Blogger-Centric Contextual Advertising

Teng-Kai Fan Chia-Hui Chang
Department of Computer Science and Information Engineering,
National Central University, Chung-Li, Taiwan 320, ROC
tengkaifan@gmail.com, chia@csie.ncu.edu.tw

ABSTRACT

This paper addresses the concept of *Blogger-Centric Contextual Advertising*, which refers to the assignment of personal ads to any blog page, chosen in according to bloggers' interests. As blogs become a platform for expressing personal opinions, they naturally contain various kinds of statements, including facts, comments and statements about personal interests, of both a positive and negative nature. To extend the concept behind the Long Tail theory in contextual advertising, we argue that web bloggers, as the constant visitors of their own blog sites, could be potential consumers who will respond to ads on their own blogs. Hence, in this paper, we propose using text mining techniques to discover bloggers' immediate personal interests in order to improve online contextual advertising. The proposed **BCCA** (**B**logger-**C**entric **C**ontextual **A**dvertising) framework aims to combine contextual advertising matching with text mining in order to select ads that are related to personal interests as revealed in a blog and rank them according to their relevance. We validate our approach experimentally using a set of data that includes both real ads and actual blog pages. The results indicate that our proposed method could effectively identify those ads that are positively-correlated with a blogger's personal interests.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning. H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval.

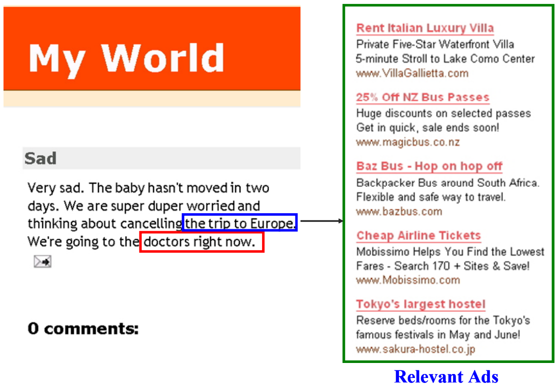
General Terms Measurements, Algorithms, Experiments.

Keywords Web Advertising, Sentiment Detection, Marketing, Blog, Intention Recognition, Machine Learning.

1. INTRODUCTION

Contextual advertising is based on studies that show that 80% of Internet users are interested in receiving personalized content on sites they visit [4]. Since the topic of a page somehow reflects the interest of visitors, ads delivered to visitors should depend upon page content rather than upon stereotypes created according to their geographical locations or upon other demographic features, such as gender or age [7]. As shown in previous studies, strong relevance increases the number of click-throughs [1] [3] [10] [13]. Some studies [6] [14] have also demonstrated that focusing on relevant topics written with positive sentiment produces high click-through rates. Although a page-relevant topic is a way to capture visitors' interest, there is no other way to determine their personal interests. However, since bloggers are constant visitors to their own blogs, and their interests or intensions are well expressed in the weblogs, an ad agency could use those expressed intentions to place interest-oriented ads. For example, Figure 1 shows a weblog with five ads related to *traveling* placed on the right. Since the content of this page describes reasons for cancelling a trip, these traveling ads are unlikely to be clicked. Instead, what the blogger needs is medical information or information on doctors. The point here is that an ad agency system could assign relevant ads according to bloggers' own interests, especially their immediate interest or intentions, for targeting ads, thus treating bloggers as the main visitors of their own blogs. To this end, even if an ad is related to the content of a linking page, an ad agency should preferentially consider the immediate interests expressed on the page (i.e., intentions) for placing ads.

In this paper, we proposed an ad matching mechanism, which we refer to as Blogger-Centric Contextual Advertising (BCCA), and which is based on latent interest detection to associate ads with blog pages. Instead of the traditional placement of relevant ads, BCCA emphasizes that the ad agency's system should provide relevant ads that are related to different levels of personal preferences in order to increase clicks. To evaluate our proposed method, we used a real-word collection comprised of ads and blog pages from Google AdSense and Google's Blog-Search Engine, respectively. Our results show that the proposed approach, based on text mining can effectively recognize the latent interests (e.g., intentions) in a blog page, or the personal interests of the blogger. In addition, we further investigated the effects of ad page matching using an ad Click-Through Rate (CTR) experiment, and our results suggest that our proposed method can effectively match relevant ads to a given blog page.



