ACAKS: An Ad-Collection-Aware Keyword Selection Approach for Contextual Advertising

Klessius Berlt[1], Marcela Pessoa[1,2], Mauro Herrera[1],
Carlos Alessandro Sena[1], Marco Cristo[1], Edleno Silva de Moura[1]

[1] Universidade Federal do Amazonas, Brazil
[2] FUCAPI, Brazil
{klessius, marcelappessoa, mauro.rh, marco.cristo}@gmail.com,
{alessandro.sena, edleno}@dcc.ufam.edu.br

Abstract. In this work we address the problem of selecting keywords from a web page in order to submit them to an ad selection system. Several previous research found in literature have proposed machine learning strategies to determine these keywords in different contexts, such as emails and web pages. Such machine learning approaches usually have the goal of selecting keywords considered as good by humans. We here propose a new machine learning strategy where the selection is driven by the expected impact of the keyword in the final quality of the ad placement system, which we name here as *ad-collection-aware* keyword selection. This new approach relies on the judgment of the users about the ads each keyword can retrieve. Although this strategy requires a higher effort to build the training set than previous approaches, we believe the gain obtained in recall is worth enough to make the *ad collection aware* approach a better choice. In experiments we performed with an ad collection and considering features proposed in a previous work, we found that the new *ad-collection-aware* approach led to a gain of 62% in the recall over the baseline without dropping the precision values. Besides the new alternative to select keywords, we also study the use of features extracted from the ad collection in the task of selecting keywords.

Categories and Subject Descriptors: H. Information Systems [**H.m. Miscellaneous**]: Databases

Keywords: contextual advertising, keyword selection

1. INTRODUCTION

Most of contextual advertising strategies use the text of the web page the user is browsing to help on the task of discovering what ads to display. Although, the entire textual content of a web page is a very noisy source of information, there are several terms that do not have a direct relation with the content of the page. The usage of such terms decreases the quality of the advertising systems while increases the communication and latency costs [Anagnostopoulos et al. 2007]. To solve this problem, these systems use algorithms to estimate the importance of the keywords in the text to be processed. Such estimates are then used as input for other algorithms, such as those employed in ad matching, as well as to select subsets of keywords. This work proposes a new strategy to select from a web page the keywords that would be more useful on an advertising system.

A simple strategy to determine the importance of a keyword would be use classical information retrieval statistics such as term frequency and inverse of document frequency [Baeza-Yates and Ribeiro-Neto 1999; Salton et al. 1974]. Such statistics, however, might not capture specific aspects of the problem, for instance, the marketing appeal of a term. To address this problem of determining which sequences of terms from a textual content would make good ad keywords, more sophisticated ideas have been proposed in literature. In particular, previous work has adopted machine learning strategies in which candidate terms are described by several features. The aim in these cases is to learn specific patterns which would clearly distinguish ad keywords.

Such machine learning approaches usually have the goal of selecting keywords considered as good by humans. We here propose a new strategy where the selection is driven by the expected impact of the keyword in the final quality of the ad placement system, which we name as *ad collection aware* also referred to as *ACAKS*. Our intuition is that the keywords selected by the users as the best for advertising are not always the most appropriate for contextual advertising systems.

More specifically, on this work we take advantage of the ad collection by changing the strategy used to compose the training collection which guides the learning process. Instead of asking users to directly giving examples of what are the good keywords found in the training pages (which we call *traditional approach*), we gather the ads which have a match with any sequence of one or more word candidates to be a keyword (*keyword candidate*) found in the training pages and ask the users to evaluate their relevance. As we show, this strategy provides competitive results. Further, while at a first glance it appears to be prohibitively expensive, we show that it is possible to perform a training with the *ACAKS* method and that it produces, at least in our experiments, results superior to the baseline, being an interesting alternative to select keywords from web pages for ad placement purposes.

Besides studying the performance of the features proposed by the authors in [Yih et al. 2006] on keyword selection, we also assess the impact of using a new set of ads derived from the ad collection. While these new features are not studied in [Yih et al. 2006], which is our baseline, previous research articles indicate that ad collection information is useful to improve the quality of results in ad placement systems [Ribeiro-Neto et al. 2005; Chakrabarti et al. 2008], which motivated us to investigate its use also in the keyword selection problem.

The inclusion of ad collection features also makes the comparison between the methods fairer, since we can say that the *ACAKS* approach indirectly uses information about the ad collection as part of its keyword selection method. The selection of keywords in the method *ACAKS* is guided by the relevance of ads they can bring, and we can say it implicitly uses ad collection information in its keyword selection process. When including ad features in the experiments, we also provide information about ad collection to the traditional approach, thus allowing both methods to take advantage of this information.

This work is organized as follows. Section 2 discusses the related work. Section 3 describes our proposed strategy to keyword selection. Section 4 presents an evaluation of our proposal, comparing it to other approaches found in the literature. Finally, section 5 presents conclusions and future research directions on the topic.

2. RELATED WORK

The design of web search advertising systems and its variations have motivated research related to their key tasks, such as, extracting keywords from textual sources, suggesting keywords to be included in an ad, matching keywords and ads, and methods to deternine when placing ads or not.

One of the first problems addressed was the task of matching keywords and ads. Ribeiro-Neto et al [Ribeiro-Neto et al. 2005] propose and investigate several match strategies. Their most successful method consisted in minimizing the mismatch between ad and page vocabularies by expanding the page vocabulary using the content of similar pages. Broder et al [Broder et al. 2007] addressed the same problem proposing the use of a topic classification as a ranking factor. In a follow up work, Broder et al [Broder et al. 2008] also proposed the expansion of keywords and ads, but this time, using additional knowledge derived from the result pages of a search engine. On [Karimzadehgan et al. 2011], the authors proposed a stochastic learning to rank algorithm using simulated annealing and show that the proposed approach has a good performance on ranking ads related to a given web page. Radlinski et al [Radlinski et al. 2008] focused on another aspect of the matching problem, that is, the selection of ads with the largest bids. Among the research efforts which focus on the problem of placing ads or not, we cite Broder et al [Broder et al. 2008], Yih and Meek [Yih and Meek 2008], and Lacerda et al [Lacerda et al. 2006].

