Topic Cor-relevance Model within Latent Semantic Space for Contextual Advertising

Yu Zou, Chunming Wang, Lei Wang, Hao Zheng
Beijing Yahoo! Global R&D Center
22nd FL, Building C, SP Tower,
Tsinghua Science Park, Beijing, 100084
{yuzou, wangcm, wanglei, hzheng}@yahoo-inc.com

ABSTRACT

Vector-Space Model (VSM) is one of most widely used models for retrieval task. Probabilistic Latent Semantic Indexing was proposed to address issues of synonymy as well as polysemous words when retrieving documents. While those models assume relevance only originates from identical words or topics from page side and ad side. However, relevance should also exist in cases that words or topics between page and ad are not identical. The subtle relevance, called as Cor-relevance in this paper, is proven to be another strong signal of user clicks, which should be leveraged to predict click-ability more precisely for ad retrieval task. In this paper, we propose Cor-relevance Model (CRM) which is a natural extension of VSM. The CRM is able to model causation or relevance of a unit in page and another unit in ad no matter whether the two units are the same or not. Thus the model has stronger representability to reveal more subtle relevance relation than VSM.

Moreover embedding CRM into latent topic space, we propose Topic Cor-relevance Model (TCRM) within latent topic space. We may interpret the model as a prediction of click-ability for a given page and ad pair under probability framework. With the explanation, the meaning of cor-relevance between a page topic and an ad topic is P(click | page_topic, ad_topic), which gives us an insight on how relevance is connected to click-ability. We also propose a Maximum a Posterior learning method to learn cor-relevance from click/view logs automatically and even to address data sparseness issues in the learning. Our experiments showed TCRM improved AUC of P-R curve significantly than a VSM baseline model. And live traffic test also showed lift of CTR in a contextual advertising system.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Commercial Services; H.3.3 [**Information Search and Retrieval**]: Relevance models, Text mining, Relevance feedback; I.5.2 [**Design Methodology**]: Classifier design and evaluation.

General Terms

Algorithms

Keywords

Contextual advertising, sponsored search, click predication, cor-relevance model, semantic match, click-ability and relevance, vector-space model, word-pair model, maximum a posteriori

1. INTRODUCTION

Web advertising derives a large portion of revenue of Internet companies and continues to support blooming of web world. While Sponsored Search is main revenue income of search engine providers, Contextual Advertising, which refers to the placement of textual advertisements within the content of a web page, is another quickly developing revenue source of Internet business. Contextual advertising exploits traffic of sites that range from individual bloggers to large publishers such as major newspapers, and provides remarkable additional revenue besides of Sponsored Search and supports their businesses. In contextual advertising, usually there is an ad platform which is source of ads, web publishers are able to retrieval ads from the platform and post the ads on their web pages. Clicks of ad bring revenue which is shared between publisher and ad platform.

1.1 Ad retrieval models

Similar to the treatment in Sponsored Search, it is preferable to display ads related to page content to provide a better user experience. And it is assumed that better relevance also usually increases the probability of clicks and confirmed by user studies [1]. Different with Sponsored Search where query is strong signal of user intention, keywords are extracted from the page content, then the keywords are used to retrieve displaying ads. Popular approaches that estimated the ad relevance are based on co-occurrence of the same words or phrases within the ad and within the page. One of those famous approaches from this category is Vector-Space Model [2, 16].

However ads matching, as a specific document retrieval problem, is relatively difficult because ad content contains very few terms and the language of ads is sparse. It is frequent that a same concept mentioned with some terms in ad side is described with different terms in query side which is publishers' web pages in contextual advertising. Thus match based on word or phrase within page and ad leads to problems in case of synonymy or polysemous words. For example of polysemous word, 'apple' is a famous brand of digital device, while it is also a kind of fruit in another context. For example of synonymy case, pure syntactic

match also fails to find ad if only 'motel' exists in page content while advertiser bids 'cheap hotel'. Mismatching caused by issues of synonymy or polysemous words is essentially rooted from the sparse language of ads. Many researchers have tackled the problem of sparse language from different perspectives.

To solve the problem especially the synonymy case, query rewriting technology [3, 4, 5] is popular starting from Sponsored Search domain and also be applicable for contextual advertising. Query rewriting is to find frequent alternatives, also called as rewrites, for a query. Not only the original query but also several rewrites are used to retrieval related ads. Therefore, the synonymy problem can be solved partially. For contextual advertising, a data-driven word-pair mining approach was proposed in [6]. In the paper, some word-pairs by mining are also considered good indicator features of relevance or click. And their experiments showed that word-pair match outperformed same-word match a lot. However, it becomes another problem how to mine word-pairs meaningful for match task, because the amount of word-pair candidates is huge while training data are always limited due to so many long-tailed words in a language.

As a contrast of syntactic based match where the key is how to find proper transformations of a query, semantic based match was proposed in [7, 8, 10]. Several statistical topic models have been applied to many document retrieval tasks and became popular. Two popular statistical topic models are the Probabilistic Latent Semantic Analysis (PLSA) [8] and the Latent Dirichlet Allocation (LDA) [9]. Both of them map documents as well as queries to points in latent topic space. Probabilistic Latent Semantic Indexing (PLSI) was proposed by Thomas Hofmann [10]. In this paper, a method called PLSI-Q was proposed to build index for topics and evaluate similarity with VSM in low dimension latent topic space. The synonymy or polysemous issues can be addressed by match mechanisms like PLSI, because the relevance is based on semantic similarity within latent topic space. PLSA approach can be traced back to Latent Semantic Analysis (LSA) which was proposed early as in 1990 [11]. The difference is that PLSA is probabilistic while LSA is a SVD application.