Sad

Very sad. The baby hasn't moved in two days. We are super duper worried and thinking about cancelling the trip to Europe. We're going to the doctors right now.

0 comments:

Rent Italian Luxury Villa
Private Five-Star Waterfront Villa
5-minute Stroll to Lake Como Center
www.VillaGallietta.com

25% Off NZ Bus Passes
Huge discounts on selected passes
Get in quick, sale ends soon!
www.magicbus.co.nz

Baz Bus - Hop on hop off
Backpacker Bus around South Africa.
Flexible and safe way to travel.
www.bazbus.com

Cheap Airline Tickets
Mobissimo Helps You Find the Lowest
Fares - Search 170 + Sites & Save!
www.Mobissimo.com

Tokyo's largest hostel
Reserve beds/rooms for the Tokyo's
famous festivals in May and June!
www.sakura-hostel.co.jp

Figure 1: Example of a blog page with ads

*CIKM'*09, November 2–6, 2009, Hong Kong, China.
Copyright 2009 ACM 978-1-60558-512-3/09/11...$10.00.

2. BCCA FRAMEWORK

Traditional contextual advertising processes a given page a user visits to find related topics for matching ads, while blogger-centric contextual advertising would assign ads to a given blog page in accordance with the blogger's interests. Before demonstrating the proposed framework, we will explain how bloggers' personal interests could be obtained. Generally speaking, an individual blog often contains profile information, tags and posts, which we could classify as indicating different levels of interest.

Profile-level: Blog service providers (e.g., BlogSpot.com, Technorati.com) usually ask bloggers' to enter interests to build their profiles (e.g., music, reading) when they register as a member of the service. In addition to generic interests collected at registration time, the tags on posts or the archive of past posts can be used to construct bloggers' profiles, showing their specific interests. Since these kinds of interest continue for a period of time, we view them as long-term interests. In this paper, we assumed that each blog-site has an interest profile containing either generic or specific long-term interests.

Triggering-level: Blog posts are media in which bloggers express their opinions and interests, as well as their intensions. For example, the sentence, "The Nokia N95 is a good cell phone for several reasons," express the sentiments of the author toward the object in question, a Nokia N95. For blogger-centric contextual advertising, we argue that recognizing the intentions of authors could be even more effective. For instance, the sentence, "*We're going to the doctor right now*," in Figure 1 indicates that the author has an immediate intention to see a doctor. As another example, the sentence, "*I am looking for a new laptop*," implies that the author probably will purchase a laptop. As such, targets are immediate interests; ads centered around them might increase click-through.

In BCCA framework, the advertising system analyzes the content of the page to recognize intention and detect sentiment for triggering-level interests. If no such targets are found, the system uses targets from the blogger's profile and searches the ad database to find the best matching ads. These four modules: intention recognition, sentiment detection, term expansion and target-ad matching are the main components in our BCCA framework. The first two components analyze triggering-level interests, while the last two components enhance the target-ad matching procedure. Note that we place a priority on ad assignment based on different levels of interests. That is, triggering-level interests (i.e., immediate interests) have higher priority than profile-level interests (long-term interests). Thus, the sentiment detection module is invoked only when no positive intention is detected. If neither module detects intention or positive sentences, the system should then use targets from the blogger's profile.

Next, the system proceeds to term expansion and the target-ad matching agency. Due to the short form of ads and targets, we designed a term expansion component to enhance the likelihood of intersection between targets and available ads. Finally, a retrieval function based on a query likelihood language model is deployed for the target-ad matching strategy to rank the ads.

2.1 Intention Recognition

Given a triggering page, our aim in this section is to explain the process of recognizing whether there exist any intention-bearing sentences. By modeling this problem one of classification, our job here is the preparation of training sentences which must be labeled as intentional or non-intentional.

