Sentiment-Oriented Contextual Advertising

Teng-Kai Fan
Department of Computer Science
National Central University
No. 300, Jung-Da Rd., Chung-Li,
Tao-Yuan, Taiwan 320, R.O.C.
tengkaifan@gmail.com

Chia-Hui Chang
Department of Computer Science
National Central University
No. 300, Jung-Da Rd., Chung-Li,
Tao-Yuan, Taiwan 320, R.O.C.
chia@csie.ncu.edu.tw

Abstract Web advertising (Online advertising), a form of advertising that uses the World Wide Web to attract customers, has become one of the world's most important marketing channels. This paper addresses the mechanism of *Content-based advertising* (*Contextual advertising*), which refers to the assignment of relevant ads to a generic web page, e.g. a blog post. As blogs become a platform for expressing personal opinion, they naturally contain various kinds of expressions, including both facts and comments of both a positive and negative nature. Besides, in line with the major tenet of Web 2.0 (i.e., *user-centric*), we believe that the web-site owners would be willing to be in charge of the ads which are positively related to their contents. Hence, in this paper, we propose the utilization of sentiment detection to improve Web-based contextual advertising. The proposed **SOCA** (**S**entiment-**O**riented **C**ontextual **A**dvertising) framework aims to combine contextual advertising matching with sentiment analysis to select ads that are related to the positive (and neutral) aspects of a blog and rank them according to their relevance. We experimentally validate our approach using a set of data that includes both real ads and actual blog pages. The results indicate that our proposed method can effectively identify those ads that are positively correlated with the given blog pages.

Keywords *Web Advertising, Sentiment Detection, Marketing, Machine Learning.*

1 INTRODUCTION

Due to its rapid growth and popularization, the World Wide Web has become one of the most essential media channels for advertising. According to the Interactive Advertising Bureau (IAB)[1], Internet advertising revenues exceeded 5.2 billion USD for the third quarter of 2007, representing yet another historic quarterly record and a 25.3 percent gain over Q3 2006. The statistics clearly show that advertisers are increasingly using the Web since it not only produces consuming interaction, but is also highly flexible in terms of both geography and time. Furthermore, multiple forms of Web advertising are available, including plain text, video, and e-mail (including spam). Generally, for text-based ads, there are two main categories [1] [5]:

Sponsored Search (*Paid Placement Advertising* or *Keyword Targeted Marketing*): an advertiser bids a reasonable price using certain keywords or phrases in order to appear at a certain position in lists of advertisers. Then, a list of ranked ads are triggered by user's search keyword (query) and placed on the result pages from a search engine [9] [38].

Content-based Advertising (*Content-Targeted Advertising* or *Contextual Advertising*): Information Service Providers (ISP) (e.g., Google, Yahoo, and Microsoft) often support another textual ad format known as Content-based Advertising. This uses an intermediate ad matching system that parses the content of a page and returns those ads that are most relevant to currently-viewed pages, either through ad placements or pop-ups. For instance, if a user is browsing a web page about mobile phones, ads related to the content of this page may be selected by Google's Content-based advertising system (Google AdSense[2]). These can be arbitrarily displayed in predefined positions, as shown in Figure 1.

Received: Oct 26, 2008
Revised: Apr 24, 2009
Accepted: May 09, 2009
[1] http://www.iab.net/
[2] https://www.google.com/adsense/

Figure 1: Correlation ads conflicting with blog content.

The techniques differ for these two approaches, in that sponsored search analyzes only the user's query keywords while Content-based advertising parses the contents of a web page to decide which ads to show. However, the goals of each approach are identical. The intent is to create a triple-win commercial platform. In other words, an advertiser pays a low price to purchase valuable advertisements, the ad agency system shares advertising profits with the web site owner (publisher), and consumers can easily respond to ads to purchase products or services.

Some prior studies have suggested that strong relevance increases the number of click-through [7] [22] [36]. However, existing Content-based advertising strategies only attempt to match relevant ads to a given web page, which we refer to through the term *correlation*, but they neglect to distinguish *positive* from *negative* correlations between ads and the pages on which they are placed. For example, as shown in Figure 1, an ad agency system might place the three most relevant ads about *mobile phone* on the top of a web page. However, the content of this page may describe reasons why someone should consider not using a mobile phone. We hypothesize that such ads that conflict with the negative orientation of the page are less likely to trigger click-through. In other words, even if an ad is related to the content of a triggering page, the ad agency system should avoid placing ads on pages which discuss the product/service from a negative point of view. Moreover, it is likely that blog owners themselves may not be happy about advertisements that directly conflict with their opinions.

Consistent with the Web 2.0 paradigm (e.g. *user-centric*) [11] [20], we focused our dataset on the blogosphere. As well as publishing opinions, news, comments, diaries, and personal photos, people can design blog pages to appear any way they please. Hence, in this paper, we proposed an ad matching mechanism, which we refer to as Sentiment-Oriented Contextual Advertising (**SOCA**), based on sentiment detection to associate ads with blog pages. Instead of traditional placement of relevant ads, **SOCA** emphasizes that the ad agency's system should provide relevant ads that are related to the positive (and neutral) aspects of the page in order to best attract consumers. To evaluate our proposed method, we used a real-word collection comprising ads and blog pages respectively from Google AdSense and Google's Blog-Search Engine[3]. Our results suggest that our proposed method can effectively match relevant ads to a given blog page.

The rest of this paper is organized as follows: Section 2 provides background information on current on-line advertising and sentiment detection. Section 3 introduces our methodology. The experimental evaluations are presented in Section 4. Section 5 outlines some related work. Finally, we present conclusions and future directions in Section 6.

2 BACKGROUND

In this section, we briefly describe current on-line Content-based advertising and review the main concepts in opinion detection and sentiment classification.

[3] http://blogsearch.google.com

2.1 Content-based Advertising

Content-based advertising involves an interaction between four players [1] [5], as shown in Figure 2. The publisher, an owner of a web-site, usually provides interesting pages on which ads are shown. The publishers typically aim to engage a viewer, encouraging them to stay on their web page and, furthermore, attracting sponsors to place their ads on the page. The advertiser supplies a series of ads to market or promote their products. Generally, the activity of the advertisers is organized around campaigns, which are defined by a set of ads and a theme over a particular period. In addition to traditional media (e.g., TV, magazines and direct mail), the advertisers register certain characteristic keywords to describe their products or services. The ad agency system is a mediator between the advertisers and the publishers, i.e. it is in charge of matching ads to pages. End Users not only browse the contents of a web page, but interact with the ads to engage in commercial activities.

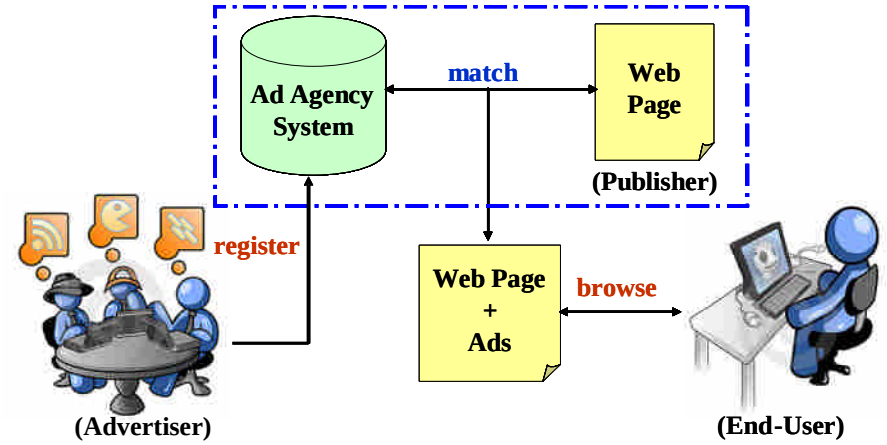

Figure 2: Key players in Content-based advertising.

Without loss of generality, in Content-based advertising, an ad generally features a *title*, a text-based *abstract*, and a *hyperlink*. For example, as shown in Figure 1, the first ad is selected by Google, the title is "AT&T Wireless," usually depicted in bold or a colorful font, the abstract is "Large Selection of Free Cell Phones visit AT&T Wireless Official Site." The latter is generally concise due to space limitations. The hyperlink is "www.att.com/wireless," which links to an ad web page, known as the *landing page*.