1.2 Cor-relevance information

However all those models still assume relevance only originates from identical words or topics from page side and ad side. And VSM similar match is adopted to measure similarity or relevance in the models. However, relevance should also exist in cases that words or topics between page and ad are not identical. For example, if a user is reading a web page about new car, he/she is maybe also interested in GPS. Although 'new car' and 'GPS' are from two different topics, there is certain stable causality between page topic 'new car' and ad topic 'GPS'. In other word, relevance not only comes from same-topic pairs, but also comes from cross-topic pairs. In this paper we call the relevance originated from cross-topic pairs as **Cor-relevance** in order to differentiate it with traditional concept of relevance.

Several topic pairs with high cor-relevance are illustrated in table 1. Those topic pairs were selected from millions candidates by mining which is also our training process with two conditions: first high enough cor-relevance score from page topics to ad topics in our training click/view data collection and second enough view number on the topic pairs. As we can see intuitively, although the two topics in any of the pairs are not identical, there exists certain causality which leads users to click the ads when they read the pages. The first example in the table shows that users who are interested in PC games have strong intention to click ads of electricity adaptor category. The fifth example in the table is the example of topic pair of 'new car' and 'GPS'. We strongly believe the topic cor-relevance information is very helpful to predict click-ability more precisely for ad retrieval task.

Table 1. Examples of high cor-relevant topic pairs

Page Topic Id	Hot Words in Page Topic	Ad Topic Id	Hot Words in Ad Topic
941	alliance, task, monster, mage	828	power, switch, voltage, electricity
690	fingernail, pipe, painting, beauty	706	indoor, decorating, house, decoration
837	fund, net-value, investment, report	830	bank, finance, credit, BOA
612	university, dean, advisor, student	535	computer, purchase, repair, second-hand
578	feature, BMW, auto, model	783	temperature, location, satellite, GPS

Although semantic match, such as PLSI model, is able to match 'motel' in page content to ads with bid term of 'cheap hotel', in general the match in the model only happens when the page and the ad have a same latent topic in both side. To our best knowledge, cor-relevance of different topics was still not taken into existing topic based relevance models reported in papers. In Benyah's paper [6], they considered some word-pairs to be strong signal of user clicks, However it is not trivial to mine those word-pairs significant to click/non-click classification from huge amount of pair candidates but very sparse click/view logs, which makes word-pair match still be hard to implement.

In order to leverage the topic cor-relevance information to build better relevance model we propose **Topic Cor-relevance Model (TCRM)** within latent topic space. Further, TCRM is a probabilistic instance of a more general model - **Cor-relevance Model (CRM)** that we propose formally in this paper. The CRM is able to model causation or relevance of a unit in page and another unit in ad no matter whether the two units are the same or not. As we can see, CRM is a natural extension of VSM. With CRM framework we can review existing VSM based methods from a broader perspective. The rest of this paper is organized as follows. In Section 2, we formally propose and formulate the general cor-relevance model and discuss its relation with VSM. Section 3 presents topic cor-relevance model in details, discusses its probabilistic meaning, and gives a MAP based learning method. Section 3 presents experimental results by comparing VSM, PLSI, diagonal-matrix TCRM and full-matrix TCRM on click/view data and relevance data. Finally we discuss the related work in Section 5 and give conclusions in Section 6.

2. GERNAL COR-RELEVANCE MODEL

In document retrieval, keyword based match mechanism has been pervasive for many years in several areas like page search, sponsored search, and contextual advertising etc. VSM model and other similar models assume relevance only comes from identical words from query side and ad side. However, we believe that more subtle relevance also exists in word pairs between page and ad although the words in pairs are not identical. But the cor-relevance of different units from two sides was ignored by most of

existing models. In order to exploit the cor-relevance to help the retrieval problem, we re-formulate match of page and ad as follows.

Documents including both pages and ads form a n-dimension vector space, a document is a vector composing of n document bases within the document space;

$$Document = x_1 \cdot doc\_base_1 + \ldots + x_n \cdot doc\_base_n \ldots (1)$$

Relevance of a page and an ad is calculated as below equation, if the page is $(q_1,\ldots,q_n)$ and the ad is $(a_1,\ldots,a_n)$. COR matrix of below is cor-relevance to measure how important (page, ad) unit pairs contribute to overall relevance.

$$\begin{aligned} Relevance &= \vec{q}^{\,T} \cdot COR \cdot \vec{a} \\ &= (q_1,\ldots,q_n) \begin{pmatrix} Cor_{1,1} & \cdots & Cor_{1,n} \\ \vdots & \ddots & \vdots \\ Cor_{n,1} & \cdots & Cor_{n,n} \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \ldots (2) \end{aligned}$$

In CRM framework, relevance not only originates from pairs of identical base of document space but also comes from pairs of cross base (Figure 1).

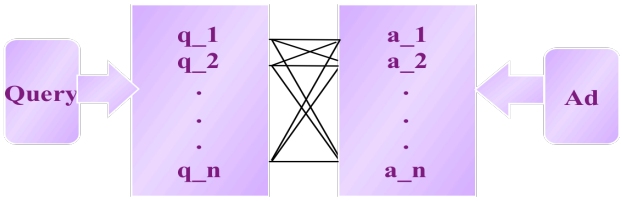

Figure 1. CRM is to model cor-relevance of a unit in page and another unit.

CRM is a natural extension of VSM that is able to model causation or relevance of a unit in page and another unit in ad no matter whether the two units are the same or not. If COR matrix is an identity matrix and word is document base and TF-IDF is to weight word importance, then the CRM is degenerated into a VSM (Figure 2). In latent topic based indexing approaches like PLSI, document space is latent semantic space and topic probability distribution of a document is a vector representation of the document. Moreover identity COR matrix is still adopted. Word-pair approach proposed in paper [6] also regards a document as a bag of words and uses words as document bases, but its feature value is an indicator if a word exists in the document. The approach assumes that the COR matrix is a very sparse one, uses mutual information method to select non-trivial matrix cells, then applies Maximum Entropy learning to train the matrix. Hence VSM, PLSI, and word-pair match can be viewed as instances of CRM model with different assumptions or simplifications.