Labeling each sentence as intentional or non-intentional is a time-consuming and costly task. In this study, we propose a novel technique to semi-automatically label training data for this task. Since many auction web sites (e.g., *ebay.com*, *yahoo.com*) provide special forums for buyers to post their needs, it naturally becomes a source of intention-filled sentences. In this study, we collected a large set of posts containing buyers' requirements from *ebay.com* [1] . However, since many buyers describe their requirements (e.g., product name, product type) in a concise sentence, the labeling system filters simple sentences that do not contain intentions by the following criteria.

Part-of-Speech (*POS*) *tag*: Since an intentional sentence usually contains a verb that is not a form of "to be", we keep candidate sentences that contain terms tagged as verbs, and remove sentences that contain only noun phrases.

The length of the sentence: short polite words are usually used in the forums. Hence, we simply disregard sentences whose lengths are less than three words.

All the candidate sentences that conform to the above rules are regarded as intentional data. As for non-intentional training data, we manually construct some queries that will collect entry pages from Wikipedia[2].

For feature selection, we considered a subset of word unigrams chosen via the chi-square feature selection. In this study, we selected the top-*K* features with a high chi-square value as input features. To apply these machine learning algorithms on our dataset, we used the standard bag-of-features framework.

2.2 Sentiment Detection

Our aim in this section is to apply a contextual sentiment technique for detecting the sentiment of a blog page. 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. For sentiment detection, a good Web resource is *epinions.com*, where pros and cons of products or other topics are discussed by reviewers. Such information is thus used by Kim and Hovy to prepare their training data [8]. Although this method is fully automatic, such an idea based on pro and con phrases does not perform well (F-measure: 0.65). As indicated in [5], opinionated content is most often carried by parts of speech used as modifiers (i.e., adverbs and adjectives) rather than parts of speech used as heads (i.e., verbs, nouns). Thus, we combined the concept proposed in [5] and modified Kim & Hovy's method as follows. For each term tagged as adjective or adverb (e.g., JJ, RB) in pro and con sets, our labeling system checks each sentence to find sentences that

[1] http://pages.ebay.co.uk/wantitnow/

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

contain those adjectives or adverbs. Then the system annotates each sentence with appropriate sentiment labels. As for feature selection, we similarly adopted the chi-square statistic method to select the top-*K* features for both the sentiment identification and sentiment classification steps. For the identification step, all the subjective and objective sentences are represented as feature-presence vectors, where the presence or absence of each feature is recorded. For the classification step, all the positive and negative sentences are similarly represented as feature-presence vectors to train the classifiers.

2.3 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. In this paper, we followed three term-expansion methods proposed in [6] to expand the specific terms in a blog page.

2.4 Page-Ad Matching

We can regard the blogger-centric 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*** as a user's query ***q*** and corresponding documents ***d***, respectively. In this paper, the query likelihood language model is adopted as our ad retrieval model. The language modeling approach to IR models the following idea: A document ***d*** is a good match to a query ***q*** if the document model is likely to generate query ***q***, which will in turn happen if the document contains the query words often [11]. Hence, we constructed from each ad ***a*** in the collection a language model M_a. Our goal is to rank ads by $P(a|q)$, where the probability of an ad ***a*** is interpreted as the likelihood that it is relevant to the query ***q***.

3. EXPERIMENTAL RESULTS

In this section, we focus on our three experiments: intention recognition, sentiment detection and page-ad matching. We begin by describing the dataset, and we then proceed to a discussion of the experimental results.

3.1 Datasets

To evaluate intention recognition performance, we retrieved data from *ebay.com* and *wikipedia.com* and then used these as our training dataset for building our learning classifier model. For sufficient detail, we first collected 36,151 buyer's posts and then randomly examined 10,000 sentences. For non-intentional training data, we similarly examined 10,000 sentences from the Wikipedia dataset. Additionally, to evaluate sentiment detection performance, we collected data from *epinions.com* and used this as our training dataset for building learning classifier models. We gathered many different types of reviews from *epinion.com*. For sufficient detail, we examined about 30,000 reviews (about 900,000 sentences), with an average number of 30 sentences per review document. For page-ad matching, we collected a total of 138,907 ads. Our triggering page collection was comprised of 200 blog pages on various topics. It included a range of opinions and was comprised of various subjective articles (100 positive and 50 negative articles) and 50 general articles including intentional sentences.