Evaluating Content-based advertising is different from assessing traditional media. A major advantage is that it is relatively easy to measure user response. In general, a functional estimation is to calculate the number of click-through or judge whether a user's activity is consistent with the relevant ad guidelines. Three common pricing models through which online advertising is purchased are **CPM**, **CPA** and **CPC**. **CPM** (Cost Per Thousand Impressions) applies when advertisers pay for exposure of their message to a specific audience. **CPA** (Cost Per Action) is based on the user completing a certain set of actions or placing an order. The most widespread model for contextual ads is **CPC** (Cost Per Click), also known as **PPC** (Pay Per Click) [9]. Advertisers pay every time a user clicks on their ads and is redirected to their web site. They do not actually pay for the ads, but instead they pay only when the ads are clicked. This approach allows advertisers to refine search keywords and gain information about their market. Generally, user's clicks generate profits for the publishers and the ad agency system. Advertisers receive access to a network channel between the end-user and the target web-site. A number of studies have suggested that strong relevance definitely increases the number of ad clicks [7] [22] [36]. Hence, in this study, we similarly assume that the probability of a click for a given ad and page is determined by the ad's relevance score with respect to the page. For simplicity, we ignore the positional effect of ad placement and pricing models, as in [1] [5] [18] [28].

2.2 Blogosphere, Sentiment Detection

Blogosphere is a novel term used to denote all blogs and their interconnections. A blog is a web-site where entries are commonly displayed in reverse chronological order. A typical blog combines text, photos, videos, and links to other blogs, web pages, and other media related to its

theme (e.g., YouTube, Picasa). People publish blogs in the publishing style of serial journal web page, which provide news, comments, opinion, diaries, and interesting articles. In addition to basic functions, a blog may integrate various convenient applications encoded in JavaScript or HTML. For instance, the ads displayed in Figure 1 are developed using JavaScript. Users can autonomously express their opinions on a blog space and can embed related elements from other web sites. Again, from the user's perspective, we believe that bloggers generally expect to have ads on their page that are related to their positive blog content. Hence, in this study, we concentrated on sentiment-based contextual advertising and on pairing relevant ads with a given blog page to increase the Click-Through Rate.

Recently, TREC[4] (**T**ext **Re**trieval **C**onference) developed a blog task (called TREC-BLOG) that focuses on information retrieval from blog documents [44]. A core task is opinion retrieval that focuses on a specific aspect of blogs: the opinionated nature of the writing (e.g., products, movies and political candidates). The opinion retrieval task involves locating blog posts that express an opinion about a given target. It can be summarized as "What do people think about <***target***>." The target is not restricted to be a named entity (e.g., name of a person, location, or organization) but it can also be a concept (such as a type of technology), a product name, or an event. Note that the topic of the post does not necessarily have to be the target, but an opinion about the target must be present in the post or in the comments on the post. Furthermore, the definition of opinions is interpreted broadly to include expressions of personal opinions (e.g., "I love Skype") as well as reports of other people's opinions (e.g., "Most women think…"). Another main task is polarity opinion retrieval (sentiment classification) that aims to determine the polarity of the opinions in the retrieved documents/sentences, namely whether the opinions are positive, negative or mixed. The evaluation employs IR measures such as **MAP** (Mean Average Precision). The assessments can be classified into three groups: positive, negative and mixed opinions, each of which is self-explanatory.

3 SOCA FRAMEWORK

In our sentiment-oriented contextual advertising framework (**SOCA**), the advertising system processes the content of the page, detects sentiment, and then searches the ad collection to find the best matching ads. Thus, we focused on how to construct a scoring function to determine the relationship between the ads and a blog page. Given a page $\boldsymbol{p}$, which we called a *triggering page*, and a set of ads $\boldsymbol{A}$, the task is to select ads $a_i \in \boldsymbol{A}$ related to the content of a page. As a Web page may have a certain polarity (either positive or neutral or negative), our goal is to find the ads that are related to the positive (and neutral) aspects of the page and rank them according to their relevance.

We designed three processes to assign the relevant ads to a given page, as shown in Figure 3. The sentiment detection mechanism is adopted to detect the sentiment of the blog's contents. We subsequently removed those sentences that reflect negative sentiments. Then, due to limitations in information about the ads and a certain page, we selected some specific terms in the triggering page as a set of seed words that were then used for vocabulary expansion to enhance the likelihood of intersection with available ads. A linear combination function using the traditional cosine similarity and an ontology mapping function was deployed for the page-ad matching strategy to rank the ads.

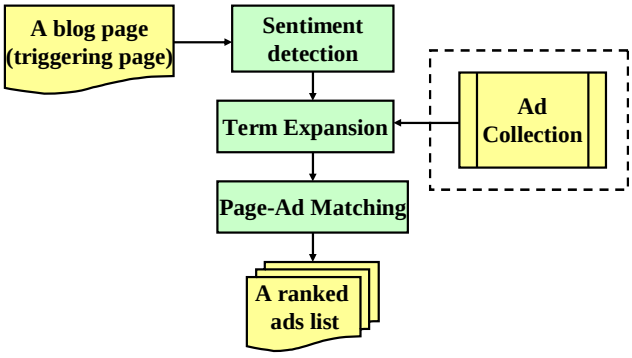

Figure 3: The SOCA Framework.

[4] http://trec.nist.gov/

3.1 Sentiment Detection

Our aim in this section is to apply a contextual sentiment technique for recognizing the intent of a blog page. Generally, researchers study opinion at three different levels: *word* level, *sentence* level, and *document* level [16] [17]. In this study, we regarded a sentence (paragraph) as the minimum unit and we tried to identify opinion-bearing sentences, and classify their sentiment into positive and negative semantic categories using machine learning algorithms and simple linear models.

The first sentiment detection method uses the SVM algorithm with different feature sets, namely uni-gram and opinion-bearing words (we manually pre-selected opinion-bearing words from [8]). We adopted C4.5 algorithm as our second approach with identical feature sets. We compared two supervised learning algorithms for sentiment detection, since SVM is a well-known learning method widely used for classification and regression, while C4.5 generates a decision tree which is more interpretable. Besides, we applied the LibSVM [6] and C4.5, which are included in the Weka machine learning tool kit[5] (a collection of data mining algorithms) for the following tasks. To implement these machine learning algorithms on our dataset, we used the standard bag-of-features framework. Let $\{f_1,\ldots,f_m\}$ be a predefined set of m feature that can appear in a document. Let $w_i(d)$ be the weight of f_i that occurs in document $\boldsymbol{d}$. Then, each document $\boldsymbol{d}$ is represented by the document vector $\vec{d} = (w_1(d), w_2(d),\ldots, w_m(d))$. As for the weighting value, it can be assigned either a boolean value or a *tf-idf* (term frequency – inverse document frequency) value. Here we used the *tf-idf* which is a statistical measure that evaluates how important a word is to a document. The *tf-idf* function assumes that the more frequently a certain term t_i occurs in documents d_j, the more important it is for d_j, and furthermore, the more documents d_j that term t_i occurs in, the smaller its contribution is in characterizing the semantics of a document in which it occurs. Besides, to build an efficient learning model, we divided the task of detecting the sentiment of a sentence into two steps, each of which belongs to a binary classification problem. The first is an identification step that aims to identify whether the sentence is subjective or objective. The second is a classification step that classifies the subjective sentences as positive or negative.

We applied a dictionary from [8] to assign each word a weight (strength) in either a positive, negative or objective direction. Then, we designed our second type sentiment detection approaches, as follows:

$$\text{Model 1: } \prod_{w=1}^{n} (\text{the sign of a word})$$

Model 1 simply considers the polarities of the sentiments. The intuition here is based on English grammar like "double negation". For example, consider a positive sentiment, "I will never be unmoved whenever I see the film." Although this sentence contains two negative words *never* and *unmoved*, it will be reversed into a positive sentiment by the above formula. As for Model 2, it mainly assesses the strength of sentence $sentence_{strength}$; hence, we can sum the strength of each word and the $sentence_{strength}$ can then be normalized by its length, as follows:

$$\text{Model 2: } \frac{\sum_{w=1}^{n} (\text{the strength of a word})}{\text{the length of sentence}}$$

In model 2, we defined a threshold ε (default value of 0.35) to determine the sentence categories according to iterative experiments.

3.2 Term Expansion

In general, a blog page can be about any theme while the advertisements are concise in nature. Hence, the intersections of terms between ads and pages are very low. If we only consider the existing terms included in a triggering page, an ad agency may not accurately retrieve relevant ads, even when an ad is related to a page. According to [1] [28], considering the ads' abstracts and titles is not sufficient to perform page-ad matching. Thus, a term expansion of the keywords in the triggering page as well as the ads is conducted to increase the overlap. For a triggering page, because not all the words included in a page are useful to carry out term expansion, we simply

[5] http://www.cs.waikato.ac.nz/ml/weka/

took the terms tagged as nouns (NN & NNS) as candidate terms from which we generated a set of seed terms according to the following rules.

$T_{Capitalization}$: Whether a candidate term is capitalized is an indication of it being a proper noun, or an important word.

$T_{hypertext}$: Whether a candidate term is part of the anchor text of a hypertext link.

T_{title}: Whether a candidate term is part of the post's title.

$T_{frequency}$: Consistent with term frequency, we considered the three most frequently occurring candidate terms as a subset of the seed terms.

