Advertising Keyword Generation Using Active Learning*

Hao Wu[1], Guang Qiu[1], Xiaofei He[2], Yuan Shi[1], Mingcheng Qu[1], Jing Shen[3], Jiajun Bu[1] and Chun Chen[1]
[1,2]College of Computer Science and Technology, Zhejiang University, China
[3]China Disabled Persons' Federation Information Center
[1]{haowu, qiuguang, shiyuan, qumingcheng, bjj, chenc}@zju.edu.cn,
[2]xiaofeihe@cad.zju.edu.cn, [3]shenjing@cdpf.org.cn

ABSTRACT

This paper proposes an efficient relevance feedback based interactive model for *keyword generation* in sponsored search advertising. We formulate the ranking of relevant terms as a supervised learning problem and suggest new terms for the seed by leveraging user relevance feedback information. Active learning is employed to select the most informative samples from a set of candidate terms for user labeling. Experiments show our approach improves the relevance of generated terms significantly with little user effort required.

Categories and Subject Descriptors: H.3.5 [Information Systems]: Information Storage and Retrieval—*On-line Information Services*

General Terms: Algorithms, Design, Experimentation

Keywords: Active Learning, Keyword Generation, Sponsored Search

1. INTRODUCTION

Sponsored search is a successful form of on-line advertising, with annual revenue exceeding billions of dollars. With the goal of branding or/and marketing, advertisers create ads and bid on relevant keywords. The auction winners have their ads displayed as sponsored links alongside the organic search results when bidded terms are queried. *Keyword generation/suggestion* methods are employed to assist advertisers in finding all the terms relevant to their products or services. The term relevance is crucial to capture valid queries and clicks from their potential customers.

Provided an initial keyword which represents a concept such as "*shoes*" (i.e. a *seed term*), some typical keyword generation methods make use of a set of elements (such as URLs, frequent queries and meta-tags) that enrich the meaning of the seed to select high co-occurrence or similar terms as *suggestions*. To be effective, keyword generation requires even hundreds of relevant terms to be suggested for a seed. Some popular keyword generation tools (e.g. Google's Adwords Tool) mainly rely on query log mining and have the drawbacks of failing to suggest terms that don't contain the seed and ignoring the semantic similarity between terms.

Recent related work tends to exploit semantic relationships between terms. For keyword generation, Joshi and Motwani [3] present TermsNet, which captures semantic similarity between terms as a directed graph; Abhishek and Hosanagar [1] use a web based kernel function to establish semantic similarity between terms; Chen and et al. [2] exploit the semantic knowledge among concept hierarchy. However, there is no effort taking user relevance feedback information, especially that from advertisers, into account.

In this paper, we propose an interactive model to explore relevance feedback for keyword generation. Our approach also differs from previous work in formulating the ranking of relevant terms as a supervised learning problem, in which the term relevance is learned from several representative features rather than determined using a specific measure.

2. KEYWORD GENERATION

2.1 The Interactive Model

We establish semantic relevance between terms by leveraging relevance feedback, which conveys direct semantic information. For a seed term, our system first employs search engines to match it with a large set of ranked *candidate terms* that are determined using TFIDF measure. The candidates achieve a high coverage of relevant terms, but a poor precision. Users are then required to provide *relevant/irrelevant* labeling on just a few candidates selected by the system. Given these labeled candidates, a regression model is trained to predict the *relevance scores* of unlabeled candidates to the seed, from some representative features about them. Based on the relevance scores, we re-rank all the candidate terms to improve the precision. Finally, top ranked terms are selected as suggestions for the seed and returned to users.

Due to the consideration of the performance and user satisfaction, there are two requirements: (1) The samples to be selected for labeling should be as few as possible since labeling is labor-intensive. (2) The quality of training samples affects the performance directly. A natural strategy is to select the most informative samples using active learning.

2.2 Candidate Term Generation

To be efficient, we represent each seed term using a *characteristic document* which contains retrieved snippets from top search-hits. It is similar to [3]. Given a set of seed terms, the corresponding characteristic documents form a corpus. We remove stop words from these documents and stem the terms using Porter's stemmer as preprocessing. The TFIDF of all terms in the corpus is used as the initial weight of relevance. Top weighted terms in a characteristic document are selected as the candidates for the corresponding seed term.