Previous works have also addressed the problem of selecting keywords from textual content to make suggestions for bidders. Chen, Xue and Yu [Chen et al. 2008] have addressed the problem of suggesting keywords exploring the co-occurrence of the keywords selected by the bidders to determine new keywords for an advertiser. They also proposed the use of a taxonomy and strategies to deal with ambiguity. Also, Irmak et al [Irmak et al. 2009] studied the problem of using clickthrough to rank the keywords of a document. Although this problem is very closely related to the problem we try to solve on the present work, their main focus is different. While the goal of [Irmak et al. 2009] is to rank a list of pre-determined entities of a document according to their interestingness and relevance, our work aims to select from a full document and the best keywords to be used to select ads on an advertiser system. The work of Irmak et al cannot be included in our baselines also because we do not have a collection with clickthrough information.

The usage of external sources of information to help on the selection of keywords to describe a document has achieved good results as shown by the authors in [GM et al. 2011], which use a technique to enrich the content of the page using pages from the same website. The inclusion of this approach would not be feasible in our experiments, since the only collection we have is not composed of web sites, containing only a few web pages. In [Grineva et al. 2009], the Wikipedia is used to build a graph with the terms of a document and then discover what terms are more related to the main topic of the document. These two previous work are not included in our baseline, since our method does not take advantage of external sources of information, such as Wikipedia or from any other web pages, to select keywords, thus a comparison with them would not be fair.

Our work is based on previous ideas proposed by Goodman and Carvalho [Goodman and Carvalho 2005] and Yih et al [Yih et al. 2006] for determining keywords to be used to place ads in emails and web pages. In these approaches, the authors used logistic regression to learn good keywords for advertising. They studied a large number of features to determine the importance of a keyword. Among these features, we cite the frequency of each keyword candidate, its rareness in the target collection, the content section where it occurs (for instance, metadata section, title, etc) and its presence in search query logs. From their empirical study, they found that the presence of keywords in the query log was the most important feature to determine the importance of a keyword. They also concluded that other features, when taken into combination, were also useful to determine the best keywords. While the focus was to select keywords for advertising, both articles do not present experiments to evaluate the impact of the proposed methods in terms of precision and recall when retrieving ads.

The approach adopted by Wu and Bolivar [Wu and Bolivar 2008] is very similar to the one proposed by [Yih et al. 2006]. However, as their focus is to select good keywords for displaying ads from ebay[1], they take advantage of a set of features derived from proprietary data from this web site. These extra features are not available for experiments and, as our work is not focused on any particular website, we decided to use the work on [Yih et al. 2006] as our baseline.

Unlike previous approaches for learning ad keywords from web pages, we guide the learning process using an *ad-collection-aware* keyword selection approach. Besides this important difference, we also exploit features derived from the ad database that were not considered by Goodman and Carvalho [Goodman and Carvalho 2005] and neither by Yih et al [Yih et al. 2006]. Previous research articles indicate that such information is useful to improve the matching and ranking functions that associate ads to pages [Ribeiro-Neto et al. 2005; Chakrabarti et al. 2008].

[1] http://www.ebay.com

3. SELECTING KEYWORDS

Most of Contextual Advertising methods can be divided into two main steps: (i) discovering the context into the user is inserted. In this phase, keyword selection methods are usually used to discover what terms would describe better the content of the web page the user is browsing. (ii) recovering ads related to a given context. The information obtained on the first step is used to search on a set of ads for the ads that would be more interesting to the user. The algorithms used in this phase are usually based on matching of keywords (i.e.: VSM based algorithms).

The main focus of this work is to improve the quality of the method used to select the keywords on the first step of the process described above. We believe that such improvement will result on a better overall system quality. In this section, we model the keyword selection task as a classification problem, present the approaches we adopted for selecting keyword candidates and defining keyword relevance for advertising, as well as the features used to represent the keyword candidates.

3.1 Keyword Selection as a Classification Problem

As Yih et al [Yih et al. 2006], we address the problem of determining the advertising relevance of a keyword (sequence of terms) as a classification problem. Thus, let $K = \{k_1, k_2, ..., k_n\}$ be a set of keyword candidates. Each keyword k_i is represented by a set of m features $F = \{F_1, F_2, ..., F_m\}$, such that $k_i = (f_{i1}, f_{i2}, ..., f_{im})$ is a vector representing k_i, where each f_{ij} is the value of feature F_j in keyword k_i. Note the term *feature* describes a statistic that represents a measurement of some advertising relevance indicator associated with a keyword candidate.

We assume that we have access to some *training data* of the form

$$\{(k_1, r_1), (k_2, r_2), ..., (k_n, r_n)\} \subset K \times \{0, 1\}$$

where each pair (k_i, r_i) represents a keyword candidate and its corresponding relevance value, such that if $r_i = 1$, then the candidate k_i is a keyword. Otherwise, it is not a keyword.

Using this learning approach, the solution to this problem consists in: (a) determining the set of features $\{F_1, F_2, ..., F_m\}$ used to represent the keyword candidate in K; and (b) applying a classification method to find the best combination of the features to predict the relevance value r_i for any given keyword k_i.

To accomplish this, we use a logistic regression that, by means of the logistic function (see eq.(1)), computes the probability of a keyword being relevant for advertising as a function of the values of its relevance indicators.

$$f(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

Note that, in eq.(1), given a keyword k_i, $z = \beta_0 + \beta_1 f_{i1} + \beta_2 f_{i2} + ... + \beta_m f_{im}$. The values $\beta_0, \beta_1, ..., \beta_m$ are the regression coefficients which indicate how important are the relevance indicators f_{ij} to the probability of k_i being relevant.

Finally, we perform a regression for each class, setting the output equal to one for training instances that belong to the class and zero for those that do not. Then, given a candidate, we calculate the value of the logistic regression expression both for keywords and not keywords and choose the one that is largest. This value is also used to rank the keywords. We decided to use logistic regression in the classification task because it was also used by Yih et al [Yih et al. 2006], work that will be used as baseline for our experiments.

3.2 Definition of Keyword Candidates

In the two methods we study in this article, we use a keyword candidate definition that follows the same settings of the monolithic combined candidate selector described in [Yih et al. 2006] since this was found to be the best keyword selector in that work. More specifically, a keyword candidate is any word or phrase (consecutive words up to length 5) that appears in a page, in any of the sections: title, body, and meta-tag.