Subsequently, the set of seed terms ($Seed_{Term} = T_{Capitalization} \cup T_{hypertext} \cup T_{title} \cup T_{frequency}$) undergoes three term expansion methods. Two methods are dictionary-based operations that utilize the WordNet[6] and Wikipedia[7] thesauruses, respectively. The third method is a web-based search that identifies pages related to a triggering page to construct a co-occurrence list using the specific terms on a triggering page.

For the first method, we submitted each seed term to WordNet to acquire its synonyms. For instance, a noun *car* has the synonym *automobile*. However, since many product names and acronyms are not covered by WordNet, we introduced another open collaborative dictionary (Wikipedia) to further expand the term. For example, the term *Nokia* has no corresponding information stored in WordNet; on the other hand, Wikipedia contains an entry page about this term. By ranking the words (excluding stop-words) in the entry page based on frequencies, we can select the top five terms to be part of our list of expanded terms. In this example, those terms would be *mobile*, *phone*, *network*, *company*, and *telecommunication*. In addition, Wikipedia provides natural language recognition (e.g., abbreviations, acronyms). As another example, the term *ID4*, which is the acronym of the movie Independence Day, has an entry page in Wikipedia and can be expanded to the movie category using the most frequent terms, namely, *alien*, *film*, *movie*, *novel*, and *president*.

Co-occurrence in the linguistic sense can be interpreted as an indicator of semantic proximity. Certain words often appear together with other words and phrases. We assumed that the Web documents ***D*** similar to the triggering page would share various common topics ***t***. Then by inspecting the terms in these documents we can construct a co-occurrence list for topic ***t***. Hence, the final term expansion technique we applied is a web-based method. For efficiency, we simply took the $T_{title} \cup T_{frequency}$ to represent the blog's topics ***t***. For each topic, we used this to retrieve the top ***k*** (the default is 3) ranked documents ***D*** from the search engine. Then, we adopted the **LODR** (Logarithm of the Odds Ratio)[10] formula to recognize topic-related words in ***D***, as shown in Figure 4.

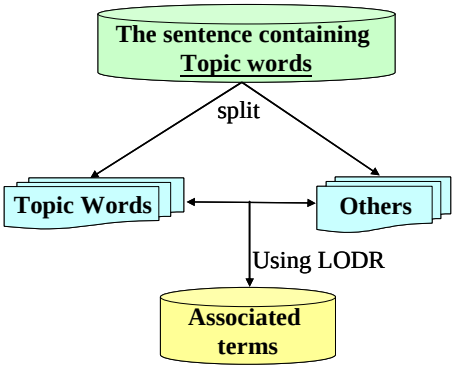

Figure 4: Process for generating associated terms.

We first obtained a sentence set containing the topic words of ***D***. Then, the sentence set is split into two subsets comprising topic words and other words, respectively. Finally, the LODR formula is used to generate associated terms that occur frequently together with a topic word, as follows:

$$LOD_r(w_i, T) = \log \frac{p/(1-p)}{q/(1-q)} = \log \frac{p(1-q)}{q(1-p)}$$

where, T is the topic word, w_i is any word in a set of other words. Let p be the probability that a word, w_i, co-occurs with any of the topic words. That is,

[6] http://wordnet.princeton.edu/

[7] http://en.wikipedia.org

$$p = \mathrm{P}_r\,(w_i \mid \text{sentence containing topic words})$$

Also, let q be the probability that w_i co-occurs with non-topic words. The formula for the occurrence probability of the word w_i with non-topic words is shown below:

$$q = \mathrm{P}_r\,(w_i \mid \text{sentence excluding topic words})$$

If LODr (w_i, T) > δ (where, δ is a threshold), then w_i is considered an associated term that topic words co-occur with.

Following term expansion, we merged the expanded terms generated from the three aforementioned expansion methods into a set of expanded terms ($Expanded_{Term} = E_{WordNet} \cup E_{WiKi} \cup E_{LODR}$) for a given page.

3.3 Page-Ad Matching

We can regard the sentiment-oriented contextual advertising issue as a traditional information retrieval problem, that is, given a user's query ***q***, the IR (Information Retrieval) system returns relevant documents ***d*** according to the query content. Hence, we intuitively model a triggering page ***p*** and relevant ads ***a*** with a user's query ***q*** and corresponding documents ***d***, respectively. In general, the most frequently used data representation in text mining is the bag-of-words approach. Since this representation excludes information about the order of the words, its major merits are conceptual simplicity and relative computational efficiency. Thus, for our data representation, we used the vector space model (VSM), which is a way of representing documents through the words that they contain. Pages and advertisements are represented as weights in an *n*-dimension space. Let $w_{i,p}$ be a weight associated with a term t_i on a page ***p*** and let $w_{i,a}$ be a weight associated with a term t_i in an ad a_j. Then, the page vector $\vec{p}$ is defined as $\vec{p} = \{w_{1,p}, w_{2,p},\ldots,w_{t,p}\}$ and the vector for an ad a_j is defined as $\vec{a} = \{w_{1,a}, w_{2,a},\ldots,w_{t,a}\}$. Moreover, the weight of each vector can be assigned either a boolean value or a *tf-idf* value. Here we adopted the *tf-idf* weight that is a statistical measure used to evaluate how important a word is to a document in a collection. The vector model evaluates the degree of similarity between two documents in terms of the correlation between two vectors. Hence, the ranking of the page ***p*** with regard to the ad ***a*** is computed by the cosine similarity function, that is, the cosine of the angle between the vector $\vec{p}$ and the vector $\vec{a}$:

$$Sim_{Cos}\,(a_j, p) = \frac{\vec{a}_j \bullet \vec{p}}{\left|\vec{a}_j\right| \times \left|\vec{p}\right|} = \frac{\sum_{i=1}^{t} w_{i,a} \times w_{i,p}}{\sqrt{\sum_{i=1}^{t} w^2{}_{i,a}} \times \sqrt{\sum_{i=1}^{t} w^2{}_{i,p}}}$$

In addition, weights computed by *tf-idf* techniques are often normalized so as to counter the tendency of *tf-idf* to emphasize long documents. The type of *tf-idf* that we used to generate normalized weights for data representations in this study is

$$tf\text{-}idf = tf\,(t_i, d_j) \cdot \log \frac{|\mathrm{D}|}{\#_D\,(t_i)}$$

where the factor $tf(t_i, d_j)$ is called the term frequency, the factor $\log \frac{|D|}{\#D(t_i)}$ is called the inverse document frequency, while $\#\mathrm{D}(t_i)$ denotes the number of documents in the document collection ***D*** in which term t_i occurs at least once and

$$tf\,(t_i, d_j) = \begin{cases} 1 + \log \#(t_i, d_j), & \text{if } \#(t_i, d_j) > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\#(t_i, d_j)$ denotes the frequency of t_i in d_j. Weights obtained from the *tf-idf* function are then normalized by means of cosine normalization, finally yielding

$$w_{i,\mathrm{j}} = \begin{cases} \dfrac{tfidf\,(t_i, d_j)}{\sqrt{\sum_{\mathrm{k}=1}^{|\mathrm{T}|} tfidf\,(t_s, d_j)^2}} \end{cases}$$

Generally, the goal of ontology mapping is to compute the domain of two ontologies which are semantically related at a conceptual level. Hence, in addition to the cosine similarity score, we

also considered the terms ontology mapping between a page and an ad to estimate the degree of similarity. Unfortunately, there is no generally accepted ontology for a certain domain. Since the page and the advertisement may match many categories, it is difficult to design a common ontology that is appropriate in every case. Therefore, we simply took a dictionary (WordNet) that contains general terms as an intermediate ontology to map the nouns included in the page and in the ad. Our proposed mapping function mainly considers the relationships and distances between terms. The *relationship* can be divided into *symmetric* and *asymmetric* relationships. An asymmetric relationship Rel_{asym} is one whose source and target synonym sets have lineages with a definite divergence point. The common parent index is the index of the node in the relationship that represents this divergence point. For example, in finding a hypernym relationship between *dog* and *cat*, the relationship is dog -> canine -> carnivore; cat-> feline -> carnivore. Both the ancestry of "dog" and "cat" diverge at "carnivore," so the common parent index is 2; and the distance between them is 4, so we used the inverse distance for the Rel_{asym} value. In the case of a symmetric relationship Rel_{sym}, an example of a relationship would be the synonym. Symmetric relationships differ from asymmetric relationships in that there is no definite divergence point between the ancestry of the source and the target synonym set. For example, the term *gondola* is one of the synonyms for a *car*, which occurs at a depth of four consistent with the conceptual hierarchy. The distance is 4 and the inverse depth is chosen for the Rel_{sym} value.
Given a page-ad pair, the degree of similarity of term ontology mapping is calculated through:

$$Sim_{Onto}(a, p) = \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha Rel_{sym}(t_i, t_j) + (1-\alpha) Rel_{asym}(t_i, t_j)$$

where, t_i and t_j is any noun included on the page and the ad respectively, and the parameter α (default value of 0.2) determines the relative weight of the symmetric and asymmetric relationships.
According to above the description of our similarity formula, we formally defined the relevance score of an ad and a page as a linear combination of the cosine function score and ontology mapping score:

$$Score(a, p) = \beta Sim_{Cos} + (1 - \beta) Sim_{Onto}$$

where the parameter β determines the relative importance of the cosine similarity and ontology mapping score.

