2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

Matching and Ranking with Hidden Topics towards Online Contextual Advertising

Dieu-Thu Le, Cam-Tu Nguyen, Quang-Thuy Ha
Coltech, Vietnam National University
144 Xuan Thuy, Cau Giay, Hanoi, Vietnam
{dieuthu, ncamtu, thuyhq}@vnu.edu.vn

Xuan-Hieu Phan, Susumu Horiguchi
GSIS, Tohoku University
Aobayama 6-3-09, Sendai, 980-8579, JAPAN
{hieuxuan, susumu}@ecei.tohoku.ac.jp

Abstract

In online contextual advertising, ad messages are displayed related to the content of the target Web page. It leads to the problem in information retrieval community: how to select the most relevant ad messages given the content of a page. To deal with this problem, we propose a framework that takes advantage of large scale external datasets. This framework provides a mechanism to discover the sematic relations between Web pages and ad messages by analyzing topics for them. This helps overcome the problem of mismatch due to unimportant words and the difference in vocabularies between Web pages and ad messages. The framework has been evaluated through a number of experiments. It shows a significant improvement in accuracy over word/lexicon-based matching and ranking methods.

1. Introduction

In contextual advertising, ad messages are delivered based on the content of the Web pages that users are surfing. It can therefore provide Internet users with information they are interested in and allow advertisers to reach their target customers in a non-intrusive way [3, 4]. In order to suggest the "right" ad messages, contextual ad matching and ranking techniques are needed to be used. This has posed new challenges to Web mining and IR researcher. Firstly, as words can have multiple meanings and some words in the target page are not important, they can lead to mismatch in lexicon-based matching method. Moreover, a target page and an ad can still be a good match when they share no common words or terms but belong to the same topic.

To deal with these problems, we present a framework that can discover the semantic relatedness between Web pages and ads by analyzing *implicit* or *hidden* topics for them. After that, both Web pages and advertisements are expanded with their most relevant topics, which helps reduce the sparseness and make the data more topic-focused. The framework can therefore overcome the limitation of word choices, deal with a wide range of Web pages and ads, as well as process future data, that is, previously unseen ads and Web pages, better. It is also easy to implement and general enough to be applied in different domains of advertising and in different languages.

2. Related Work

The success of sponsored search in Web advertising has motivated IR researchers to study content match in contextual advertising. Thus, one of the earliest studies in this area was originated from the idea of extracting keywords from Web pages [12]. Considering both semantic and syntactic feature, Andrei Broder et al. [2] proposed a framework for matching ads using a large taxonomy. Our framework also tries to discover the semantic relations of Web pages and ads, but instead of using a classifier with a large taxonomy, we use hidden topics discovered automatically from an external dataset. It does not require any language-specific resources, but simply takes advantage of a large collection of data, which can be easily gathered on the Internet.

To overcome the difference between the vocabularies of Web pages and ads, Ribeiro-Neto et al. [11] tried to use additional pages by means of a Bayesian model. It is similar to ours in the idea of expanding Web pages with external terms to decrease the distinction between their vocabularies.

Hidden topics discovered from data using latent topic analysis models, such as pLSA [7] and LDA [1], can be useful in many applications. An example of them is using LDA to build classifiers that deal with short & sparse data [10]. Our framework is also based on hidden topics discovered from a large-scale dataset to capture the semantic relations between Web pages and ads.

3. Page-Ad Matching and Ranking Framework

Our general framework for contextual page-ad matching and ranking with hidden topics discovered from external large-scale data collections is illustrated in figure 1.

Given a set of n target Web pages $\mathbf{P} = \{p_1, p_2, \ldots, p_n\}$, and a set of m ad messages (ads) $\mathbf{A} = \{a_1, a_2, \ldots, a_m\}$.