Phrases are not selected as candidates if they cross sentence or block boundaries. Further, phrases are taken as individual entities, such that features do not describe statistics about phrase constituent words but about the entire phrase. Note that no stemming normalization or stopword filtering was applied. In spite of that, phrases were not selected if they start or finish with stopwords.

3.3 Keyword Relevance

As previously mentioned, we assume that we have access to some training data of the form

$$\{(k_1, r_1), (k_2, r_2), ..., (k_n, r_n)\} \tag{2}$$

where each pair (k_i, r_i) represents a keyword candidate and its corresponding relevance value. To obtain the relevance values r we use two strategies. In the first one, human judges chose the keywords, as it is proposed in [Yih et al. 2006]. This first strategy is used here as a baseline. In second one, the candidates were taken as keywords according to their capability to trigger relevant ads. This second strategy is our new proposal to select relevant keywords.

In the first approach, we ask volunteers to label as keyword the words or phrases they judge relevant for advertising in a test collection. Volunteers are instructed to select keywords respecting the definition of keyword candidates presented in Section 3.2. Given a candidate k_i, it is considered as relevant ($r_i = 1$) if, at least, one user labels it as a keyword[2]. Otherwise, it is considered as irrelevant ($r_i = 0$). In this work, we referred to the keywords selected using this strategy as the baseline.

In the second approach, which is our proposal and we name as *ad-collection-aware* keyword selection (also referred as $ACAKS$) we retrieve the most similar ads for each keyword candidate in the training pages. We then ask users to evaluate the relevance of the ads for the page where the keyword candidate was extracted from. More specifically, for a given keyword k_i, five ads are retrieved according to their similarity to k_i. Candidate k_i is considered as relevant ($r_i = 1$) if at least one of these ads is considered as relevant for being presented in the page. Otherwise, k_i is considered as irrelevant ($r_i = 0$). As in the baseline, we also experimented with higher threshold values to determine the relevance, but again the best results were achieved with threshold 1.

Note that this new method is also based on learning from human evaluation of relevance, but requires a different kind of information. Based on an anecdotal analysis of our data, some factors which contribute for human judges selection errors are (a) their tendency to avoid keyword candidates in text fragments of peripheral importance, (b) their limited expertise on some page subjects, and (c) their general lack of knowledge about the ad database, in particular, regarding its vocabulary and advertising opportunities. We believe some of these problems are smoothed by the $ACAKS$ approach.

Further, a natural advantage of this new approach is that reference collections adopted for evaluating the performance of ad systems already contain the training information we require, since to create

[2] We also considered in preliminary experiments the possibility of using as thresholds the values 2 and 3, but these threshold resulted in worse quality when compared to threshold 1. While requiring more relevant judgments to consider an ad as relevant could improve the precision of the method, such constraint results on a small set of positive examples. With few examples, the learner is not able to build a good model which leads to low accuracy. On future work, we intend to increase the amount of pages used in the training to obtain a more accurate result.

such reference collections it is necessary to evaluate the relevance of ads given a web page. Reference collections are also available if the ad system uses any learn-to-advertise approach [Lacerda et al. 2006]. Thus, in practice, the change of focus in the selection of keywords may reduce the cost for training. Also, click-through information may be used as an approximation for human judgment relevance for ads since, for most of the keywords. This information is already available for companies that operate sponsored search systems.

One could argue that if no reference collection is available, the cost of training in a *ACAKS* keyword selection is higher. While this is not a practical situation, still the cost is not so high to avoid the application of the method, since a large number of keyword candidates do not have a match to the ad collection in practical systems, a phenomenon that is referred to in the literature as *impedance* between the web page vocabularies and the terms founds in the ad inventories [Ribeiro-Neto et al. 2005; Yih et al. 2006].

To retrieve the most similar ads to a keyword candidate we used the ADKW method, described on [Ribeiro-Neto et al. 2005]. This model was adopted in literature as a ranking method in contextual advertising [Ribeiro-Neto et al. 2005]. This model considers an ad as the concatenation of all the terms on the fields title, description and keywords of the ad and applies the Vector Space Model to rank the ads.

On the Vector Space Model, ads and keyword candidates are represented as vectors in a space composed of index terms. Thus, an ad a_j is represented as a vector of t term weights $a_j = (w_{1j}, w_{2j}, ..., w_{tj})$. Each w_{ij} weight reflects the importance of term k_i in an ad a_j and is computed as $w_{ij} = tf_{ij} \times log\frac{N}{ni}$, where tf_{ij} is the number of times the term k_i occurs in an ad a_j, n_i is the number of ads in which k_i occurs, and N is the total number of ads in the ad collection. Note that tf_{ij} is known as the *TF factor* whereas $log\frac{N}{ni}$ is known as the *IDF factor*. To calculate the similarity between a keyword candidate c and an ad a_j, we used the cosine value of the angle between c and a_j, as:

$$sim(a_j, c) = \frac{\sum_{i=1}^{t} w_{ij} \times w_{ic}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \times \sqrt{\sum_{i=1}^{t} w_{ic}^2}} \quad (3)$$

We selected this simple algorithm because it has already been used to rank ads in literature and our main focus here is to validate our keyword selection method.

3.4 Keyword Representation

In this section, we describe the features used to represent the keywords. These features are extracted from the textual content of the pages and query log. They were originally proposed and extensively studied in [Yih et al. 2006]. From the set of features used in that work, we have omitted the linguistic ones derived from the annotation obtained using a part-of-speech tagger. As observed by the authors in [Yih et al. 2006], linguistic features did not help in this domain, providing redundant information with other features, easier to calculate, such as capitalization and presence in query log.

The set of features are organized in several groups, as described as follows:

—**Capitalization:** whether the keyword is capitalized. The capitalization can indicate the keyword is part of a proper noun, or is an important word.

—**Hypertext:** whether a candidate phrase or word is part of the anchor text for a hypertext link.

—**Meta section features:** whether the candidate is part of the metadata section of the HTML document.

—**Title:** whether the candidate is part of the TITLE field.

—**Meta features:** whether the candidate is part of the meta description, meta-keywords or meta-title fields.

—**URL:** whether the candidate is part of the URL string.

—**Information retrieval features:** the TF (term frequency) and DF (document frequency) values of the candidate. The document frequency is the number of documents in the web page collection that contains the candidate. In addition to the original TF and DF, $log(TF + 1)$ and $log(DF + 1)$ are also used as features.