3.4 Ad-Content Indexing

In the section 3.3 we discussed the scoring function for an ad given the triggering blog page. The top-*k* ads with the highest score are assigned by the ad agency system. The process of score calculation and ad selection is to be done at retrieval time and therefore must be very efficient.

We adopted a basic inverted index framework including *postings* and *dictionary*, where there is one posting list for each distinct term. The ad contents are tokenized into a list of terms via linguistic preprocessing including stemming, stop word removing to produce a list of normalized tokens, called *indexing terms*. For each indexing term, we had a list that records which ads the term occurs in, and its associated weight (*tf***idf* value). The list is then called a *postings list*. The posting lists contain one entry per indexing term/ad combination. For more details, gentle readers are referred to [3].

To be able to search the ads based on a combination of keywords and ontologies, we implemented ontology match via an intermediate index table which records the terms and their *related terms* included in WordNet. The related terms are determined by the symmetric relationship (i.e., synonym) and the asymmetric relationship (the terms included in a default distance range, in here the distance is 5). We then use these related terms to retrieve ads through keyword match indexing to calculate the ontology score.

4 EXPERIMENTAL RESULTS

In this section, we focus on our two experiments, namely, sentiment detection and page-ad matching. We begin by describing the dataset and text preprocessing, and we then proceed to a discussion of the experimental results.

4.1 Datasets and Text-Preprocessing

To evaluate sentiment detection performance, we collected data from *epinions.com* and used this as our training dataset for building our learning classifier models. The sentiment dataset from *epinions.com* expresses positive and negative user experiences on specific fields of web pages. Hence, we adopted the label processing method proposed in [17] to automatically mark the positive, negative and neutral sentences. Kim & Hovy's method firstly extracts comma-delimited phrases from each pro and con field in a review document, generating two sets of phrases: $\{P_1, P_2,\ldots,P_n\}$ for pros and $\{C_1, C_2,\ldots,C_m\}$ for cons. For each phrase in two sets, the system checks each sentence to discovery a sentence that covers most of the words in the phrase. Then the system annotates this sentence with appropriate "pro" or "con" label. All remaining sentences with neither label are marked as "neither". They used these data and maximum entropy model with different feature categories (such as unigram, opinion-bearing word) to train pro and con sentence recognition system. In this study, we gathered many different types of reviews from *epinion.com* (e.g., 3C products, hotels, movies, travel, theme parks, and second-hand cars). For sufficient detail, we examined 32,304 reviews (938,621 sentences), with an average number of sentences in per review document of 29.

For page-ad matching, because of the lack of large-scale ad databases, we first chose certain general topic words (e.g., alcohol, book, clothes, cosmetics, culture, game, laptop, medicine, mobile phone and sport) as query terms to request web pages from search engines such Google and Yahoo!. About 10,000 pages were retrieved and we placed these pages on an ad-crawler platform to obtain the corresponding ads assigned by Google AdSense. Our ad-crawler was similar to a generic blog website that can be embedded in a JavaScript module (e.g., Google AdSense). We firstly extracted the content of each retrieved page (about 10000 pages) and then regarded this content as a blog post to get the corresponding ads assigned by Google AdSense JavaScript model. Then, we extracted the ads by a simple program mainly coded in regular expression. For each blog post, we can obtain about 10 unique ads. It is a reasonable way to collect real-world advertisements; we totally collected 104,094 ads. Unfortunately, we were unable to obtain hidden information about the ads, such as the ad keywords (bid phrases). We will briefly describe how to generate simulated advertiser keywords. For each ad and its *landing page*, we first generated a set of seed terms consistent with the rules discussed in Section 3.2; we then submitted them to keyword suggestion toolkits (e.g., Google AdWords Keywords[8] and Yahoo! Search Marketing[9]) to obtain the top 5 most popular terms and thereby approximate the likely advertiser keywords. We collected 280,613 unique keywords in our collection and the average numbers of keywords in per advertisement was 8. Our triggering page collection comprised 150 blog pages on various topics. It included a range of opinions and comprised various subjective articles (100 positive (neutral) and 50 negative articles). We selected triggering pages according to the ratio of positive and negative sentences, that is, if the ratio of positive to negative sentence was over 4:1, we regarded a triggering page as positive, and vice versa.

To acquire the POS tag, we adopted the GENIA Tagger developed by the University of Tokyo[10]. In addition, we preprocessed the full text of a triggering page with its expanded terms (as discussed in Section 3.2) and the ads (including the full text of the landing page and its abstract) with their expanded terms by removing stop words and one-character words, followed by stemming [26].

4.2 Evaluation of Sentiment Detection

The goals of this section are similar to [17]: the first is to explore how well the positive and negative detection model with different approaches on the data collected from *epinions.com*; and the second is to investigate how well the trained model performed on a different data source (150 triggering pages).

For the *epinions.com* data, we compared various methods as described in Section 3.1 for sentiment detection. We adopted a ten-fold cross validation mechanism for learning algorithms. The precision (proportion of identified sentences that are marked as subjective sentence to all the identified sentences), recall (proportion of the marked subjective sentences that are identified, out of all marked subjective sentences available) and F-measure (weighted harmonic mean of precision and recall) are used as the evaluation measures. Table 1 shows subjective sentence

[8] https://adwords.google.com/select/KeywordToolExternal
[9] http://help.yahoo.com/l/us/yahoo/ysm/sps/
[10] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

identification results. As can be seen in this table, C4.5 gave average 63% precision, 62% recall and 63% F-measure, whereas SVM classifier gave better results, yielding average 64% precision, 65% recall, and 64% F-measure. Besides, our results show that best precision (64.17%) and recall (65.1%) performance appeared using SVM with opinion-bearing words features. Moreover, our results also suggest that there is no significant different in terms of precision, recall and F-measure between an SVM that uses unigrams and an SVM with opinion-bearing words.

Table 1: Subjective sentence identification results.

Approach	Precision	Recall	F-measure
SVM with unigram	64.14 %	64.8 %	64.5 %
SVM with opinion-bearing words	**64.17 %**	**65.1 %**	**64.6 %**
C4.5 with unigram	63.60 %	62.6 %	63.1 %
C4.5 with opinion-bearing words	61.60 %	60.7 %	61.1 %

For the sentiment classification experiment, we used the results of sentiment identification as input. The goal of this experiment was to classify the subjective sentence into a suitable class (i.e., positive or negative). Since our two linear models are not classified into two steps in the task of sentiment detection, the meaning of our final results in linear models is identical to that of sentiment classification. The precision (proportion of classified sentences that are marked as negative sentence to all the classified sentences), recall (proportion of the marked negative subjective sentences that are classified, out of the entire marked negative sentences available) and F-measure (weighted harmonic mean of precision and recall) are similarly used as the evaluation metrics. Table 2 shows results for sentiment classification experiment. The results clearly show that best precision (59.9%) and recall (62.3%) performance are respectively produced by the SVM with unigram and by an SVM with opinion-bearing words. As for C4.5, the better precision (58.8%) and recall (58.8%) are generated by C4.5 with unigram. However, our linear model 1 and model 2 respectively produced the F-measures of 33.64 % and 32.71 %. Moreover, as can be seen in this table, the learning algorithms can outperform the linear models around 20 % in terms of precision, recall and F-measure.

Table 2: Positive and negative sentence classification results.

Approach	Precision	Recall	F-measure
SVM with unigram	**59.9** %	61.2 %	**60.1** %
SVM with opinion-bearing words	54.2 %	**62.3** %	57.9 %
C4.5 with unigram	58.8 %	58.8 %	58.8 %
C4.5 with opinion-bearing words	46.1 %	43.5 %	44.8 %
Linear model 1	39.10 %	29.52 %	33.64 %
Linear model 2	32.90 %	32.53 %	32.71 %

According to the above results, it seems reasonable to infer that the SVM with unigram feature model has no significant effect to one which uses opinion-bearing words. Furthermore, using learning algorithms to detect sentiment seems more appropriate than using linear models. As for feature selections in the sentiment classification task, the learning models with unigram features have more significant effects than learning models which adopted opinion-bearing words features. In the future, in addition to reformulating our linear models, we plan to incorporate scores from linear models as one of the features of our learning frameworks.

