Advertising Effectiveness and Content Evaluation in Print and on the Web," *Journal Of Advertising Research*, Vol. 41, No. 4:57-70, 2001.
IAB: Interactive Advertising Bureau, "IAB Internet Advertising Revenue Report," available at http://www.iab.net/resources/adrevenue/, 2006
Joachims, T., "Optimizing Search Engines Using Clickthrough Data," *Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 133-142, 2002a.
Joachims, T., "Learning to Classify Text Using Support Vector Machines," Dissertation, Kluwer, 2002b.
Kendall, M.G., "A New Measure of Rank Correlation," *Biometrika*, Vol. 30:81-93, 1938.
Krovetz, R., "Viewing Morphology as an Inference Process," *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, pp. 191-202, 1993.
Lacerda, A., M. Cristo, M.A. Goncalves, W. Fan, N. Ziviani, and B. Ribeiro-Neto, "Learning to Advertise," *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, pp. 549-556, 2006.
Levin, A.M., I.P. Levin, and J.A. Weller, "A Multi-Attribute Analysis of Preferences for Online and Offline Shopping: Differences Across Products, Consumers, and Shopping Stages," *Journal of Electronic Commerce Research*, Vol. 6, No. 4:281-290, 2005.
Manning, C.D. and H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge Massachusetts, 1999.
Mu, E. and D.F. Galletta, "The Effects of Meaningfulness of Salient Brand and Product-Related Text and Graphics on Web Site Recognition," *Journal of Electronic Commerce Research*, Vol. 8, No.2:115-127, 2007.
Murdock, V., M. Ciaramita, and V. Plachouras, "A Noisy Channel Approach to Contextual Advertising," *Proceedings of the 1st International Workshop on Data Mining and Audience Intelligence for Advertising (ADKDD'07)*, 2007.
NIST Report. "Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics," www.nist.gov/speech/tests/mt/doc/ngram-study.pdf (as of March 2007), 2002.
OneUpWeb, "How Keyword Length Affects Convertion Rates," Available at http://www.oneupweb.com/landing/keywordstudy_landing.htm, 2005.
Ounis, I., G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma, "Terrier: A High Performance and Scalable Information Retrieval Platform," *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, Seattle Washington, USA, 2006.
Papineni, K., S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," *Proceedings of ACL-2002: 40th annual meeting of the Association for Computational Linguistics* pp. 311-318, 2002.
Ribeiro-Neto, B., M. Cristo, P.B. Golgher, and E.S. De Moura, "Impedance Coupling in Content-targeted Advertising," *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, pp. 496-503, 2005.
Ribeiro-Neto, B. and R. Muntz, "A Belief Network Model for IR," *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, pp. 253-260, 1996.
Robertson, S.E., S. Walker, M.M. Hancock-Beaulieu, M. Gatford, and A. Payne,"Okapi at TREC-4, *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*," Nov. 1-3, 1995, pp.73-86, 1995.
Schapire, R.E. and Y. Singer, "BoosTexter: A Boosting-based System for Text Categorization," *Machine Learning*, Vol. 39, No. 2/3:135-168, 2000.
Vapnik, V.N., *The Nature of Statistical Learning Theory*. Springer, 1995.
Vestergaard, T. and T. Schroeder, *The Language of Advertising*. Oxford, Blackwell. 1985.
The Yahoo! Research Team, "Content, Metadata, and Behavioral Information: Directions for Yahoo! Research," *IEEE Data Engineering Bulletin*, December 2006.
Yih, W., J. Goodman, and V.R. Carvalho, "Finding Advertising Keywords on Web Pages," *Proceedings of the 15th international conference on World Wide Web*, pp. 213-222, 2006.
Yoo, C.Y., "Preattentive Processing of Web Advertising"', *Ph.D. Thesis*, University of Texas at Austin, 2006.
Zhou, L., L. Dai, and D. Zhang, "Online Shopping Acceptance Model – A Critical Survey of Consumer Factors in Online Shopping," *Journal of Electronic Commerce Research*, Vol. 8, No. 1:41-62, 2007.