—**Relative location of the candidate:** the beginning of a document often contains an introduction with important words and phrases. Therefore, the location of the occurrence of the candidate is extracted as a feature. Since the length of a document varies considerably, we use the relative location by considering a normalized document length equal to 1. When the candidate is a phrase, its first word is used as its location. There are three different relative locations used as features: (a) $wordRatio$: the relative location of the candidate in the sentence; (b) $sentRatio$: the location of the sentence where the candidate is in divided by the total number of sentences in the document; (c) $wordDocRatio$: the relative location of the candidate in the document. In addition to these 3 features, we also use their logarithms as features. Specifically, we used $log(1 + wordRatio)$, $log(1 + sentRatio)$, and $log(1 + wordDocRatio)$.

—**Sentence and document length:** the length (in words) of the sentence ($sentLen$) where the candidate occurs, and the length of the whole document ($docLen$) (words in the header are not included) are used as features. Similarly, $log(1 + sentLen)$ and $log(1 + docLen)$ are also included.

—**Length of the candidate phrase:** the length of the candidate phrase ($phLen$) in words and $log(1 + phLen)$ are included as features.

—**Query log:** the query log of a search engine reflects the distribution of the keywords people are most interested in. We use the information to define three features: whether the phrase appears in the query log, the frequency with which it appears and the log value, $log(1 + frequency)$. In this work, we used the query log described in Section 4.1.

4. EXPERIMENTS

In this section we describe the datasets, the experimental methodology we used to conduct our empirical study and the results obtained.

4.1 Environmental Setup

To train and evaluate our ad placement framework, we used a test collection built from a set of 300 pages extracted from a Brazilian newspaper. As we have no preference for particular topics, these pages cover diverse subjects, such as culture, local news, international news, economy, sports, politics, agriculture, cars, children, computers and Internet, among others.

The IDF information we have used was obtained from a commercial search engine. We submitted each keyword candidate selected from the pages in the experiments (referred also as $kw_{candidate}$) as a query to the search engine and the number of documents retrieved was considered as the DF (Document Frequency). We then computed the IDF as $log(\frac{N}{DF(kw_{candidate})})$, where N is the total number of documents found in the search engine collection. As no search engine provides this information explicitly we estimated it by searching for some stop words, like "a" and "the", and then considering the highest number of results obtained as the value of N.

The ads used in our experiments were obtained from a real case ad collection composed of $93,972$ ads grouped in $2,029$ campaigns provided by $1,744$ advertisers. With these ads, advertisers associated a total of $68,238$ keywords[3]. In this collection, only one keyword is associated with each ad.

[3] Data in Portuguese language provided by an on-line ad company that operates in Brazil.

We need to obtain reference sets containing information about what keywords are useful to represent web pages and also containing the ads that are relevant to be placed in each of the 300 web pages of the collection. These two information will be used in the training and test phases of the experimented keyword selection approaches.

To train and test the baseline method, we need to construct a set of keywords that should be manually labeled by users. To obtain such training we present each of the 300 pages of the test collection to volunteers (60 volunteers contributed to all phases of our experiments), asking them to select keywords from these pages following the keyword candidate guidelines described in Section 3.2, and considering that the purpose of the keyword selection is to associate pages with relevant ads.

Note that the decision about the selection of keywords depends exclusively on the judgment of the volunteers. The result of this process is a set of keywords associated with each of the 300 web pages of the collection. Given a page p, we name this set as the *human tagged keywords* of p, denoted as $(HT_{Kw}(p))$. In the experiments, users tagged an average of 14.33 keywords per page. From these keywords, very few were labeled for more than one user. More specifically, 13.88% were labeled by two or more volunteers and 3.28% were labeled by the three volunteers. Table I summarizes the information about the training data used on the baseline approach.

baseline
300 pages 60 volunteers 14.33 keywords/page (average)

Table I. Training details of baseline approach.

The reference collection for *ACAKS* method also requires a set of relevant ads related to each of the 300 pages. To obtain this set, we extracted keyword candidates. As the textual content of some web pages is very large (1000 or more words), we take into account only the first 400 words in each page. Such constraint does not affect most of web pages and reduce the number of keyword candidates considered for these very large pages. As we consider that eliminating or rising this threshold could improve even more the results obtained by our method, we intend to study the impact of using higher values as future work. The average number of *keyword candidates* per page we found by using this approach was 279.76. The result of this process is a set of *keyword candidates* associated with each of the 300 web pages of the collection. Given a page p, we name this set as the *keyword candidates set* of p, denoted as $(C_{kw}(p))$.

We submit each *keyword candidate* as a query to the indexed collection of ads and take the top five answer results, using the *ADKW* [Ribeiro-Neto et al. 2005] as the ranking method. For each page p, we create a set of ads AD_P composed by the union of the answer results obtained from all the *keyword candidates* found in $C_{kw}(p)$. As a result of the above process, we selected a total of 95,327 distinct pairs of ads and pages corresponding to an average of roughly 317 ads per page.

Then, for each pair (p,a), p being a page and $a \in AD_P$, three human volunteers judged whether a is relevant to p or not. Thus, note that we have $285,981$ evaluation of pairs. To reduce the costs of this phase, each volunteer evaluated a set of 15 pages and the ads associated with them. Using this strategy, each volunteer spent at most three hours labeling ads as relevant or not. Note that this effort would not be necessary if clickthrough information were available for the ad collection adopted. It could be also avoided if we had obtained a reference collection with the complete set of relevant ads for each page.

We consider as relevant to p an ad labeled as relevant by at least one volunteer. Only 20.90% of the relevant ads were labeled as relevant by two or more volunteers and 6.88% of them, labeled as relevant by the three volunteers. Finally, a *keyword candidate* is considered relevant to p if at least one of the ads retrieved by it is relevant. The average number of relevant ads per page obtained with this process was 41.83, and the number of relevant keywords per page obtained in the reference set according to the ACAKS method was 21.35.

The result of this process is also a set of *keywords* associated with each of the 300 web pages of the collection. Given a page p, we name this set as the *score tagged keywords* of p, denoted as $(ST_{kw}(p))$. Table II summarizes the information about the training data used on $ACAKS$ approach.

ACAKS
300 pages
279.76 keyword candidates/page (average)
317 ads/page (average)
Total of 95,327 pairs (ad,page) evaluated
41.83 relevant ads/page (average)
21.35 keywords/page (average)

Table II. Training details of ACAKS approach.