For the dataset of triggering pages, owing to a lack of training data, according to above results, we subsequently chose learning algorithms with different feature sets to train the models on the *epinion.com* dataset. We then applied these models to our triggering page dataset. In order to get a reasonable evaluation, human experts had to annotate the entire triggering page set. The human experts were divided into two groups to generate gold-standard independently; besides, the average pair-wise human agreement measure between two groups was 0.89 and the Kappa coefficient value was 80.8%. The resulting numbers of objective and subjective labeled sentences were 3593 and 1434, respectively. The subjective sentences contained 976 positive and 458 negative sentences. We conducted the identical sentiment identification and classification step and the results are shown in Table 3. As for the sentiment identification task, the best precision (69.10%), recall (66.10%) and F-measure (67.33%) performance are respectively generated by C4.5 with opinion bearing words and by SVM with unigram. For the sentiment classification task, SVM with unigram obtained the best 56.17 % precision, 62.12 % recall and 58.99 % F-measure. As can be seen in our results, although the triggering pages covered various topics (few intersected

with *epinions.com*'s), the effects are consistent with the *epinions.com* data. It seems reasonable to conclude that our trained sentiment classifier models are not limited to a few specific domains.

Table 3: System results for triggering pages.

Approach	Detection Task	Precision	Recall	F-measure
SVM with unigram	Identification	64.76 %	**66.10 %**	65.42 %
	Classification	**56.17 %**	**62.12 %**	**58.99 %**
SVM with opinion bearing words	Identification	61.92 %	62.97 %	62.44 %
	Classification	53.07 %	57.10 %	55.00 %
C4.5 with unigram	Identification	66.80 %	65.15 %	65.96 %
	Classification	50.45 %	52.42 %	51.42 %
C4.5 with opinion bearing words	Identification	**69.10 %**	65.65 %	**67.33 %**
	Classification	55.75 %	48.25 %	51.73 %

4.3 Evaluation of Page-Ad Matching

The goal of this section is to investigate to what extent the ad placements are actually related to the positive (and neutral) aspects of the triggering pages. To evaluate our page-ad matching framework, we compared the top-10 ranked ads provided by three different ranking methods, namely our proposed approach with sentiment detection (i.e., **SOCA** (Sentiment-Oriented Contextual Advertising)), our proposed approach without sentiment detection (i.e., **CA** (Contextual Advertising)), and Google AdSense. Here we used the SVM with unigram features as sentiment detection model according to our results of sentiment detection. No more than 30 ads were retrieved and inserted into a pool for each triggering page. All the advertisements in each pool were manually judged by experts. The experts mainly evaluated each page-ad pair according two principles, *correlation* and *intention*. The correlation principle is that whether this ad is positively related to the content of a given blog page. Another principle is that whether the experts have any intention to click this ad. An ad judged as golden-standard has to comply with both correlation and intention principles. The human experts were divided into two teams to individually label gold-standard; moreover, the average pair-wise human agreement measure and Kappa coefficient value between two teams were 0.94 and 0.87%, respectively.

By comparing with Google AdSense, we only measured the accuracy for this experiment. The experts regarded an ad related to the positive (neutral) aspects of the triggering pages as a target. In other words, on a triggering page that features negative opinions with respect to a particular topic, an advertisement on that topic will be judged inappropriate if it appears on that page. In order to investigate the effectiveness of our proposed method on a different sentiment dataset, we further selected positive sentiment dataset (100 documents) from our triggering pages. Table 4 shows the results for three page-ad matching methods across various types of datasets. In the case of the positive sentiment dataset, there are no significant differences among the three page-ad matching approaches. Our SOCA framework and CA can respectively produce 65.2% and 67.0% in terms of precision. Google AdSense achieves about 64.5% accuracy. Regarding all triggering pages, our results show that the proposed SOCA approach can yield best performance (68.2 %) than other approaches (57.1% for CA approach and 52.3% for Google AdSense). According to Table 4, these results lead us to the conclusion that our sentiment-oriented contextual advertising framework (SOCA) can place ads that are related to the positive (and neutral) content of triggering pages.

Table 4: Accuracy of page-ad matching.

Dataset	Method	Accuracy
Positive sentiment dataset	SOCA	65.2 %
	CA	**67.0 %**
	Google AdSense	64.5 %
All Triggering pages	SOCA	**68.2 %**
	CA	57.1 %
	Google AdSense	52.3 %

Although the results generated by our proposed method are better than Google's, in this paper we did not emphasize this conclusion on the basis of two reasonable reasons. One of the reasons is that Google AdSense needs to select the recommended ads out of an ad pool that is vastly larger

than the one used by us. Another possible reason is due to the ad categories. That is, Google AdSense considers more various ad categories than the categories adopted by our.

The goal of our next experiment was to explore our SOCA framework in detail. We selected the top 10 ranked ads provided by three matching mechanisms, namely cosine, ontology function and a combination of cosine and ontology (Cos+Onto). We used all triggering pages as our dataset. We thereby ensured that no more than 30 ads would be retrieved and inserted into a pool for that triggering page. All the advertisements in each pool were manually judged by experts. The human experts similarly were divided into two groups to independently judge gold-standard; furthermore, the average pair-wise agreement measure and Kappa coefficient value between two teams can achieve 0.95 and 0.91, respectively. The average number of relevant advertisements was 9 per triggering page. To quantify the precision of our results, we applied an 11-point averaged figure. Since it is quite difficult to evaluate our entire ad collection, our recall values were only relative to the set of judged advertisements.

The results of our proposed page-ad matching approach are shown in Figure 5. Each data point corresponds to the precision value calculated at a certain percentage of recall. The results clearly indicated that using the combination of cosine and ontology (Cos+Onto) with default β = 0.8 can achieve better performance than the use of cosine (β = 1) and ontology (β = 0) alone.

In addition to the precision-recall curve, we also used another presentation involving two quality measures (Precision@K and mean average precision (MAP)) to assess matching results:

※ We calculated the average retrieval precision computed at recall level K as follows:

$$\text{Precison@}(K) = \frac{\sum_{i=1}^{N_q} P_i@(K)}{N_q}$$

where Precision@(K) is the average precision at recall level K, N_q is the number of queries used, and P_i is the precision at recall level K for the i-th query.

※ To compare the precision-recall curves across the three page-ad matching functions, we computed MAP. For a single query, Average Precision is the average of the precision value derived for the set of k top documents that exist after each relevant document is retrieved. This value is then average over all queries. That is, if the set of relevant documents for a query $q \in Q$ is $\{d_1, d_2, \ldots, d_{mj}\}$ and R_{jk} is the set of ranked retrieval results from the top result until the retrieval system returns the documents d_k, then

$$\text{MAP}(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}\text{Precision}(R_{jk})$$

where, Q $\{q_1, q_2, \ldots, q_m\}$ is a set of queries. Since we have three page-ad matching strategies, we computed the MAP score for three query sets.

The results are displayed in Figure 6. It is clear that the cosine and ontology approach can generate MAP of around 39% and 29%, respectively; besides, improved performance (of around 43%) can be produced by Cos+Onto. As shown in these figures, the results based on cosine similarity are apparently very powerful and consistently superior than using ontology alone. However, the combination of cosine and ontology does offer a slightly positive effect.

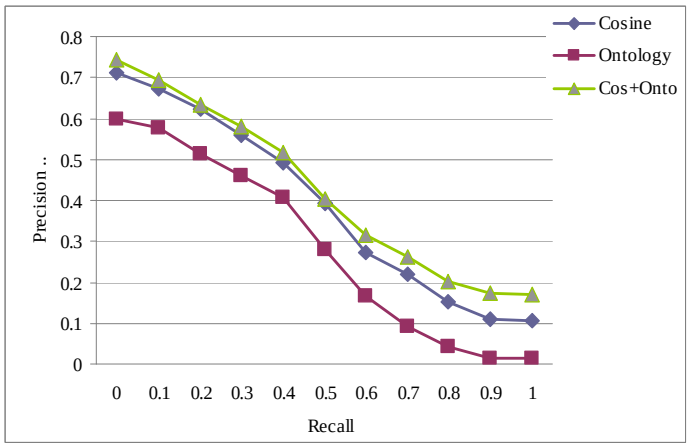

Figure 5: Precision-Recall 11-point curve.

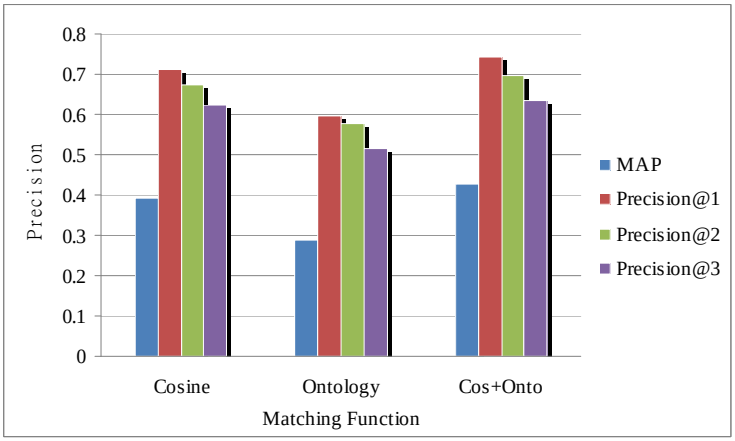

Figure 6: The performance of the three matching strategies.