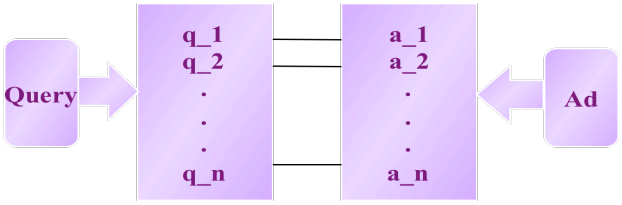

Figure 2. VSM is a special case of CRM when COR matrix is a diagonal matrix.

3. TOPIC COR-RELEVANCE MODEL

Topic cor-relevance model is also an instance of CRM with certain setting. We choose latent semantic space as the document space of CRM for several reasons: 1) Semantic match already addresses issues of synonymy and polysemous words to certain extend. We still want to keep the benefit in our new model. 2) Topic-level cor-relevance information is our goal that we try to leverage to build a better relevance or click predication model. 3) Topic model can naturally play a role of dimension reduction to cut down parameter space greatly. The scale of word pairs is at 100M if a vocabulary of 10 thousands of word is used, while several thousands of topics are enough for both page side and ad side in most cases. By mapping documents into latent semantic space we may reduce COR matrix from a scale of 100M to 1M. Less model complexity then better model generalizability, moreover the dimension reduction also makes the learning more feasible and robust.

Although clustering documents into topics relieve the click/view data sparseness problem greatly, it is still challenging to learning a robust cor-relevance model with 1M parameters using click/view data because data is still very sparse in long tailed topics. In order to overcome the data sparseness issue we give a Maximum a Posteriori (MAP) estimation based method to learn the cor-relevance between page topic and ad topic.

TCRM model consists of two parts: one is a latent topic model to map documents including both pages and ads into a latent semantic space; the other is a cor-relevance model to measure how cor-relevant a topic in page side is with another topic in ad side.

3.1 Formulation of topic cor-relevance model

Latent semantic is the document space for TCRM match of pages and ads. Firstly, documents are represented as vectors in latent semantic space, more specifically the vector is a probability distribution over topics $\{P(t_i \mid d\}_{i=1}^{k}$ of the document. Mapping documents into latent topic space, we may reduce COR matrix scale from 100M to 1M and relieve the data sparseness problem greatly. Moreover, intuitively it is reasonable to tie documents into categories based on their topics. For example, given both two ads are about computer products, if a user is reading a web page about a new laptop, it is reasonable to believe that two cor-relevance of the page and each of two ads are similar.

Within TCRM we incorporate a click probability of topic level $P(click \mid page\_topic, ad\_topic)$ to concretize the meaning of cor-relevance of the model. Thus we propose TCRM as follows:

1. Documents including both pages and ads are located in a latent topic space;

2. Relevance of a page and an ad is calculated as below equation, if the page is $(q_1,...,q_n)$ and the ad is $(a_1,...,a_n)$ and TCOR is topic COR matrix of topic level.

$$P(click \mid page, ad) = \vec{q}^T \cdot TCOR \cdot \vec{a}$$

$$= (q_1,...,q_k)\begin{pmatrix} m_{1,1} & \cdots & m_{1,k} \\ \vdots & \ddots & \vdots \\ m_{k,1} & \cdots & m_{k,k} \end{pmatrix}\begin{pmatrix} a_1 \\ \vdots \\ a_k \end{pmatrix} = \sum_{i,j} q_i m_{i,j} a_j ...(3)$$

$$where \ \ m_{i,j} = P(click \mid page\_topic\_i, ad\_topic\_j)$$

The cell of COR matrix $m_{i,j}$ is defined as click-ability of topic level which is cor-relevance in our TCRM.

For a given (page, ad) pair, calculation in formula (3) consists of two steps. First, we map the page and the ad into topic space; second, the click-ability of the (page, ad) pair is a probabilistic combination of $P(click \mid page\_topic\_i, ad\_topic\_j)$ over all possible page topics and ad topics. The calculation procedure is more straightforward to understand, if hard classification is adopted for topic categorization instead of soft classification. The first step is to map a document into category id, the second step is to search a lookup table with key (page_topic, ad_topic) for a corresponding click-ability on topic level. Essentially, we can attribute TCRM into smoothing methods using latent topic based tying.

3.2 Topic extraction

Two latent topic models are the Probabilistic Latent Semantic Analysis and the Latent Dirichlet Allocation. For PLSA, the training data collection *C* is a collection of occurrence of (word, doc) pairs, then the log likelihood of the data collection *C* to be generated with PLSA model is:

$$L(C) = \sum_d \sum_w c(w,d) \log \sum_{i=1}^{k} p(\theta_i \mid d) p(w \mid \theta_i) ...(4)$$

The parameters of PLSA are estimated with Expectation Maximization algorithm (EM) by maximizing the expected log likelihood *L(C)*. LDA is a Bayesian estimation version of PLSA by assuming that the all topic distributions of documents $p(\theta_i \mid d)$ are generated from a Dirichlet distribution, so that the number of parameters does not grow linearly with the training data size.