978-0-7695-3496-1/08 $25.00 © 2008 IEEE
DOI 10.1109/WIIAT.2008.180

For each Web page p_i, we need to find a corresponding ranking list of ads: $\mathbf{A}_i = \{a_{i1}, a_{i2}, \ldots, a_{im}\}, i \in 1..n$ such that more relevant ads will be placed higher in the list. These ads are ranked based on their relevance to the target page and the keyword bid information. However, in the scope of our work, we only take linguistic relevance into consideration and assume that all ads have the same priority, i.e. the same bid amount.

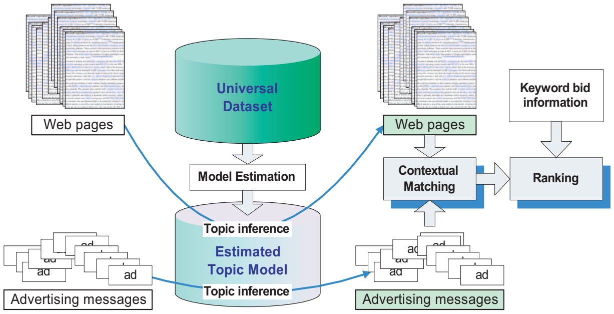

Figure 1. A General Framework for Page-Ad Matching & Ranking with Hidden Topics

(a) *Choosing an appropriate "universal dataset"*
(b) *Doing topic analysis for the universal dataset*
(c) *Doing topic inference for Web pages and ads*
(d) *Page-Ad Matching and Ranking*

As depicted in Figure 1, the first important thing to consider in this framework is collecting an appropriate external large-scale document collection (a) which is called Universal Dataset. The dataset must be large enough and its vocabulary must be consistent with that of Web pages and ads. It will make sure topics analyzed from this data can overcome the vocabulary difference of Web pages and ads.

In general, we can apply any topic models, such as pLSI [7] or LDA [1], to analyze the Universal Dataset. The result of the step (b) is an estimated topic model that includes hidden topics discovered from the dataset and the distributions of topics over terms. Steps (a) and (b) will be presented more details in section 4. After the estimating process (b), we can again do topic inference for both Web pages and ads based on this model to discover their meanings and topic focus (c). This information will be integrated into the corresponding Web pages or ads for matching and ranking (d). Both steps (c) and (d) will be discussed more in section 5.

4. Hidden Topic Analysis of Universal Dataset

Our framework is largely inspired by the recent success of topic modeling research in Machine Learning and NLP. Topic models [1, 5, 7] allow discovering semantic information from data based on the idea that each document is a probability distribution over topics and each topic, in turn, is a mixture distribution over words/terms.

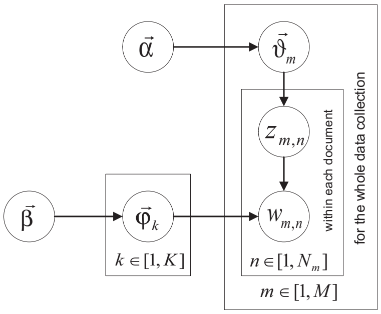

Figure 2. Graphical structure of LDA

4.1. Latent Dirichlet Allocation (LDA)

LDA is a generative graphical model introduced by Blei et al. [1]. Its graphical structure is shown in Figure 2. It is based upon the idea that a document is a mixture of topics. It can be modeled as a process of generating new documents. First, to make a new document, one can choose a topic distribution for the document, that means the document is composed of different topics with different distribution. Then, in order to generate words for the document, one can choose some words randomly based upon the distribution of words over those chosen topics.

To estimate LDA, some approximate methods can be used, such as Variational Methods [1] and Gibbs Sampling [5]. Gibbs Sampling is a special case of Markov-chain Monte Carlo (MCMC) and often yields relatively simple algorithms for approximate inference in high-dimensional models such as LDA [6].

4.2. News Collection as Universal Dataset

4.2.1 Data preparation

With the purpose of using a large scale dataset for Vietnamese contextual advertising, we chose VnExpress, one of the highest ranking e-newspaper in Vietnam[1] as the Universal Dataset. The data was preprocessed using JVnTextPro[2] before being analyzed (Table 1).