In another experiment, we used only the negative sentiment articles and identical experimental parameters as mentioned previously to perform page-ad matching. The precision-recall curve and bar chart are respectively shown in Figures 7 and 8. The results are roughly similar to other results which we generated using all the available triggering pages. The Cos+Onto function similarly was able to achieve better performance than cosine or ontology alone. In detail, the best MAP performance (of around 47%) can be generated by Cos+Onto, besides, the cosine and ontology approach can achieve MAP of around 38% and 35%, respectively.

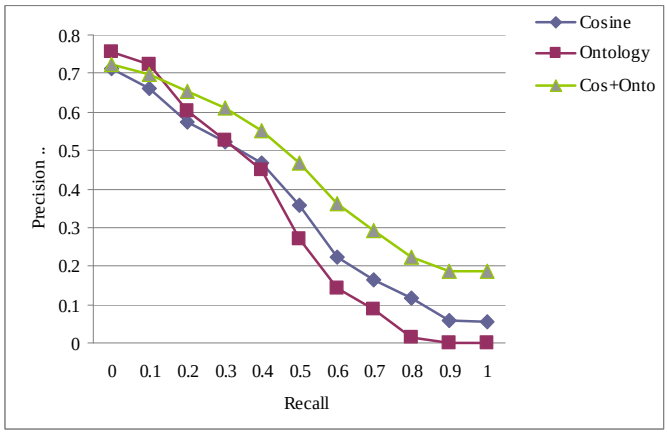

Figure 7: Precision-Recall curve for the negative sentiment dataset.

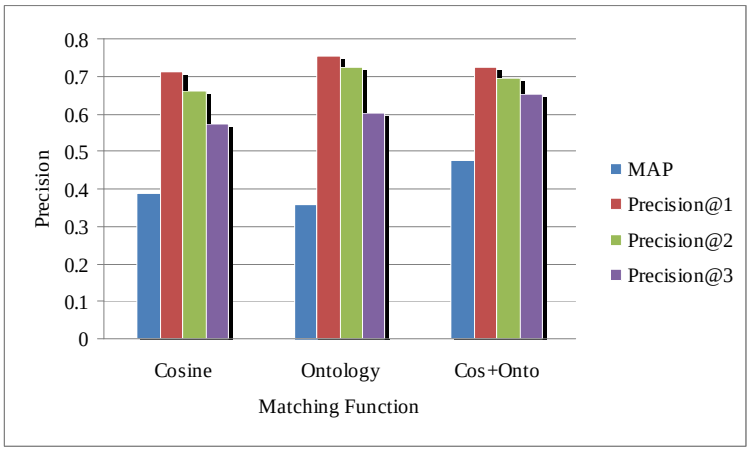

Figure 8: The performance of the three matching strategies on the negative sentiment dataset.

In our last experimental analysis, we carried out the same experiment for the positive sentiment dataset. However, the results displayed in Figure 9 and 10 were essentially the same as those obtained using all triggering pages and just the negative sentiment dataset, as shown in Figure 9 and 10. For sufficient detail, the MAP of around 40% and 27% can be achieved by Cosine

and Ontology, respectively, moreover, the Cos+Onto can generate better MAP performance of around 43%.

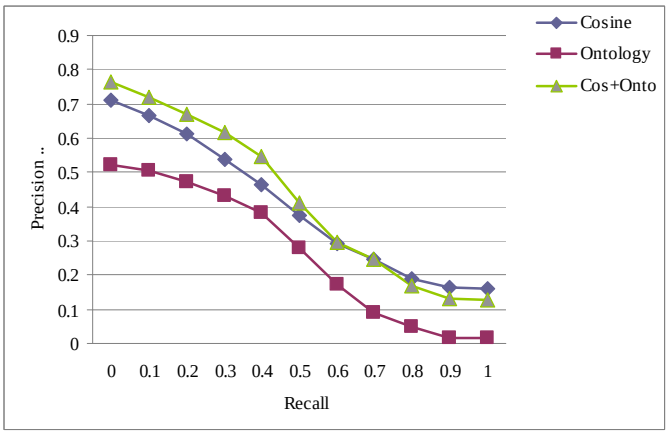

Figure 9: Precision-Recall curve for the positive sentiment dataset.

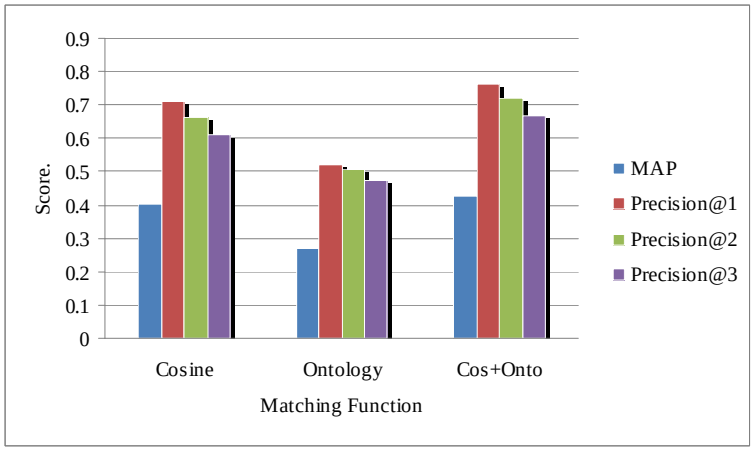

Figure 10: The performance of the three matching strategies on the positive sentiment dataset.

According to our evaluation of page-ad matching, these results lead us to the conclusion that our SOCA framework can place ads that are related to the positive (and neutral) content of triggering pages. In addition, the combination of cosine and ontology can yield the best performance in terms of MAP, no matter what type of sentiment datasets (such as, positive and negative) are used.

4.4 Deployment of SOCA & Performance Issues

In a sponsored search advertising system, since ads are assigned at query time, the major concern of the system is the efficiency. For contextual advertising system, we can associate ads with a blog post at publishing time. In addition, we might consider assigning new ads to already published blog pages in offline mode. In other words, we might devise the ad agency system based on the frequency of new blog pages and the frequency of ads update. Besides, the related information (i.e., the related web page crawling, term expansion and ad indexing) can be gathered and processed in offline mode.

As for the performance of term expansion, two dictionary-based expansion methods can be performed within mille-second. For web-based expansion method, we might design an expansion term database to store (or index) the corresponding web pages of the seed terms and the expanded term generated by LODR in advance. However, if a triggering page containing some processed seed terms and a new seed term comes in, we firstly might retrieve the existing information (i.e., the expanded terms of pre-processed seed terms) included in the expansion term database to perform page-ad matching task and then assign the related ads. In the meantime, we might do the term expansion on the new seed term then update both expanded term database and related ads of the given triggering page. Thus, this environment deployment might not only ease on-line latency

issue, but also does not affect the end-user browsing.

5 RELATED WORK

Several prior research studies are relevant to our work, including efforts in online advertising, sentiment classification and web application.

Several studies pertaining to advertising research have stressed the importance of relevant associations for consumers [19] [36] and how irrelevant ads can turn off users and relevant ads are more likely to be clicked [7] [25]. They show that advertisements that are presented to users who are not interested can result in customer annoyance. Thus, in order to be effective, the authors conclude that advertisements should be relevant to a consumer's interests at the time of exposure. Novak *et al*. [21] reinforce this conclusion by pointing out that the more targeted the advertising, the more effective it is. As a result, certain studies have tried to determine how to take advantage of the available evidence to enhance the relevance of the selected ads. For example, studies on keyword matching show that the nature and number of keywords impact the likelihood of an ad being clicked [22]. As for contextual advertising, Ribeiro-Neto *et al*. [28] proposed a number of strategies for matching pages to ads based on extracted keywords. The first five strategies proposed in this work match pages and ads based on the cosine of the angle between their respective vectors. To identify the important parts of the ad, the authors explored the use of different ad sections (e.g., bid phrase, title and body) as a basis for the ad vector. The winning strategy required the bid phrase to appear on the page, and then ranked all such ads using the cosine of the union of all the ad sections and the page vectors. While both pages and ads are mapped to the same space, there exists a discrepancy (called "impedance mismatch") between the vocabulary used in the ads and on the pages. Hence, the authors achieved improved matching precision by expanding the page vocabulary with terms from similar pages. In follow-up research [18], the authors proposed a method to learn the impact of individual features using genetic programming to generate a matching function. The function is represented as a tree comprised of arithmetic operators and functions as internal nodes, and different numerical features of the query and ad terms as leaves. The results show that genetic programming can identify matching functions with significantly improved performance compared to the best method proposed in [28]. Besides, Papadopoulos *et al*. [24] propose the use of lexical graphs created from web corpora as a means of computing improved content similarity metrics between ads and web pages. The results indicated that using lexical graph can provide evidence of significant improvement in the perceived relevance of the recommended ads.