The query log features used on this work were derived from the query log of the WBR03 collection, a database extracted from the Brazilian web which contains queries submitted to TodoBR[4], a real case search engine. The log consists of $12,795,101$ queries and $2,987,745$ distinct queries.

4.2 Evaluation Methodology

To perform the experiments, we used the 10-fold cross validation method [Mitchell 1997]. All the results reported are average values of the 10-fold runs and for all comparisons reported in this work, we used the Student's t-test [Fisher 1925] for determining if the difference in performance was statistically meaningful. We consider statistically meaningful results with a p-value ≤ 0.01. We assessed the performance of each keyword selection method proposed through four distinct experiments, described in the following paragraphs.

First, we measured the quality of our keyword classifier using the accuracy measure, which is defined as the proportion of correctly classified examples for this purpose. Although this experiment is interesting to measure the quality of the classifier on each approach, it is important to note that the main goal of this work is to select keywords to improve the quality of advertising systems.

In the second experiment, we evaluated the quality of ads retrieved by the keywords obtained by each approach. The set of keywords returned by each method was used as a query submitted to a system which returned a ranking of ads based on the $ADKW$ method [Ribeiro-Neto et al. 2005]. This experiment aimed at asserting the impact of the keyword selection strategies studied, when used in an ad selection system. The metric adopted was precision and recall considering the top 3 ads retrieved by each method. We also adopt the $pavg$@3 as described on [Lacerda et al. 2006].

We calculated Precision at 3 (referred also as p@3) as:

$$p@3 = \frac{|rel \cap answers|}{|answers|} \tag{4}$$

where rel is the set of relevant ads associated with the web page in the pool and $answers$ is the set of ads displayed to this page by the evaluated method. Note that, as we consider only the top 3 results for each page, the maximum value of $|answers|$ is 3. Indeed, in some cases, the system retrieves less then 3 ads for the set of keywords used as query.

A problem found when computing p@3 in our experiments is that it is common to find cases where a method do not provide answers. This is a quite common problem in real case advertising collections,

[4] TodoBR is a trademark of Akwan Information Technologies, which was acquired by Google in July 2005.

since they may not cover the whole set of keywords and topics found on the web. In these cases, the precision for such specific queries cannot be determined. To cope with that, we could define $p@3$ as 1 or 0 if no answer is provided, however, we think such a definition does not reflect the real precision. We then calculate $p@3$ as an average over only the pages where at least one ad was returned by the method. However, by adopting only that strategy we do not provide full information about quality. For instance, if method A returns just one ad for one page (out of the 300 pages) and this ad is relevant, $p@3$ for A would be 100%. Displaying only such result would hide the fact that the method could not associate any ad (relevant or not) with the other 299 documents.

Thus, to provide more insight about the performance of each approach, we also introduced the computation of Recall at 3 (referred also as $r@3$). Metric $r@3$ is calculated taking the number of relevant ads found in the pool as the set of relevant answers, but limiting this number to the maximum of relevant ads that could be shown by each method, as follows.

The $r@3$ of a method given a page p is:

$$r@3 = \frac{|rel \cap answers|}{min(|rel|, 3)} \tag{5}$$

where rel and answer are defined as in the previous equation and min is a function that returns the minimum value of two arguments.

The third experiment shows the performance of each approach with training sets of different sizes. The objective of such experiment is to discover the behavior of each method while increasing and decreasing the size of the training set.

Note that both the second and third experiments can bring new pairs of ad and pages not evaluated in our initial pool. Thus, a second round of evaluation was required to complete the set of relevant ads associated with each page. In this second round we found an average number of 1.34 ads not evaluated per page. After evaluating then, we found an average of 0.65 extra relevant ads per page.

Finally, the fourth experiment includes ad collection features in the learning process in order to check if their inclusion changes the comparative results between the ACAKS and the baseline approaches.

4.3 Experimental Results

This section presents the results of experiments we conducted to evaluate our proposed methods and compare them to baselines found in literature.

In all tables of this section, we present the results obtained while using *ACAKS* and baseline approaches and the performance of *IDEAL-baseline* and *IDEAL-ACAKS*. We refer as *IDEAL-baseline* the method that classifies as keywords exactly the ones (and only them) labeled as such by the human evaluators, i.e., the ones present in $HT_kw(p)$, for each p in the test set. In the same way, we refer as *IDEAL-ACAKS* to the method which classifies as keywords exactly the ones (and only them) taken as relevant by the *ACAKS* approach, i.e., the ones found in the set $ST_{kw}(p)$, for each page p in the test set.

Our first experiment aims at evaluating the quality of the keyword selection methods, taking the human judgments as our gold standard. Note that in this scenario, it is expected that the baseline approach reaches better accuracy than the *ACAKS* one, since in the first approach the keywords were learned taken as training set *exactly* the gold standard, whereas in the second approach, the keywords were selected based on their performance on triggering ads. In fact, the resulting training sets are quite different. Out from 10,444 keywords provided to train the methods, only about 10% occurred on both training sets. Table III depicts the accuracy results of the studied keyword detection approaches. Note that the results presented in this Table consider as relevant keywords of a given page p only the human tagged keywords, thus $HT_{kw}(p)$.

Method	Accuracy
baseline	29.16%
ACAKS	29.16%
IDEAL-baseline	100%
IDEAL-ACAKS	22.41%

Table III. Accuracy of each method in the task of selecting good keywords, the keywords in the test sets were labeled by humans.

While there was little intersection between the keywords labeled as relevant in the training sets used by the *ACAKS* and the baseline approaches, the methods achieved the same performance. After a careful inspection of the keywords used for training and the ones selected by the methods in the test, we noticed twice as many keywords in common in the test than in the training. However, the difference in the output of the methods is still large. For instance, from the total of keywords selected by both the baseline and *ACAKS* methods, about 80% were different. Among the examples of keywords selected only by the baseline approach, we cite "rig" and "whiz kid". These keywords, in general, are not found at all or have little importance in the ad collection. On the other hand, examples of keywords selected only by the *ad-collection-aware* approach are "advising", "team", and "work of art". These keywords normally have peripheral importance in the pages. As we show in the next experiment, many of these candidates not selected by baseline, but caught by *ACAKS*, have triggered interesting ads.