Table 1. VnExpress as Universal Dataset

After removing html, doing sentence and word segmentation:
$size \approx 219M$, $\lvert docs \rvert = 40,328$
After filtering and removing non-topic oriented words:
$size \approx 53M$, $\lvert docs \rvert = 40,268$ $\lvert words \rvert = 5,512,251$; $\lvert vocabulary \rvert = 128,768$

4.2.2 Topic Analysis of VnExpress Collection

After preprocessing, the dataset was analyzed using GibbsLDA++[3]. We carried out topic analysis of three models:

[1] VNExpress e-newspaper: http://VnExpress.net/
[2] JVnTextpro: http://jvnsegmenter.sourceforge.net/
[3] GibbsLDA++: http://gibbslda.sourceforge.net

60, 120 and 200-topic, respectively. Some samples of hidden topics of 200-topic model are illustrated in Figure 3. The full lists of topics are available online[4].

Topic 3	Topic 15	Topic 44
bác_sĩ (doctor)	**thời_trang** (fashion)	**thiết_bị** (equipment)
bệnh_viện (hospital)	**người_mẫu** (model)	**sản_phẩm** (product)
thuốc (medicine)	**mặc** (wear)	**máy** (machine)
bệnh (disease)	**trang_phục** (clothes)	**màn_hình** (screen)
phẫu_thuật (surgery)	**thiết_kế** (design)	**công_nghệ** (technology)
điều_trị (treatment)	**đẹp** (beautiful)	**điện_thoại** (telephone)
bệnh_nhân (patient)	**váy** (dress)	**hãng** (company)
y_tế (medical)	**sưu_tập** (collection)	**sử_dụng** (use)
ung_thư (cancer)	**mang** (wear)	**thị_trường** (market)
tình_trạng (condition)	**phong_cách** (style)	**usd** (USD)
cơ_thể (body)	**quần_áo** (costume)	**pin** (battery)
sức_khoẻ (health)	**nổi_tiếng** (famous)	**cho_phép** (allow)
đau (hurt)	**quần** (trousers)	**samsung** (Samsung)
gây (cause)	**trình_diễn** (perform)	**di_động** (mobile)
khám (examine)	**thích** (like)	**sony** (Sony)
kết_quả (result)	**quyến_rũ** (charming)	**nhạc** (music)
căn_bệnh (illness)	**sang_trọng** (luxurious)	**máy_tính** (computer)
nặng (serious)	**vẻ_đẹp** (beauty)	**hỗ_trợ** (support)
cho_biết (inform)	**gái** (girl)	**điện_tử** (electronic)
máu (blood)	**gương_mặt** (figure)	**tính_năng** (feature)

Figure 3. Sample hidden topics and their most likely words from 200-topic model

5. Matching and Ranking with Hidden Topics

5.1. Topic Inference for Ads & Target Pages

Given an estimated LDA model as described in the previous section, we can now do topic inference for web pages and ad messages by a similar sampling procedure. In particular, we have a set of web pages and ad messages $\underline{W}$. Topic inference process will discover the probability distribution of topics over each document in $\underline{W}$.

After performing topic sampling, the topic distribution of a new document $\overrightarrow{w_{\underline{m}}}$ is $\overrightarrow{\vartheta_{\underline{m}}} = \{\vartheta_{\underline{m},1}, \ldots, \vartheta_{\underline{m},k}, \ldots, \vartheta_{\underline{m},K}\}$. Topics that have high probability $\vartheta_{\underline{m},k}$ will be added to the corresponding Web page/ad $\underline{m}$ in order to make the data more topic-focused and reduce the vocabulary difference between Web pages and ads. Each topic integrated into a Web page/ad will be treated as an *external term* and its frequency is determined by its probability value. Technically, the number of times a topic k is added to a Web page/ad $\underline{m}$ is decided by two parameters *cut-off* and *scale*:

$$\mathit{Frequency}_{\underline{m},k} = \begin{cases} \mathit{round}\,(\mathit{scale} \times \vartheta_{\underline{m},k})\,, \text{if } \vartheta_{\underline{m},k} \geq \mathit{cut\text{-}off} \\ 0, \text{if } \vartheta_{\underline{m},k} < \mathit{cut\text{-}off} \end{cases}$$

where *cut-off* is the topic probability threshold. *scale* is a parameter that determines the frequency of each topic added.