After getting a latent topic model, folding-in method [8, 10] is used to compute a topic distribution for a new document that was not contained in the original training dataset. In PLSA, the topic distribution can be computed by an EM similar iteration, where parameters of the PLSA model $\{p(\theta_i), p(w \mid \theta_i)\}$ are fixed such that only the mixing proportions of latent topics $p(\theta_i \mid d')$ are adapted iteratively as follows in the EM iteration.

$$p(\theta_i \mid d', w) = \frac{p(w \mid \theta_i) p(\theta_i \mid d')}{\sum_j p(w \mid \theta_j) p(\theta_j \mid d')} ...(5)$$

$$p(\theta_i \mid d') = \frac{\sum_w c(d', w) p(\theta_i \mid d', w)}{\sum_w c(d', w)} ...(6)$$

With proper limitation of iteration number to folding-in method, we are able to map a new page or ad into topic space even in a prompt mode.

3.3 Machine learning of cor-relevance

The next coming problem is how to learn cor-relevance matrix of TCRM. In equation (3), we defined cor-relevance of TCRM as click-ability over pair of page topic and ad topic and a predicted click-ability of a given pair of page and ad can be computed with equation (3). In the backend of an ad platform, click/view events of ads are logged, and the log data is a key evidence for ad system or search system to improve their relevance models or click models [13]. In click/view log of a real system, a lot of information is recorded. For a simplification, we formulate click/view log data of a contextual advertising system as a collection of tuples <page, ad, imp, clk>. Then the log likelihood of the log data is given as follows:

$$L = \sum_{q,a} \Big[clk \times \log\big(P(click \mid page, ad)\big) + (imp - clk) \times \log\big(1 - P(click \mid page, ad)\big) \Big]$$

With a fixed latent topic model, q_i and a_j are known parameters for all click/view event data, then we use Maximum Likelihood (ML) estimation to learn cor-relevance matrix $m_{i,j}$:

$$= \sum_{q,a} \left[clk \times \log\left(\sum_{i,j} q_i m_{i,j} a_j \right) + (imp - clk) \times \log\left(1 - \sum_{i,j} q_i m_{i,j} a_j \right) \right] ...(7)$$

In implementation, Gradient Descent method is applied to find minimum of –L with following gradient:

$$\frac{\partial(-L)}{\partial m_{i,j}} = -\sum_{q,a} \left[\left(\frac{clk}{\sum_{i,j} q_i m_{i,j} a_j} - \frac{imp - clk}{1 - \sum_{i,j} q_i m_{i,j} a_j} \right) \times q_i a_j \right] ....(8)$$

Because feasible ranges of all $m_{i,j}$ are (0, 1), in order to prevent optimized variables from being out of range (0, 1) in our Gradient Descent algorithm, we incorporate below parameter transform with Sigmod function and its inverted function, then do Gradient Descent optimization within the new parameter space.

$$m_{i,j} = \frac{1}{1 + e^{-x_{i,j}}} ...(9)$$

However, severe data sparseness still exists at topic level. From equation (8), we can see that we will have a delta modification for variable $m_{i,j}$, only if $q_i a_j > 0$ which means pages and ad in event data have projections in cell (i, j) of topic matrix. Therefore $\sum_{q,a} q_i a_j$ is a measure how many data will contribute to iterative update of $m_{i,j}$. In our statistics analysis, the distribution of available event data over topic pairs is still a long-tailed distribution (cf. Figure 3.) due to two reasons: 1) Topic pair distributions of pages and ads are naturally long-tailed; 2) Event log data are from an optimum existing ad system so that impressions of ads are not uniform on topics. According to our analysis, about only 16% topic pairs occupies 90% event data.

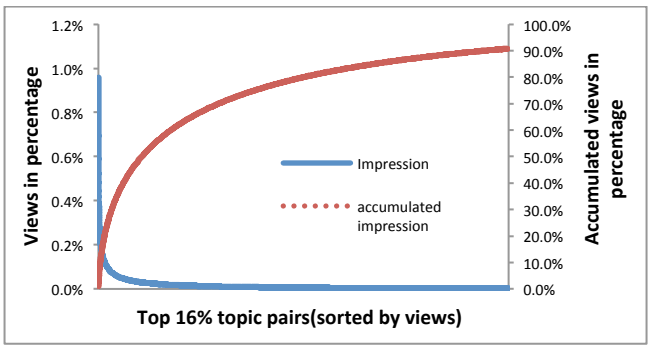

Figure 3. Distribution of view number on page-ad topic pairs

In order to dealing with the data sparseness, we employed a Bayesian prior of Beta distribution for each $m_{i,j}$ to regularize the estimation of $m_{i,j}$ in case that there is no enough training data for the parameter. For simplification and to be practical, we assume the priori distributions of different $m_{i,j}$ are independent to each other. Then we use Maximal a Posteriori (MAP) estimation to do the learning. With MAP method, the new log likelihood of the log data is:

$$L = \sum_{i,j} \log(Beta(m_{i,j};\alpha,\beta)) + \sum_{q,a}\left[clk \times \log\left(\sum_{i,j} q_i m_{i,j} a_j\right) + (imp - clk) \times \log\left(1 - \sum_{i,j} q_i m_{i,j} a_j\right)\right] \quad ...(10)$$

Gradient of –L becomes

$$\frac{\partial(-L)}{\partial m_{i,j}} = -\left\{\left(\frac{\alpha-1}{m_{i,j}} - \frac{\beta-1}{1-m_{i,j}}\right) + \sum_{q,a}\left[\left(\frac{clk}{\sum_{i,j} q_i m_{i,j} a_j} - \frac{imp - clk}{1 - \sum_{i,j} q_i m_{i,j} a_j}\right) \times q_i a_j\right]\right\} \quad ...(11)$$

It is easy to see that the only stable point of first part of equation (11) is $m_{i,j} = \frac{\alpha-1}{\alpha+\beta-2}$ which is exactly a maximum of the Beta priori. Thus, the estimation of $m_{i,j}$ will be convergent into the maximum of the Beta priori, if $\sum_{q,a} q_i a_j$ is very small. In our practice, α and β were chose by two criteria: 1) such that $\frac{\alpha-1}{\alpha+\beta-2}$ is equal to average CTR on all event data; 2) proper scales of α and β to balance the impacts of Beta priori and training data. Moreover, our experimental results also proved that MAP based method was much better to avoid the over-fitting problem brought from data sparseness than ML based method.