The performance presented by the approaches indicates that the *ACAKS* approach may be used as a more general annotation method to find keywords of web pages. While this is not the focus here, we plan to study this possibility as a future work.

A second important point to observe from Table III is that *IDEAL-ACAKS*, a classification method that would take exactly the correct keywords according to the *ACAKS* approach, would in fact result in a classifier with less relevant keywords according to the human evaluation performed. However, as we show in the next experiment, such result does not necessarily imply a worse keyword selection for the ad placement system. This result reinforces our initial intuition that a *ACAKS* keyword selection approach may be better than the baseline approach in ad placement systems.

Although the results presented in Table III are important to reinforce our initial intuition and better understand the behavior of both methods, the main objective of the proposed approach is to select keywords that are useful on the task of retrieving relevant ads. The objective of the following experiment is to assess the performance of each method on such scenario. Table IV shows the average precision and recall at the top three results retrieved by each method. Note we consider as relevant ads chosen by, at least, one user.

Method	p@3	r@3	pavg@3
baseline	0.4478	0.1933	0.3680
ACAKS	0.4774	0.3133	0.4100
IDEAL-baseline	0.5872	0.5833	0.4620
IDEAL-ACAKS	0.7678	0.7678	0.7280

Table IV. p@3, r@3 and pavg@3 for each method. An ad is considered as relevant to a page if at least one user label it as relevant.

We first note in Table IV that the results obtained by the *ACAKS* approach were better than those achieved by the baseline, confirming our assumption that our approach is quite better on retrieving relevant ads. The precision achieved by both methods was quite close (0.4478 for the baseline and 0.4774 for *ACAKS*) and the difference between them in terms of p@3 were not statistically meaningful. Thus, we can conclude that in terms of p@3, both methods are equivalent.

A similar behavior is achieved when considering the $pavg$@3 results. Considering this metric, the results obtained by the *ACAKS* approach were 11% better than the baseline. Although, the statistical test shows that this difference is not significant. Thus, we can conclude that also in this scenario, the methods are equivalents.

When considering the recall, the *ACAKS* approach improves the result obtained by the baseline by more than 62%. It indicates that the proposed approach is able to select a higher number of ads when compared with the baseline, achieving this improvement without dropping precision. These results show the importance of choosing keywords driven by the quality of the ads they will retrieve and not by only what humans believe to be good keywords.

By using the *ACAKS* approach, 593 ads were displayed, for a total of 267 relevant ads. From the total of pages, 133 have received at least one relevant ad. By using the baseline, 339 ads were displayed for a total of 157 relevant ads. Only 86 pages have received at least one relevant ad.

We can also observe in Table IV the precision results obtained by the perfect versions of the baseline and *ACAKS*. If such classifiers could be used, we would obtain a *ACAKS* result more than 60% better in terms of $p@3$ and 145% better in terms of $r@3$. Similarly, we would improve the baseline results by more than 30% in $p@3$ and more than 200% in $r@3$. Such findings indicate we have much room for improvements by enhancing the accuracy of our automatic classifiers. Further, the gain obtained by an *IDEAL-ACAKS* when compared to an *IDEAL-baseline* would be 31.73%. Such results indicate that the *ACAKS* approach is a fair better alternative strategy for extracting keywords in ad selection systems.

A possible reason for the best performance of the *ACAKS* method is the fact that the intersection between the keywords selected by this method and the ads vocabulary is high. While only 67.82% of the keywords selected by the baseline approach were found on at least one advertisement, this number rises to 97.10% while considering the *ACAKS* approach.

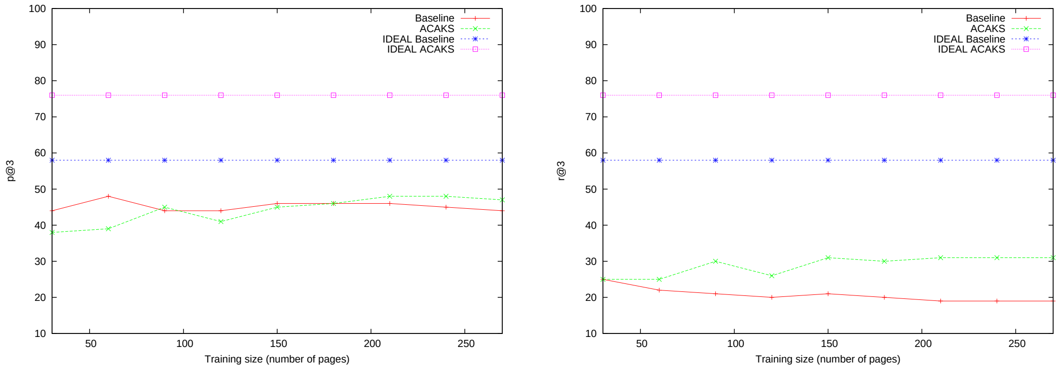
Another important aspect to be considered is the impact of the size of the training sets on the results obtained by each approach. The *ACAKS* approach relies on judgments about the ads related to each *keyword candidate*, so the number of ads associated with a page to be evaluated is quite high, while the effort to label keywords in the baseline approach tends to be smaller, since users need only to label the keywords in the training pages. As the effort to produce both training collections is quite different, one could argue that this is the cause of the difference in precision of the results obtained by the *ACAKS* and baseline approaches.

Despite the fact that the training set of the *ACAKS* approach can be obtained using clickthrough information or even a reference collection, we evaluated the quality of the results obtained by each approach with training sets of different sizes. Our goal with this final experiment is to measure the importance of the size of training set on the final results obtained by each method.

Figure 1(a) shows the $p@3$ value of each approach with training sets of different sizes. As it can be seen, the performance of both approaches did not increase in the experiments for training sets with more than 150 pages. These results indicate that extra efforts to increase the training set might not be worth.

Figure 1(b) shows the $r@3$ value of each approach with training sets of different sizes. Both methods do not show improvements on $r@3$ value using training sets with more than 150 pages. In this case, the performance of the baseline is quite worse than the performance obtained by the *ACAKS* approach. As a conclusion, we can say that the adoption of *ACAKS* represents an important practical advantage to an ad selection system, since gains in recall may also represent an increasing in revenue that certainly justifies the extra effort required to train.

Both the methods did not take any advantage of using more than 150 pages on the training set in any of the metrics adopted. Also, besides having a higher cost of training, *ACAKS* approach outperformed the baseline score in terms of $r@3$. As the $r@3$ seems to stabilize with training sets with more than 150 pages, the results indicate that even increasing the training set to more than 270



(a) p@3 values for training sets of different sizes.