3.2 Evaluation of Intention Recognition

For the dataset of triggering pages (i.e., 50 intentional articles), owing to a lack of training data, we chose SVM learning algorithms with different feature sets to train the models on the *ebay.com* and Wikipedia datasets. We then applied these models to our triggering page dataset. The results are shown in Table 1. SVM with features selected by chi-square gave the better results, yielding an average 94.2% F-measure for the non-intentional class and an average 80.5% F-measure for the intentional class.

3.3 Evaluation of Sentiment Detection

The goals of this section are similar to [6] [8]: the target is to investigate how well the trained model performed on a different data source (200 triggering pages). We compared the modified method to the one (regarded as the baseline system) used in [6].

We chose SVM 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. We conducted the sentiment identification and classification step and the results are shown in Table 2. For the sentiment identification task, the SVM with features selected by chi-square gave the better results, yielding an average 87.1% F-measure in the objective class and an average 69.9% F-measure in the subjective class. For the sentiment classification task, the best F-measure for the positive class (74.8%) and the best F-measure for the negative class (59.1%) are generated by the SVM with features selected by chi-square and the SVM with unigrams respectively.

Table 1: Intension Recognition by various feature sets

Feature Set (# of features)	Class	Pre. (%)	Recall (%)	F-measure (%)
Unigram (45030)	Non-Intentional	94.5	93.5	94.0
	Intentional	77.8	80.9	79.4
Chi-Square (10000)	Non-Intentional	95.5	92.9	**94.2**
	Intentional	77.1	84.2	**80.5**

Table 2: Sentiment Detection for triggering pages

Feature Set	Task	Class	Prec. (%)	Recall (%)	F-measure (%)
Uni-grams	Identification	Obj.	81.1	88.9	84.8
		Sub.	75.8	62.8	67.7
	Classification	Neg.	80.6	46.6	**59.1**
		Pos.	70.0	65.0	67.4
Chi-Square	Identification	Obj.	80.5	95.0	**87.1**
		Sub.	86.6	58.8	**69.9**
	Classification	Neg.	63.8	52.6	57.6
		Pos.	96.0	61.3	**74.8**
Baseline	Identification	Obj.	78.5	64.5	70.8
		Sub.	50.1	67.7	58.2
	Classification	Neg.	56.1	53.1	54.6
		Pos.	56.2	71.1	68.2

Table 3: Click-Through Rate (CTR) for triggering pages

Dataset	Method	CTR
Positive dataset	BCCA	52%
	PCA	**53%**
	Google AdSense	47%
Intention dataset	BCCA	**58%**
	PCA	32%
	Google AdSense	42%
All triggering pages	BCCA	**57%**
	PCA	42%
	Google AdSense	40%

3.4 Evaluation of Page-Ad Matching

The goal of this section is to investigate to what extent ad placements are actually related to personal interests evident on the triggering pages. To evaluate our page-ad matching framework, we compared the top-five ranked ads provided by three different ranking methods: our proposed approach with a personalization mechanism (i.e., **BCCA**), our proposed approach without a personalization mechanism (i.e., **PCA** (Plain **C**ontextual **A**dvertising) excluding intention recognition and sentiment detection), and Google AdSense. Here we used the SVM with features selected by chi-square method as intention and sentiment models according to experimental results. Since it is difficult to invite the original authors of blog actually to participate in an Ad Click-Through-Rate (CTR) experiment, 14 volunteers participated in our CTR experiment. All the advertisements in each pool were manually clicked by volunteers. To compare with Google AdSense, we only measured the CTR for this experiment. The measure is the fraction of retrieved ads that are clicked. Table 3 shows the results for three page-ad matching methods. In the case of the positive sentiment dataset, there are no significant differences among the three page-ad matching approaches. For the intention dataset, our results show that the proposed BCCA approach can yield a better performance (58%) than other approaches (32% for PCA approach and 42% for Google Adsense). Regarding all triggering pages, our results show that the proposed BCCA approach can yield a better performance (57%) than other approaches (42% for PCA approach and 40% for Google AdSense). According to Table 3, these results lead us to the conclusion that our BCCA framework can place ads that are related to the personal interest content of triggering pages.