3.4 Discussions

According to the definitions in equation (3), we are able to interpret the cor-relevance of topic level to be P(click | page_topic, ad_topic), then P(click | page, ad) is a probabilistic combination of P(click | page_topic, ad_topic) over all possible page topics and ad topics. Essentially TCRM can be viewed as a smoothing method dealing with click/view data on latent topic level to come up with a more robust cor-relevance model than word level, where the assumption with which we are able to do the smoothing is that click patterns of (page, ad) from a same topic-pair are similar and robust statistically .

In word-pair based approach, it is almost impossible to take all possible pair candidates into calculation due to huge amount of candidates and severe data sparseness. Moreover it is not trivial to mine word-pairs meaningful to click/non-click classification from click/view event log due to data sparseness. Smoothing methods are well applied to tackle data sparseness problems in many areas. A typical way is to build a category above first layer elements, and collect statistical information on category level. If elements are from a same category, they will share statistics among them. An example of the method is that in speech recognition HMM states are tying and share data occurrence among tied states in HMM training [12].

Comparing with mining co-occurrences of word-pairs directly, it is a better way to cluster words into topic categories and do smoothing of click/view data with the topic categories. In statistical sense the behavior of click through from a page topic to another ad topic is more robust than click through from a word in page to another word in ad. With the principle of smoothing, for a new (page, ad) pair we predicate the click-ability with two steps: first is to map the page and the ad into topic categories, second is to combine all possibilities to obtain a comprehensive estimation on topic level. The two steps are embodied in equation (3) which is our proposed TCRM model.

4. EXPERIMENTS

In this paper we focus on TCRM's application only to contextual ads retrieval. In this section we discuss the dataset and metrics that we use to evaluate the model's performance, and then give the experiment design, finally we give the evaluation results and comparison of different models.

4.1 Topic model training

We will not discuss much about topic model training, because it is not key work in the paper to train a latent topic model. PLSA training was be used in our experiments. Regarding the training corpus we used a mixture of publishers' web pages and search result pages of ad bid terms, so that our model can cover the semantic topics of both query side and ad side. We did not use ads' descriptions directly considering the ad descriptions were too shorter than web pages. And we select a proper ratio of web pages and search result pages of bid terms in order to prevent that one side's topics overwhelm the other side's topics. The number k of latent topics was selected just empirically. In our experiment a topic model of about 1,000 topics was trained such that the parameter number of COR matrix was about 1M. And the same topic model was used for different settings of COR matrix.

4.2 Datasets and evaluation metrics

The click-ability is the learning objective in TCRM training and click/view data are consumed accordingly. From our point of view, click-ability is a kind of generalized relevance. For example, if a user is reading a web page about new car, he/she is maybe also interested in GPS. The generalized relevance is cor-

relevance for which we try to model with TCRM. However, 'new car' and 'GPS' are usually labeled as irrelevant in a editorial relevance dataset. Thus the generalized relevance is often missed in the editorial relevance data. In the other side, intuitively user's click behavior is good and relatively objective evidence of the generalized relevance. Accordingly we have two kinds of datasets. One kind is click/view dataset which is click/view event log of several months from a commercial contextual advertising system. In academy it is also a hot topic how to use click-through data in training and evaluation of document retrieval models [6,17,18,19]. A log entry of view event records a page id, an ad id, and with a flag whether any user clicked on the ad. Naturally entries with click flag are positive samples and entries with non-click flag are negative samples in training and evaluation. The second kind is relevance dataset with relevance labels 'good', 'fair' or 'bad' which are given by human manually. Using the two kinds of datasets for evaluation, we consider two related but subtle different evaluation criteria – click-ability and editorial relevance.

In our experiments, we randomly separated the click/view dataset into two parts, one was for training and the other was for testing. The relevance dataset was only used as testing dataset.

P-R curve and NDCG[14] are two key metrics that we used to compare different models. P-R curve is a measure how precise and efficient ad retrieval system is, NDCG measures a model's capability to correctly rank the ads for a page against the editorial labeled ads order. We did evaluations with P-R curves on both the click/view dataset and the relevance dataset to compare performances of the models against click-ability and editorial relevance. Testing with the relevance dataset, we just threw 'fair' cases away and took 'good'/'bad' cases as positive/negative samples correspondingly. Further we did NDCG evaluation with the relevance dataset to compare ranking capabilities of models. We did NDCG-3 and NDCG-5 evaluations because usually only a short list of ads was returned by ad platforms different with the case of page search result.

4.3 Models and evaluations

We trained TCRM models with the click/view training data, and evaluated the models with two testing datasets against click-ability and relevance objectives. Two baseline models were used in the comparison. The first was a word-based VSM model in which TF-IDF was weight of word feature. The second was a PLSI-Q model in which topic probability was weight of topic feature and cosine similarity was the match score as well. Because we believed that the generalized relevance - topic cor-relevance – was missed in the editorial relevance dataset, for a fair comparison with PLSI-Q we also trained a special TCRM model with a diagonal TCOR matrix and evaluated the special model. With the treatment of diagonal TCOR matrix, we assumed the learning goal of the diagonal-matrix TCRM model was only against relevance without cor-relevance. And at last we of course had a full-matrix TCRM model trained with MAP estimation.