However, due to the vagaries of phrase extraction, and the lack of context, approaches based on "bid phrases" leads to many irrelevant ads. To overcome this problem, Broder *et al*. [5] proposed a system for contextual ad matching based on a combination of semantic and syntactic features, that is, the authors define the relevance score of an ad and page as a convex combination of the keyword (syntactic) and classification (semantic) score. The keyword score is modeled in vector space such that both the pages and ads are represented in n-dimensional space (one dimension for each distinct term). The score is then defined as the cosine of the angle between the page and the ad vectors. The semantic score relies on the classification of pages and ads into about 6000 nodes within a commercial advertising taxonomy to determine topical distance. The experimental results show that the semantic-syntactic approach outperformed the syntactic (keyword) approach over a set of pages with various types of contextual advertising. In their follow-up work [1], they consider that analyzing the entire body of such pages on-the-fly entails prohibitive communication and latency costs. Hence, the authors propose to use text summarization techniques paired with external knowledge (exogenous to the page) to craft short page summaries in real time for solving either the low-relevance or high-latency challenges. The ad retrieval function is formulated as a linear combination of similarity scores based on both vector space model and classification feature, similar to [1]. The major difference between [1] and [5] is that [1] only considers an excerpt from the document to perform classification as a means of reducing transmission and analysis time.

Even if several prior studies have proved that the relevance has a definite impact on contextual advertising and on the proposed effective ranking function that matches ads with pages, they neglect sentiment in the assignment of relevant ads. In this study, in addition to considering general syntactic (keyword) and semantic (ontology) matching, we further investigate the importance of sentiment analysis for improved contextual advertising. Besides, according to [28] and [1], these authors have respectively explored the effects of page-ad matching by analyzing fragments of ads and pages. Their results indicate that page-ad matching strategies combined with more related information can achieve better performance. Hence, in this work, we adopted the full

text of a blog page including its expanded terms and ads (including the full text of the landing page, the ad's abstract and title) to design our ad delivery system.

Sentiment classification has been pursued in multiple ways. While most researchers use a supervised approach [29] [39], others use an unsupervised approach [34]. The can be classified into three different levels: words, sentences and documents [17] [40]. Hatzivassiloglou and McKeown [12] described an unsupervised learning method for identifying positively and negatively oriented adjectives with an accuracy of over 90%. They demonstrated that the semantic orientation, or polarity, is a consistent lexical property that exhibits a high inter-rater agreement. Turney [34] showed that it is possible to use only a few of those semantically oriented words (namely, "excellent" and "poor") to label other phrases co-occurring with them as positive or negative. These phrases were subsequently used to automatically separate positive and negative movie and product reviews, with accuracies of 66-84%. Pang *et al*. [23] adopted supervised machined learning with words and n-grams as features to predict orientation at the document level; they achieved up to 83% precision. Yu and Hatzivassiloglou [43] presented a Bayesian classifier for discriminating between documents such as editorials, and they described three unsupervised, statistical techniques for detecting opinions at the sentence level. For opinion/fact sentence classification, they consider various features including n-grams, parts of speech, and polarity words. In addition, for each opinion, they adopted the aggregate polarity score to classify a positive or negative sentiment.

Kim and Hovy [17] presented a system that, given a topic, automatically finds those people who hold opinions about this topic and the sentiment of each opinion. They proved that not all of the sentiment words in sentences are important, and that some may run counter to the true opinion of the author. Their conclusion was that it would be preferable to only consider the author's emotions or desire about the topic as expressed in sentences. Their system used a sentiment words list and WordNet to classify the opinions at the word and sentence level. In their subsequent research [16], they focused on the identification of pro and con reasons in online reviews. Lexical (n-grams), positional and opinion-bearing word feature sets were used in a maximum entropy model to extract pros and cons from review sites. Since labeling each sentence is a time-consuming and costly task, the authors proposed an automatic labeling framework based on the existing pro and con information of some specific web-sites (such as *epinions.com*). In this paper, we adopted Kim's concept to quickly build a sentiment detection model. The major difference between Kim's model and ours is that we adopted two supervised learning algorithms for sentiment detection, since SVM is a well-known learning method widely used for classification and regression, while C4.5 generates a decision tree which is more interpretable. Besides, we further compared the effects by using learning algorithm with different feature sets and linear models.

Due to the growth of the World Wide Web, many studies combine their proposed method with web-based resources (e.g., Wiki-pedia, search engine, or Blogs). Wang *et al*. [37] construct a thesaurus of concepts from Wikipedia then introduce a unified framework to expand the "Bag of Words" representation with semantic relations (synonym, hyponymy, and associative relations) for text classification. Sato and Sasaki [31] [32] propose an automatic web-based method of collecting technical terms that are related to a given seed term. Turney [35] and Wong *et al*. [41] respectively apply the search engine and Wiki-pedia as distance measures to calculate the mutual information between words. Sahami *et al*. [30] propose a novel approach for measuring the similarity between short text snippets by leveraging web search results to provide greater context for the short texts. In addition, more and more studies have been conducted on measuring the semantic similarity between words using web information [2] [35]. It is noteworthy that some studies [13] [14] [45] present interesting subjects that are based on user-oriented information, such as user's review, opinion and blogs. It is clear that the importance of web-application cannot be overemphasized. Thus, in this paper, to implement term expansion effectively, we chose some characteristic terms as the seed terms to collect related documents by web mining.

Taxonomy construction and e-marketing are other issues relevant to our work, Woon and Madnick [42] apply term co-occurrence frequencies as an indicator of the semantic closeness between terms. In order to automatic taxonomy construction they not only propose a modification to the basic distance measure, but also describe a set of procedures by which these measures may be converted into estimates of the desired taxonomy. For e-marketing, Tran [33] describe a framework for modeling the trustworthiness of sellers in the context of an electronic marketplace where multiple selling agents may offer the same goods with different qualities and selling agents may alter the quality of their goods. Besides, in order to offer more targeted and personalized products and services to customers, recent research [4] [27] proposed a direct grouping-based approach that combines customers into segments by optimally combining transactional data of several customers and building a data mining model of customer behavior for each group.

However, Jiang [15] proposes a new micro-targeting method that builds predictive models of customer behavior not on segments of customers but rather on the customer product groups.

6 CONCLUSION

In this study, we proposed and evaluated a novel framework for associating ads with blog pages based on sentiment analysis. Prior work to date has only examined the extent of content relevance between pages and ads. In this paper, we investigated the sentiments of blog pages and utilized this information to demonstrate sentiment-oriented contextual advertising. For sentiment detection, we compared machine learning-based algorithms with different feature sets and two linear models. We used the *epinions.com* data source for training. Our results showed that using learning algorithms can outperform linear models around 15 %. Besides, the best performance on sentiment detection may be as much as 64% in terms of F-measure.

As for page-ad matching, we evaluated our framework using 150 blog pages and over 100,000 ads sampled from Google AdSense. First, we compared SOCA with Google AdSense and found that our proposed method with sentiment detection can achieve superior performance (68.2% accuracy). To analyze our SOCA in detail, we evaluated three matching strategies (i.e., cosine similarity, ontology similarity and the combined approach). Our results indicated that the three proposed approaches can assign relevant ads to the positive (and neutral) aspects of a blog page. The combined approach has a better performance than cosine and ontology independently.

In the future, we intend to conduct a more comprehensive analysis of our model and explore the effectiveness of sentiment detection using different machine learning algorithms (such as HMM and CRF) with different data sources (e.g., *alatest.com* and *reviews.cnet.com*) and natural language processing techniques (such as name entity). We may also apply the concept of topic analysis to pages and ads to enhance their performance. In addition, we wish to explore a broader and more detailed ontology topology to improve processing efficiency.

ACKNOWLEDGEMENT

This work is sponsored by National Science Council, Taiwan under grant NSC97-2627-E-008-001.

REFERENCES

[1] Anagnostopoulous, A., Broder, A. Z., Gabrilovich, E., Josifovski, V. & Riedei, L. (2007) Just-in-Time Contextual Advertising. Proceedings of the 16th ACM conference on CIKM, pp. 331-340.

[2] Bollegala, D., Matsuo, Y. & Ishizuka, M. (2007) An Integrated Approach to Measuring Semantic Similarity between Words Using Information available on the Web. Proceedings of the NAACL HLT, pp. 340-347.