5.2. Matching and Ranking

After being enriched with hidden topics, Web pages and ads will be matched based on their cosine similarity. For each page, ad messages will be sorted in order of its similarity to the page. The ultimate ranking function will also take into account the keyword bid information. But this is beyond the scope of this paper.

6. Evaluation

6.1. Experimental Data

• For Web pages, we chose 100 pages randomly from a set of 27,763 pages crawled from VnExpress e-newspapers (exclusive from the Universal Dataset) from different topics: Food, Shopping, Cosmetics, Mom & children, Estate, Stock, Jobs, Law, etc.
• For advertisements, we collected 3,982 ad messages from Zing[5]. Each ad message is composed of four parts: title, Website's URL, its description and keywords.

6.2. Parameters & Evaluation Metrics

First, we implemented two retrieval baselines AD and AD_KW following the settings used in [11]. Then, to evaluate the contribution of hidden topics, we carried out six different experiments, which are called HT strategies with *cutoff* = 0.05. The six matching experiments using hidden topics are called HTx_y, where x stands for the number of hidden topics of the used estimated model and y is the $scale$ (Table 2).

Methods		Description
Without Hidden Topics	AD	Use only title, description of ads
	AD_KW	Use title, description and keywords of ads
With Hidden Topics	HT60_10	number of topics = 60, $scale$ = 10
	HT60_20	number of topics = 60, $scale$ = 20
	HT120_10	number of topics = 120, $scale$ = 10
	HT120_20	number of topics = 120, $scale$ = 20
	HT200_10	number of topics = 200, $scale$ = 10
	HT200_20	number of topics = 200, $scale$ = 20

Table 2. Description of 8 experiments

To evaluate the performance of the matching method using retrieval information (term frequencies) only and the matching method using hidden topics, we prepared the test data with the same methodology used in [11]. First, we started by matching each Web page to all the ad messages and ranking them to their similarities. To determine the precision of each method and compare them, we selected top four ranked ads of each method and selected from them most relevant ads and excluded irrelevant ones. In average, the number of corrected advertisements for each Web page is 6.51 ads eventually.

6.3. Experimental Result & Analysis

As illustrated in Figure 4, using hidden topics significantly improves the performance of the whole framework.

[4] The full lists of hidden topics available at: gibbslda.sourceforge.net/vnexpress-060topics.txt; vnexpress-120topics.txt; vnexpress-200topics.txt

[5] Vietnamese Zing directory: http://directory.zing.vn/directory

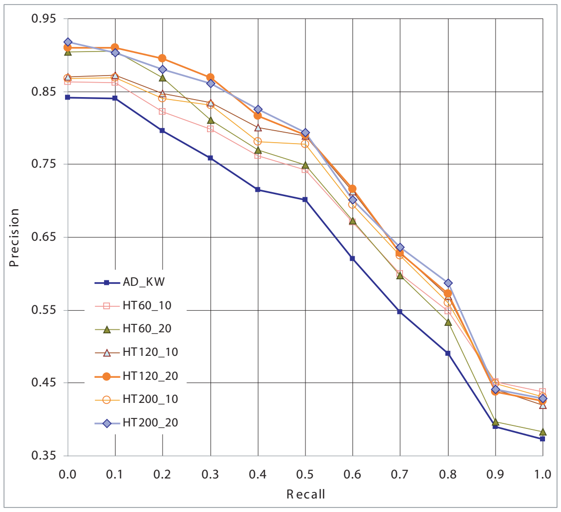

Figure 4. Precision-recall curves of matching & ranking without and with hidden topics

It increases the precision in average from 64% to 72% and reduces almost 23% error (HT200_20) (Table 3).