(b) r@3 values for training sets of different sizes.

pages, the baseline approach would not be able to outperform *ACAKS*.

4.3.1 *Including Ad Collection Features.* Previous work have shown that using statistics about the document collection could improve the quality of the ranking [Veloso et al. 2008] produced while using machine learning strategies. Also, the work on [Lacerda et al. 2006] presented good results with a set of features obtained using data from the advertising collection. As the *ACAKS* method proposed on this paper obtained good results on the task of selecting keywords using information about the ad collection, we decided to study whether the inclusion of an extra set of features derived from the Ad Collection would improve the results of both *ACAKS* and the baseline or not.

Ad collection features are extracted from the advertising database. For this extra set of features, we consider that an ad is composed of three structural parts: a title, a textual description and a set of keywords. In fact, these are the usual components of an ad in search advertising systems and comprise which is called the *ad creative*. Further, an advertiser can associate several ads with the same product or service. We refer to such group of ads as a *campaign*. Note that only an ad per campaign should be placed in a web page in order to ensure a fair use of the page advertising space and increase the likelihood that the user will find an interesting ad.

Figure 1 illustrates an ad list with three ad slots on the right side of a web page. For the ad in the first ad slot, the title is "Accommodation Cape Town", the description is "Luxury Apartments in Cape Town. Minutes To Main Venue. Enquire Now.", and the hyperlink points to the site "www.SoccerWorldCup2010s.com".

We adopted the following features to be extracted from the ad collection:

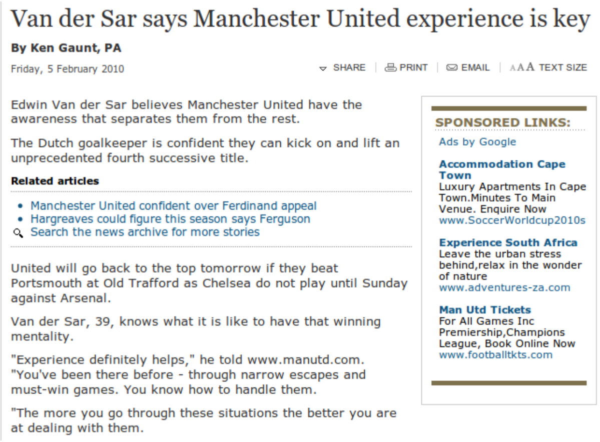
—**Ad Section TF:** candidate frequency in each of the structural sections of an ad creative. Since an ad has three sections, we use three features to represent them: Ad title TF, Ad description TF, and Ad keyword TF.

—**Ad Section Max-TF:** maximum candidate frequency in each of the structural sections of an ad creative. As for *Ad Section TF*, we then have: Ad title Max-TF, Ad description Max-TF, and Ad keyword Max-TF.

—**Ad Section Avg-TF:** average candidate frequency in the three sections of an ad creative: Ad title Avg-TF, Ad description Avg-TF, and Ad keyword Avg-TF.

—**Ad Section DF:** number of ads in which candidate occurs in a certain section of an ad creative. The three features are in this case: Ad title DF, Ad description DF, and Ad keyword DF.

—**Campaign Section Max-TF:** maximum candidate frequency in the structural sections of all the ads of a campaign. The three features in this case are: Campaign title Max-Tf, Campaign description Max-FT, and Campaign keyword Max-TF.



Van der Sar says Manchester United experience is key

By Ken Gaunt, PA

Friday, 5 February 2010

Edwin Van der Sar believes Manchester United have the awareness that separates them from the rest.

The Dutch goalkeeper is confident they can kick on and lift an unprecedented fourth successive title.

Related articles

- Manchester United confident over Ferdinand appeal
- Hargreaves could figure this season says Ferguson

Search the news archive for more stories

United will go back to the top tomorrow if they beat Portsmouth at Old Trafford as Chelsea do not play until Sunday against Arsenal.

Van der Sar, 39, knows what it is like to have that winning mentality.

"Experience definitely helps," he told www.manutd.com. "You've been there before - through narrow escapes and must-win games. You know how to handle them.

"The more you go through these situations the better you are at dealing with them.

SPONSORED LINKS:

Ads by Google

Accommodation Cape Town
Luxury Apartments In Cape Town.Minutes To Main Venue. Enquire Now
www.SoccerWorldcup2010s

Experience South Africa
Leave the urban stress behind,relax in the wonder of nature
www.adventures-za.com

Man Utd Tickets
For All Games Inc Premiership,Champions League, Book Online Now
www.footballtkts.com

Fig. 1. Example of contextual advertising in the page of an England newspaper that offers tickets to soccer games, accommodations to the World Cup and tourism in South Africa, where the next world cup will take place. The content of the page is about soccer.

—**Campaign Section Avg-TF:** average candidate frequency in each of the structural sections of all the ads of a campaign. The three features in this case are: Campaign title Avg-TF, Campaign description Avg-TF, and Campaign keyword Avg-TF.

—**Campaign Section DF:** number of campaigns in which candidate occurs in a certain section of an ad creative. The three features in this case are: Campaign title DF , Campaign description DF, and Campaign keyword DF.

These features were chosen given the success of frequency of terms in objects (**TF**) and of the number of objects where a term occurs (**DF**) as features in previous work related to information retrieval tasks [Lacerda et al. 2006; Veloso et al. 2008].

Method	p@3	r@3
baseline P.L.	0.4478	0.1933
ACAKS P.L.	0.4774	0.3133
baseline P.L.+A.C.	0.4463	0.1967
ACAKS P.L.+A.C.	0.4791	0.3144

Table V. p@3 and r@3 for each method. An ad is considered as relevant to a page if at least one user label it as relevant.

Table V presents a comparison between the results obtained by the *ACAKS* and baseline methods using only the Page and Log features proposed on [Yih et al. 2006] (referred on this table as baseline P.L. and *ACAKS* P.L.) and the results obtained by the methods using the Page and Log features and the Ad Collection features described above (referred on this table as baseline P.L.+A.C. and *ACAKS* P.L.+A.C.). The usage of Ad Collection features resulted in no gain for both methods. The difference in the results is not statistically significant. As a conclusion of this final experiment, we can say that the *ACAKS* approach is superior to the baseline even when the ad collection features are taken into account.