[3] Chapter 8: Baeza-Yates, R. & Ribeiro-Neto, B. (1999) Modern Information Retrieval: Addison Wesley.

[4] Boztug, Y. & Reutterer, T. (2006) A Combined Approach for Segment-Specific Analysis of Market Basket Data. Eur J Oper Res.

[5] Broder, A., Fontoura, M., Josifovski, V. & Riedel, L. (2007) A semantic approach to contextual advertising. Proceedings of the 30th international conference on SIGIR, pp. 559-566.

[6] Chang C.-C. & Lin C.-J. (2001) LIBSVM: a library for support vector machine.

[7] Chatterjee, P., Hoffman, D. L. & Novak, T. P. (2003) Modeling the Clickstream: Implications for Web-Based Advertising Efforts. Marketing Science 22: 520-541.

[8] Esuli, A. & Sebastiani, F. (2006) SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. Proceedings of the 5th Conference on Language Resources and Evaluation, pp. 417-422.

[9] Feng, J., Bhargava, H. K. & Pennock, D. (2003) Comparison of Allocation Rules for Paid Placement Advertising in Search Engine. Proceedings of the 5th international conference on Electronic commerce, pp. 294-299.

[10] Fleiss, J.L., Levin, B., Paik, M.C. (2003) Statistical Methods for Rates and Proportions. Statistical Methods for Rates and Proportions. Chapter 6, Wiley.

[11] Gupta, N., Yang, F., Demers, A. J., Gehrke, J. & Shanmugasundaram, J. (2007) User-centric personalized extensibility for data-driven web applications. Proceedings of the 2007 ACM SIGMOD international conference on Management of data, pp. 1125-1127.

[12] Hatzivassiloglou, V. & Mckeown, K. R. (1997) Predicting the Semantic Orientation of Adjectives. Proceedings of the 35th ACL, pp. 174-181.
[13] Hu, M. & Liu, B. (2004) Mining Opinion Features in Customer Reviews. Proceedings of the conference on AAAI, pp. 755-760.
[14] Hu, M. & Liu, B. (2004) Mining and summarizing customer reviews. Proceedings of the conference on SIGKDD, pp. 168-177.
[15] Jiang, T. & Tuzhilin, A. (2008) Dynamic micro-targeting: fitness-based approach to predicting individual preferences. Knowl Inf Syst Int J (KAIS). Doi: 10.1007/s10115-10008-10149-z.
[16] Kim, S.-M. & Hovy, E. (2004) Determining the sentiment of opinions. Proceedings of the 20th international conference on COLING, pp. 1367-1373.
[17] Kim, S.-M. & Hovy, E. (2006) Automatic Identification of Pro and Con Reasons in Online Reviews. Proceedings of the COLING/ACL, pp. 483-490.
[18] Lacerda, A., Cristo, M., Gonçalves, M. A., Fan, W. g., Ziviani, N. & Ribeiro-Neto, B. (2006) Learning to advertise. Proceedings of the 29th annual international conference on SIGIR, pp. 549-556.
[19] Langheinrich, M., Nakamura, A., Abe, N., Kamba, T. & Koseki, Y. (1999) Unintrusive customization techniques for Web advertising. The International Journal of Computer and Telecommunications Networking 31: 1259-1272.
[20] Mayora, O., Daras, P., Panebarco, M., Achilleopoulos, N., Stollenmayer, P., Williams, D., Magnenat-Thalmann, N., Guerrero, C., Pelt, M., McGrath, T., Fuenmayor, E., Salama, D., Alvarez, F., Kalapanidas, E., Shani, A. & Moine, J.-Y. L. (2008) User centric media in the future internet trends and challenges. Proceedings of the 3rd international conference on Digital Interactive Media in Entertainment and Arts, pp. 441-446.
[21] Novak, T. P. & Hoffman, D. L. (1997) New metrics for new media: toward the development of Web measurement standards. World Wide Web Journal 2: 213-246.
[22] OneUpWeb. How keyword length affects conversion rates, January 2005. Available at http://www.oneupweb.com/landing/keywordstudy_landing.htm.
[23] Pang, B., Lee, L. & Vaithyanathan, S. (2002) Thumbs up? Sentiment Classification Using Machine Learning Techniques. Proceedings of the Conference on EMNLP, pp. 79-86.
[24] Papadopoulos, S., Menemenis, F., Kompatsiaris, Y. & Bratu, B. (2009) Lexical Graphs for Improved Contextual Ad Recommendation. Proceedings of the 31st European Conference on Information Retrieval pp. 216-227.
[25] Parsons, J., Gallagher, K. & Foster, K. D. (2000) Messages in the Medium: An Experimental Investigation of Web Advertising Effectiveness and Attitudes toward Web Content. Hawaii International Conference on System Sciences.
[26] Porter, M. F. (1980). An algorithm for suffix stripping. Program, 14 (3), pp. 130-137.
[27] Reutterer, T., Mild, A., Natter, M. & Taudes, A. (2006). A dynamic segmentation approach for targeting and customizing direct marketing campaigns. Journal of Interactive Marketing 20: 43-57.
[28] Ribeiro-Neto, B., Cristo, M., Golgher, P. B. & Moura, E. S. d. (2005) Impedance coupling in content-targeted advertising. Proceedings of the 28th annual international conference on SIGIR, pp. 496-503.
[29] Riloff, E. & Wiebe, J. (2003) Learning Extraction Patterns for Subjective Expressions. Proceedings of the 2003 conference on EMNLP, pp. 105-112.
[30] Sahami, M. & Heilman, T. D. (2006) A web-based kernel function for measuring the similarity of short text snippets. Proceedings of the 15th international conference WWW, pp. 377-386.
[31] Sato, S. (2001) Automated Editing of Hypertext Resume from the World Wide Web. Proceedings of the Symposium on Applications and the Internet, pp. 15-22.
[32] Sato, S. & Sasaki, Y. (2003) Automatic collection of related terms from the web. Proceedings of the 41st ACL, pp. 121-124.
[33] Tran, T. (2009) Protecting buying agents in e-marketplaces by direct experience trust modelling. Knowl Inf Syst Int J (KAIS). Doi: 10.1007/s10115-10008-10188-10115.
[34] Turney, P. D. (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th ACL, pp. 417-424.
[35] Turney, P. D. (2001) Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. Proceedings of the ECML, pp. 491-502.
[36] Wang, C., Zhang, P., Choi, R. & D'Eredita, M. (2002) Understanding consumers attitude toward advertising. Proceedings of the 8th Americas Conference on Information Systems, pp. 1143-1148.
[37] Wang, P., Hu, J., Zeng, H.-J. & Chen, Z. (2009) Using Wikipedia knowledge to improve text

classification. Know and Inf Sys J (KAIS). Doi: 10.1007/s10115-10008-10152-10114.
[38] Weideman, M. & Haig-Smith, T. (2002) An investigation into search engines as a form of targeted advert delivery. Proceedings of the 2002 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology, pp. 258-258.
[39] Wiebe, J., Riloff, E., (2005) Creating subjective and objective sentence classifiers from unannotated texts. Proceeding of CICLing, International Conference on Intelligent Text Processing and Computational Linguistics. Vol. 3406 of LNCS. Springer-Verlag, pp. 475-486.
[40] Wilson, T., Wiebe, J. & Hoffmann, P. (2005) Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proceedings of the conference on HLT/ EMNLP, pp. 347-354.
[41] Wong, W. & W Liu, M. B. (2006) Featureless similarities for terms clustering using tree-traversing ants. Proceedings of the 2006 international symposium on Practical cognitive agents and robots, pp. 177-191.
[42] Woon, W. L. & Madnick2, S. (2009). Asymmetric information distances for automated taxonomy construction. Knowl Inf Syst Int J (KAIS). Doi: 10.1007/s10115-10009-10203-10115.
[43] Yu, H. & Hatzivassiloglou, V. (2003) Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. Proceedings of the 2003 conference on EMNLP, pp. 129-136.
[44] Zang, W., Yu, C. & Meng, W. (2007) Opinion Retrieval from Blogs. Proceedings of the 16th ACM conference on CIKM, pp. 831-840.
[45] Zhuang, L., Jing, F. & Zhu, X. (2006) Movie review mining and summarization. Proceedings of the 15th ACM international conference on Information and knowledge management, pp. 43-50.

Author biographies

Teng-Kai Fan is a Ph.D. student in Computer Science and Information Engineering department at National Central University in Taiwan. He received his M.S. in Management Information System from National Kaohsiung First University of Science and Technology, Taiwan in 2005. His research interests include information extraction, text analysis, knowledge discovery from databases, and machine learning.

Chia-Hui Chang is an associate professor at National Central University in Taiwan. She received her B.S. in Computer Science and Information Engineering from National Taiwan University, Taiwan in 1993 and Ph.D. in the same department in Jan. 1999. Her research interests include Web information integration, knowledge discovery from databases, machine learning, and data mining.