For the overall methods, we also calculated the number of corrected ad messages found in the first, second and third position of the rank lists proposed by each strategy (#1, #2, #3 in table 3). It also reflects the precision of our hidden-topic methods higher than that of the baseline matching method. Moreover, the precision at position 1 (#1) is generally higher than that of position 2 and 3 (#2, #3). If the system is ranking the relevant ads near the top of the rank list, it is possible that the system can suggest most appropriate ads for the corresponding page. It therefore partially shows the effectiveness of the ranking system.

Table 3. 11-points average precision

Methods	Correct ads found				11-point avg.
	#1	#2	#3	Totals	precision
AD	70	56	52	178	**49.86%**
AD_KW	78	69	64	211	**64.32%**
HT60_10	79	76	70	225	68.72%
HT60_20	86	75	67	228	69.02%
HT120_10	82	74	74	230	70.76%
HT120_20	89	79	69	237	**72.47%**
HT200_10	79	77	77	233	70.26%
HT200_20	88	78	79	245	**72.50%**

Finally, we also quantified the effect of the number of topics and its added amount to each Web page and ad by testing with different topic models and adjusting the scale value. However, it shows that the number of hidden topics does not cause a significant change in overall performance. As indicated in table 3, the performance of 120 and 200-topic models yield a better result than 60-topic model. However, there is no considerable change between 120-topic and 200-topic models, also in the quantities of added topics to each page and ad message. It can therefore conclude that the number of topics should be large enough to discriminate the difference of terms to better analyze topics for Web pages and ads. And since the number of topics is large enough, the performance of the overall system is quite stable.

7. Conclusions

In this work, we presented a framework to choose the most relevant advertisements for a Web page by taking advantage of hidden topics discovered from a large dataset.

This framework has shown its efficiency through a variety of experiments against the basic method using syntactic information only. In practical, the results record an error reduction of 22.9 % in the method using 200-topic model over the normal matching strategy without hidden topics. Finally, the framework is also flexible and general enough to be applied in a multilingual environment.

Acknowledgements

This work is partly supported by the Project QC.07.06 “Vietnamese named entity resolution & tracking crossover Web documents”, Vietnam National University, Hanoi and National Project 02.2006.HĐ - ĐTCT-KC.01/06-10 “Developing content filter systems to support management & implementation public security”.

References

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] A. Broder, M. Fontoura, V. Josifovski, and L. Riedel. A semantic approach to contextual ad. *Proc. ACM SIGIR*, 2007.

[3] P. Chatterjee, D. L. Hoffman, and T. P. Novak. Modeling the clickstream: Implications for web-based advertising efforts. *Marketing Science*, 22(4):520–541, 2003.

[4] R. Wang, P. Zhang, and M. Eredita. Understanding consumers attitude toward advertising. *Proc. AMCIS*, 2002.

[5] T. Griffiths and M. Steyvers. Finding scientific topics. *Proc. National Academy of Science*, 101:5228–5235, 2004.

[6] G. Heinrich. *Parameter estimation for text analysis*. TR.

[7] T. Hofmann. Probabilistic LSA. *Proc. UAI*, 1999.

[8] A. Lacerda, M. Cristo, M. Gonçalves, W. Fan, N. Ziviani, and B. Neto. Learning to advertise. *Proc. ACM SIGIR*, 2006.

[9] C. Manning, P. Raghavan, and H. Schutze. *Introduction to information retrieval*. Cambridge University Press, 2008.

[10] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. *Proc. WWW*, 2008.

[11] B. Neto, M. Cristo, P. Golgher, and E. de Moura. Impedance coupling in content-targeted advertising. *Proc. SIGIR*, 2005.

[12] W. Yih, J. Goodman, and V. Carvalho. Finding advertising keywords on web pages. *Proc. WWW*, 2006.