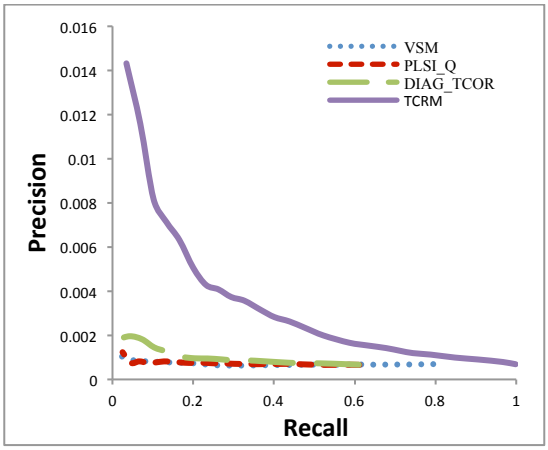

Figure 4. P-R curves of 4 models on click/view testing data.

P-R curves of several models over click/view testing dataset were showed in Figure 4. From click-ability evaluation, both full-matrix TCRM and diagonal-matrix TCRM are better on P-R curve than the two baseline models, moreover full-matrix TCRM is much better than diagonal-matrix TCRM, which indicates that cross-topic evidences – cor-relevance - contribute a lot to the accuracy of click prediction. In summary, user clicks happen not only from pages and ads of same topic, but also from pages and ads of cor-relevant topics. The kind of cor-relevance between topics should be leveraged to build more precise ad match models.

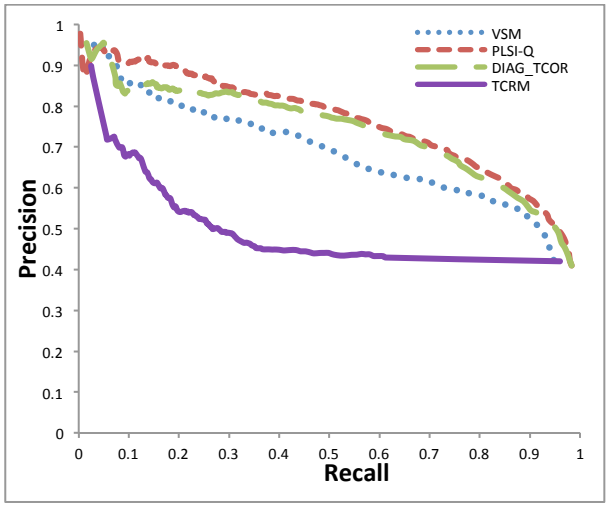

Figure 5. P-R curves of 4 models on relevance testing data.

P-R curves over relevance testing data were showed in Figure 5. NDCG evaluation for different models were showed in table (2). From both P-R curve and NDCG results, we can see that VSM and PLSI-Q were much better than full-matrix TCRM on the relevance testing data against editorial relevance goal. However diagonal-matrix TCRM's performance was closely comparable with PLSI-Q and even better than VSM although the training data of diagonal-matrix TCRM was not relevance dataset but still click/view dataset. In terms of NDCG, diagonal-matrix TCRM was a little better than PLSI-Q although the outperformance was insignificant.

Table 2. NDCG evaluations on relevance testing dataset

Model	NDCG-3	NDCG-5
VSM	0.839	0.881
PLSI_Q	0.864	0.900
DIAG_TCOR	0.870	0.904
TCRM	0.727	0.825

From the comparison our conclusions are: 1) Click-ability is not strictly correlated to traditional relevance given by editorial judgments. We think that the editorial relevance is relevance in a narrow sense, while click-ability is another kind of relevance in a generalized sense which is missed in editorial relevance data due to the difficulty of labeling click-ability. 2) We are able to set different TCOR matrices within TCRM to respectively learn either click-ability objective or narrow-sense relevance objective. Full-matrix TCRM is a model targeting click-ability and diagonal-matrix TCRM is a model targeting narrow-sense relevance. From this point of view, we can get an insight on how editorial relevance is connected with click-ability and why good CTR is able to be achieved by returning relevant ads. Moreover from the point of view we can see that it is only a sub-optimal treatment to return narrow-sense relevant ads in many traditional methods. Therefore it is a better method to take click-ability as the objective of modeling and learning, because the final goal of ad systems is to gain more clicks in a Cost per Click (CPC) pay model.

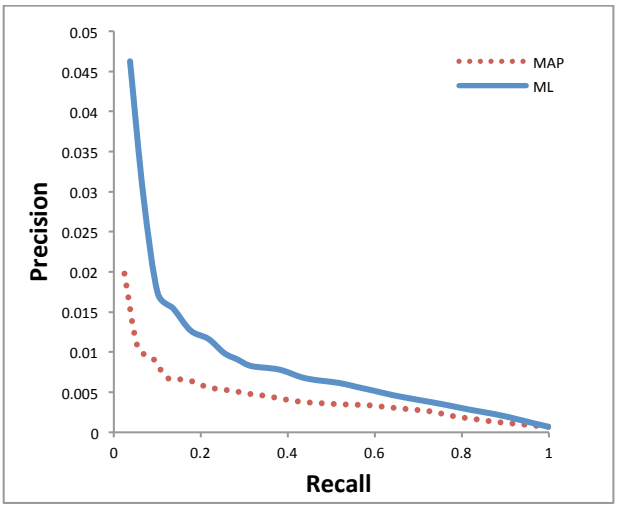

Figure 6. P-R curves of ML and MAP methods on training data

Another interesting result was the comparison of Maximum Likelihood estimation and Maximum a Posterior estimation. Figure 6 was P-R curves on training dataset in which we can see ML method was much fitter than MAP method in training dataset. In area of high precision and low recall, the precision even can be as high as 4.5% which was much higher than MAP method's result. However in testing dataset P-R curves showed in Figure 7, ML method was much worse than MAP method. Our explanation was that severe data sparseness also existed on topic-pair level, and the sparseness caused over-fitting on training data by ML method and hurt the generalizability of TCRM model. While MAP method partially relieved the over-fitting issue by applying a proper regulation of Bayesian priori on estimated parameters and improved the model's generalizability significantly.