4. RELATED WORK

Several studies pertaining to advertising research have stressed the importance of relevant associations for consumers [13] and how irrelevant ads can turn off users and relevant ads are more likely to be clicked [3]. As for contextual advertising, Ribeiro-Neto *et al*. [12] proposed a number of strategies for matching pages to ads based on extracted keywords. To identify the important parts of the ad, the authors explored the use of different ad sections as a basis for the ad vector. In follow-up research [9], the authors proposed a method to learn the impact of individual features using genetic programming to generate a matching function. However, due to the vagaries of phrase extraction, approaches based on "bid phrases" leads to many irrelevant ads. To overcome this problem, Broder *et al*. [2] proposed a system for contextual ad matching based on a combination of semantic and syntactic features. The results show that the semantic-syntactic approach outperformed the syntactic approach.

Prior studies usually focused on the interests of end-users to assign advertisements. There have been very few studies that tried to deliver personal ads geared towards publishers' interests. Besides, even if several prior studies have proven that relevance has a definite impact on contextual advertising and on the proposed effective ranking function that matches ads with pages, they neglect to consider personal intentions in the assignment of relevant ads. In this study, we further investigate the importance of personal interests mining, including intention and sentiment analysis for improved contextual advertising.

5. CONCLUSION

In this study, we proposed and evaluated a novel framework for associating ads with blog pages based on intention analysis. Prior work to date has only examined the extent of content relevance between pages and ads. In this paper, we investigated the intentions and the sentiments of blog pages and utilized this information to demonstrate blogger centric contextual advertising. For intention recognition, we adopted the information from *ebay.com* and Wikipedia as training data to recognize the latent immediate interest. As for sentiment detection, we proposed a new method to label the sentences from *epinions.com* for training data preparation. For page-ad matching, we evaluated our framework using 200 personal blog pages and over 130,000 ads sampled from Google AdSense. We compared BCCA with Google AdSense and found that our proposed method with personalized advertising mechanism can achieve superior performance (57% accuracy).

6. REFERENCES

[1] A. Anagnostopoulous, A. Z. Broder, E. Gabrilovich, V. Josifovski and L. Riedei. Just-in-Time Contextual Advertising. In CIKM'07, page. 331-340, 2007.

[2] A. Broder, M. Fontoura, V. Josifovski and L. Riedel. A semantic approach to contextual advertising. In SIGIR'07, page 559-566, 2007.

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

[4] ChoiceSteam, ChoiceStream personalization survey: consumer trends and perceptions, http://www.choicestream.com/pdf/ChoiceStream_PersonalizationSurveyResults2005.pdf, 2005.

[5] A. Esuli and F. Sebastiani. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In CLRE'06, page. 417-422, 2006.

[6] T.-K. Fan and C.-H. Chang. Sentiment-Oriented Contextual Advertising. ECIR, LNCS 5478, page 202-215, 2009.

[7] P. Kazienko and M. Adamski. AdROSA-Adaptive personalization of web advertising. Information Sciences 177, 2269-2295, 2007.

[8] S.-M. Kim and E. Hovy. Automatic Identification of Pro and Con Reasons in Online Reviews. In ACL'06, page 483-490, 2006.

[9] A. Lacerda, M. Cristo, M. A. Goncalves, W. g. Fan, N. Ziviani and B. Ribeiro-Neto. Learning to advertise. In SIGIR06, page 549-556, 2006.

[10] OneUpWeb. How keyword length affects conversion rates, http://www.oneupweb.com/landing/keywordstudy_landing.htm.

[11] M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In SIGIR'98, page 275-281, 1998.

[12] B. Ribeiro-Neto, M. Cristo, P. B. Golgher and E. S. d. Moura. Impedance coupling in content-targeted advertising. In SIGIR'05 page 496-503, 2005.

[13] C. Wang, P. Zhang, R. Choi and M. D'Eredita. Understanding consumers attitude toward advertising. In Eighth Americas Conference on Information Systems, page 1143-1148, 2002.

[14] Y. Zhang, A. C. Surendran, J. C. Platt and M. Narasimhan. Learning from multi-topic web documents for contextual advertisement. In SIGKDD'08, page 1051-1059, 2008.