5. CONCLUSIONS AND FUTURE WORK

In this work we have proposed a new approach for selecting keywords from web pages in contextual advertising systems. Our main contribution was a change in the strategy to compose the training collection to guide the learning process. Instead of asking users to directly giving examples of what are the good keywords found in the training pages, we checked which ads have a match with each keyword candidate found in the training pages, and asked the users to evaluate the relevance of the ads that would be associated with these keywords. We found this strategy provide quite competitive results when compared to a previous method proposed recently in literature.

The new approach proposed led to significant gains over the baseline, with gains of 62% in $r@3$ when considering just the features proposed by [Yih et al. 2006]. Further, our experiments indicate that even when increasing the size of the training in the baseline approach, still the $ACAKS$ presents superior results, which brings the conclusion that the $ACAKS$ approach is a viable and attractive alternative for keyword selection in ad placement systems.

As future work, we intend to expand our research in order to contemplate additional evidence and other contexts, such as video. As another future work, we intend to study the performance of other machine learning methods aiming to obtain results closer to the ideal one described here. We will particularly investigate the performance of SVM [Joachims 1998] as the classification method adopted to select keywords. Also, we intend to apply the method proposed on [Dave and Varma 2010] to drop the number of keyword candidates to be considered in each page and thus reducing also the number of ads to be evaluate in order to create the training set.

6. ACKNOWLEDGEMENTS

This research was sponsored by UOL (www.uol.com.br), through its UOL Bolsa Pesquisa program, process number 20110215215200 and partially supported by projects INWeb (MCT/CNPq grant 57.3871/2008-6), SABACO (CNPq grant575541/2008-3), Universal (CNPq grant 474099/2007-5) and by individual CNPq fellowship grant to Edleno S. de Moura (307861/2010-4).

REFERENCES

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

Baeza-Yates, R. and Ribeiro-Neto, B. *Modern Information Retrieval.* Addison-Wesley-Longman, 1999.

Broder, A., Ciaramita, M., Fontoura, M., Gabrilovich, E., Josifovski, V., Metzler, D., Murdock, V., and Plachouras, V. To swing or not to swing: learning when (not) to advertise. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management.* ACM, New York, NY, USA, pp. 1003–1012, 2008.

Broder, A., Fontoura, M., Josifovski, V., and Riedel, L. A semantic approach to contextual advertising. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, New York, NY, USA, pp. 559–566, 2007.

Broder, A. Z., Ciccolo, P., Fontoura, M., Gabrilovich, E., Josifovski, V., and Riedel, L. Search advertising using web relevance feedback. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management.* ACM, New York, NY, USA, pp. 1013–1022, 2008.

Chakrabarti, D., Agarwal, D., and Josifovski, V. Contextual advertising by combining relevance with click feedback. In *Proceeding of the 17th international conference on World Wide Web.* WWW '08. ACM, New York, NY, USA, pp. 417–426, 2008.

Chen, Y., Xue, G.-R., and Yu, Y. Advertising keyword suggestion based on concept hierarchy. In *WSDM '08: Proceedings of the international conference on Web search and web data mining.* ACM, New York, NY, USA, pp. 251–260, 2008.

Dave, K. S. and Varma, V. Pattern based keyword extraction for contextual advertising. In *Proceedings of the 19th ACM international conference on Information and knowledge management.* CIKM '10. ACM, New York, NY, USA, pp. 1885–1888, 2010.

Fisher, R. A. Applications of "student's" distribution. *Metron* vol. 5, pp. 90–104, 1925.

GM, P. K., Leela, K. P., Parsana, M., and Garg, S. Learning website hierarchies for keyword enrichment in contextual advertising. In *Proceedings of the fourth ACM international conference on Web search and data mining.* WSDM '11. ACM, New York, NY, USA, pp. 425–434, 2011.

Goodman, J. and Carvalho, V. R. Implicit queries for email. In *Second Conference on Email and Anti-Spam.* `http://www.ceas.cc/papers-2005/141.pdf`, 2005.

Grineva, M., Grinev, M., and Lizorkin, D. Extracting key terms from noisy and multitheme documents. In *Proceedings of the 18th international conference on World wide web*. WWW '09. ACM, New York, NY, USA, pp. 661–670, 2009.

Irmak, U., von Brzeski, V., and Kraft, R. Contextual ranking of keywords using click data. In *ICDE '09: Proceedings of the 2009 IEEE International Conference on Data Engineering*. IEEE Computer Society, Washington, DC, USA, pp. 457–468, 2009.

Joachims, T. Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, 1998.

Karimzadehgan, M., Li, W., Zhang, R., and Mao, J. A stochastic learning-to-rank algorithm and its application to contextual advertising. In *Proceedings of the 20th international conference on World wide web*. WWW '11. ACM, New York, NY, USA, pp. 377–386, 2011.

Lacerda, A., Cristo, M., Gonçalves, M. A., Fan, W., Ziviani, N., and Ribeiro-Neto, B. Learning to advertise. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, NY, USA, pp. 549–556, 2006.

Mitchell, T. *Machile Learning*. McGraw-Hill, 1997.

Radlinski, F., Broder, A., Ciccolo, P., Gabrilovich, E., Josifovski, V., and Riedel, L. Optimizing relevance and revenue in ad search: a query substitution approach. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, NY, USA, pp. 403–410, 2008.

Ribeiro-Neto, B., Cristo, M., Golgher, P. B., and Silva de Moura, E. Impedance coupling in content-targeted advertising. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, NY, USA, pp. 496–503, 2005.

Salton, G., Wong, A., and Yang, C. S. A vector space model for automatic indexing. Tech. rep., Cornell University, Ithaca, NY, USA, 1974.

Veloso, A. A., Almeida, H. M., Goncalves, M. A., and Jr., W. M. Learning to rank at query-time using association rules. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, NY, USA, pp. 267–274, 2008.

Wu, X. and Bolivar, A. Keyword extraction for contextual advertisement. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*. ACM, New York, NY, USA, pp. 1195–1196, 2008.

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

Yih, W.-t. and Meek, C. Consistent phrase relevance measures. In *ADKDD '08: Proceedings of the 2nd International Workshop on Data Mining and Audience Intelligence for Advertising*. ACM, New York, NY, USA, pp. 37–44, 2008.