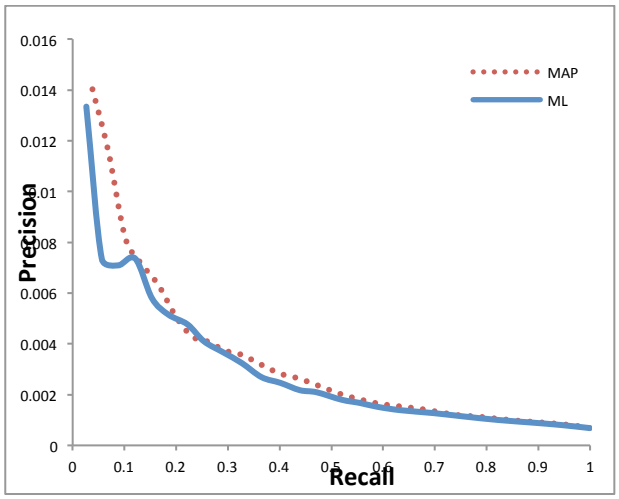

Figure 7. P-R curves of ML and MPA methods on testing data

5. RELATED WORK

Vector-Space Model [2, 16] is one of most widely used models for information retrieval task. In VSM, documents including both page and ad are represented as a vector of term (word or phrase) weights, where the weights are importance of the terms within the document context. Then a cosine of two vectors from page and ad is to measure similarity of the page and the ad. However VSM based exact word match runs into problems in case of synonymy or polysemous words. Many researchers have tackled the problem with different approaches from different perspectives. From our point of view those approaches can be classified into syntactic category and semantic category.

Query rewriting technology is a typical example of syntactic methods. For a given query not only the original query but also several rewrites are used to retrieval related ads. Different approaches have been proposed to generate rewrites for the given query with various evidences. Jones et al. [3] derived the rewrites from user sessions in query logs. In scenario of contextual advertising, B. Ribeiro-Neto etc[20] proposed a query-side expansion method which firstly find the most relevant pages to the triggering page and then extract most representative terms in those pages to expand the text of the triggering page to reduce the "vocabulary impedance" with regard to an advertisement. Benyah et al. [6] considered some word-pairs to be strong signal of clicks, used mutual information based method to mine those word-pairs and used the selected word-pairs as indicator features, and Maximum Entropy learning method was employed to learn the importance of the selected word pairs. However it is not trivial to mine word-pairs meaningful to click/non-click classification from click/view event log because of two reasons. 1) Huge problem - the amount of word-pairs is at scale of 100M with a vocabulary of ten thousands of words; 2) sparse data - there are only a few of click/view data for most of word combinations, and it is biased to derive a word-pair importance to clicks with only a few of data. The challenge makes word-pair match still be an open question.

Latent topic models have been applied to many text mining tasks. The basic idea of these models is to model documents with a finite mixture model of k topics, where each topic is a multinomial

distribution of words which can be regarded as a group of similar words inferred from word concurrences in documents. PLSI-Q proposed in [10] is an example of semantic methods. In the approach, documents are mapped into a low dimension space – latent semantic space. Each dimension of the semantic space is regarded as a topic which is a cluster of related words in some sense and a document is a vector in the space. Then similar with VSM method, cosine similarity is used to measure the similarity of two documents in the latent semantic space. Latent topic based approaches partially addressed issues of synonymy or polysemous words by mapping documents into topic space and do document matching in the new space.

Andrei Broder et al. [7] also proposed another semantic approach to contextual advertising in which a human edited taxonomy tree of around 6,000 nodes was employed. Each node in the taxonomy is represented as a collection of exemplary bid phrases or queries that correspond to that node concept. The semantic match of the pages and the ads is performed by classifying both pages and ads into the common taxonomy and in addition to relevance score from syntactic features it rated some relevance score according to the least common ancestor of the page node and ad node. What it was different with PLSI was that topic structure was a taxonomy tree and when a document was classified into a node in the tree, not only the node but also the node's ancestors were going to be counted for matching.

Our work is more similar to word-pair match proposed by Benyah etc. However what differentiate our work is: 1) Topic model is employed to reduce numbers of pairs from 100M scale to 1M so that data sparseness issue is solved greatly after the feature reduction. Although Mutual Information method was adopted in word-pair match, the results are very sensitive to the training data. For example, even if 'laptop' and 'ipad' are identified as a significant pair in a dataset, it is very possible that 'notebook' and 'ipad' are a insignificant pair in the dataset due to data bias. As a contrast, with topic based approach 'laptop' and 'notebook' are from same topic category, more data are available on category level, then better generalizability is with TCRM due to much less model complexity. 2) A more feasible MAP based learning is proposed. The learning is more robust than ML estimation or Maximum Entropy method by involving a proper setting of Bayesian prior in the case of data sparseness. Our experiments showed the learning method worked very well with a problem of 1M parameters. 3) With TCRM we explain what the difference is between a click-ability model and relevance model. From the perspective we have an insight on the connection of click-ability and relevance. 4) CRM model is proposed formally. The model is a more general and abstract framework able to cover existing VSM, PLSI models and our TCRM model, able to model causation or relevance of a unit in page and another unit in ad.

6. CONCLUSIONS

Although issues of synonymy or polysemous words have been tackled with different approaches for many years, there is information significant for ad matching which is still unveiled by existing papers – cor-relevance between different topics – to our best understanding. In another word, although two topics are totally different in semantic, there exists some caution relationship between a user's reading a page of the first topic and clicking an ad of the other topic. We think that the kind of caution is another kind of relevance in a generalized sense. In our paper we are trying to model the generalized relevance formally and leverage the information for ad retrieval problem.

Formally we proposed an abstract cor-relevance model by incorporating a cor-relevance matrix between page units and ad units to reveal more subtle relevance relation. The CRM model is able to model causation or relevance of a unit in page and another unit in ad no matter whether the two units are the same or not. The model is a natural extension of VSM. CRM is degenerated into a VSM when COR matrix is an identity matrix, which make VSM is a special case of VSM. Thus many existing VSM based approaches can still be represented with CRM, so that we can review those approaches from another broader perspective. The CRM model generalizes the definition of relevance from what is radical from identical words or topics from subtle cor-relevance existing in case that words or topics between page and ad are not identical. The generalization brings us more flexibility to model relevance in matching problems.

Employing latent topic space as document space of CRM model, we further proposed topic cor-relevance model within latent topic space. We may interpret the cor-relevance matrix in the model as P(click | page_topic, ad_topic) and interpret the output of TCRM as a prediction of click-ability for given page and ad under probability framework. The interpretation gives us an insight on how relevance is connected to click-ability. With the explanation click-ability is also a kind of relevance in a general sense. Even pages and ads can be represented in two different latent semantic spaces within TCRM model, so that we may train dedicated latent topic models for page side and ad side. It brings us more flexibility of modeling not to require pages and ads located in a same semantic space. But we did not cover more about this point in the paper.

Although TCRM brings more complexity to parameters learning because huge parameter space is also brought in with better representation, a feasible learning method based on Maximum a Posteriori estimation was proposed by this paper to automatically learn the parameters whose number is at scale of 1M. Compared with ML method, the MAP based learning greatly addresses the data sparseness problem existing in learning in a systematical way. With experiments it was proven that our method was robust to learn TCOR matrix from sparse click/view data and our TCRM model performed better to predict click-ability on an unknown testing dataset than typical VSM or PLSI model by taking cor-relevance into consideration. Even against narrow-sense relevance, diagonal TCRM achieved comparable NDCG metric on an editorial relevance dataset. Therefore TCRM is a more general model which may represent both narrow-sense relevance and generalized relevance such as click-ability with proper configurations.

TCRM does not require pages and ads to be located in a same topic space. This allows us to train dedicated a page topic model and an ad topic model respectively. Considering page topics may be more diverse and ad topics may be more commercial related, the dedicated topic models may further improve the precision of click prediction. Another interesting direction is to apply TCRM to Sponsored Search problem. Although high relevance is more important for high click-ability in Sponsored Search, intuitively it is reasonable to expect topic cor-relevance to be meaningful in the Sponsored Search scenario as well. Of course ad position bias should be considered and addressed well when applying TCRM. Those are our future work regarding TCRM.

7. REFERENCES

[1] Chingning Wang, Ping Zhang, Risook Choi, and Michael D. Eredita. Understanding Consumers Attitude toward Advertising. In Eighth Americas Conference on Information System, pages 1143–1148, 2002.

[2] Gerard Salton, A. Wong, and C.S. Yang. A Vector Space Model for Automatic Indexing. Communications of the ACM, 18:613-620, 1975.

[3] Rosie Jones, Benjamin Rey and Omid Madani, and Wiley Greiner. Generating Query Substitutions. WWW'06, 2006.

[4] Filip Radlinski, Andrei Broder, Peter Ciccolo, Evgeniy Gabrilovich, Vanja Josifovskiy, Lance Riedel, Optimizing Relevance and Revenue in Ad Search: A Query Substitution Approach. SIGIR'08, 2008.

[5] Andrei Broder, Peter Ciccolo, Evgeniy Gabrilovich, Vanja Josifovski, Donald Metzler, Lance Riedel, Jeffrey Yuan, Online Expansion of Rare Queries for Sponsored Search. WWW 2009, 2009.

[6] Benyah Shaparenko, Özgür Çetin, Rukmini Iyer, Data-Driven Text Features for Sponsored Search Click Prediction. ADKDD'09, pages 46-54, 2009

[7] Andrei Broder, Marcus Fontoura, Vanja Josifovski, Lance Riedel, A Semantic Approach to Contextual Advertising, SIGIR '07, 2007

[8] Thomas Hofmann, Probabilistic Latent Semantic Analysis, Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99), 1999

[9] David Blei, Andrew Ng, and Michael Jordan. Latent Dirichlet Allocation. J. Mach. Learn. Res., 3:993–1022, 2003.

[10] Thomas Hofmann, Probabilistic Latent Semantic Indexing. In Proceedings of SIGIR'99, pages 50–57, 1999.

[11] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 1990.

[12] Mei-Yuh Hwang, Xuedong Huang, and Fil Alleva, Predicting Unseen Triphones with Senones, IEEE Transactions on Speech and Audio Processing, 1996.

[13] Massimiliano Ciaramita, Vanessa Murdock, Vassilis Plachouras, Online Learning from Click Data for Sponsored Search, WWW'08, 2008

[14] Kalervo Jarvelin, Jaana Kekalainen, Cumulated gain-based evaluation of IR techniques, ACM Transactions on Information Systems 20(4), 422–446, 2002.

[15]

[16] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. Information Processing and Management, 24:513-523, 1988.

[17] C. Clarke, E. Agichtein, S. Dumais, and R. White. The inuence of caption features on clickthrough patterns in web search. In Proc. ACM SIGIR, pages 135-142, 2007.

[18] T. Joachims. Optimizing search engines using clickthrough data. In Proc. ACM SIGIR, pages 133-142, 2002.

[19] T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In Proc. ACM KDD, pages 154-161, 2005.

[20] B. Ribeiro-Neto, M. Cristo, P. Golgher, and E. Moura. Impedance coupling in content-targeted advertising. In Proc. ACM SIGIR, pages 496-503, 2005.