*Supported by the National Key Technology R&D Program of China (NO.2008BAH26B02).

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

2.3 Feature Representation

For each (*seed, candidate*) term pair, the following features characterizing the relevance between them are used as *predictor variables*: (1) **TF** and **TFIDF** of the candidate in the search snippets document of the seed; (2) **Inverse TF**: the frequency of the seed in the search snippets document of the candidate; (3) **Search Snippets Similarity**: the frequency with which the snippets of top search results for the seed and the candidate contain the same words. (4) **Common Search URLs**: the frequency with which the seed and the candidate share the same search URLs.

2.4 Active Learning

We apply an active learning approach called Transductive Experimental Design (TED [4]), which tends to select candidates (for labeling and training) that are *hard-to-predict* and *representative* for unlabeled candidates. Let $\mathbf{V}$ denote both the matrix $[\mathbf{v}_1, ..., \mathbf{v}_n]^T \in \mathbb{R}^{n\times d}$ and the data set $\{\mathbf{v}_i\}$, and $\mathbf{X}$ denote both the matrix $[\mathbf{x}_1, ..., \mathbf{x}_m]^T \in \mathbb{R}^{m\times d}$ and the data subset $\{\mathbf{x}_i\}$, where $\mathbf{X} \subset \mathbf{V}$ (m < n). In our case, $\mathbf{v}_i$ is the feature vector of a (candidate, seed) term pair. Define $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ as the output function learned from *measurements* $y_i = \mathbf{w}^T\mathbf{x}_i + \epsilon_i$, $i = 1, ..., m$, where $\mathbf{w} \in \mathbb{R}^d$ is *weight vector* and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is measurement error. Let y_i (label) be the *relevance score* (1 or 0) associated with the feature vector $\mathbf{x}_i$. Consider a regularized linear regression problem, the maximum likelihood estimate of $\mathbf{w}$ is given by

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\{ J(\mathbf{w}) = \sum_{i=1}^{m} (\mathbf{w}^T\mathbf{x}_i - y_i)^2 + \mu||\mathbf{w}||^2 \right\} \quad (1)$$

where $\mu > 0$ and $||\cdot||$ is the vector 2-norm. It is known that the estimation $\mathbf{e} = \mathbf{w} - \hat{\mathbf{w}}$ has a covariance matrix given by $\sigma^2\mathbf{C}_\mathbf{w}$, where $\mathbf{C}_\mathbf{w}$ is the inverted Hessian of $J(\mathbf{w})$

$$\mathbf{C}_\mathbf{w} = \left(\frac{\partial^2 J(\mathbf{w})}{\partial \mathbf{w}^2} \right)^{-1} = (\mathbf{X}^T\mathbf{X} + \mu\mathbf{I})^{-1} \quad (2)$$

The introduced regularization improves numerical stability since $\mathbf{X}^T\mathbf{X} + \mu\mathbf{I}$ is full-rank. Let $\mathbf{f} = [f(\mathbf{v}_1), ..., f(\mathbf{v}_n)]^T$ be the function values on all the available data $\mathbf{V}$, then the predictive error $\mathbf{f} - \hat{\mathbf{f}}$ has the covariance matrix $\sigma^2\mathbf{C}_\mathbf{f}$ with

$$\mathbf{C}_\mathbf{f} = \mathbf{V}\mathbf{C}_\mathbf{w}\mathbf{V}^T = \mathbf{V}(\mathbf{X}^T\mathbf{X} + \mu\mathbf{I})^{-1}\mathbf{V}^T \quad (3)$$

In contrast to $\mathbf{C}_\mathbf{w}$, $\mathbf{C}_\mathbf{f}$ directly characterizes the quality of predictions on the target data $\mathbf{V}$. Therefore, the total predictive variance on the complete data set $\mathbf{V}$ is given by

$$\sum_{i=1}^{n} E\left[(f(\mathbf{v}_i) - \hat{f}(\mathbf{v}_i))^2 \right] = \sigma^2 Tr(\mathbf{C}_\mathbf{f}) \quad (4)$$

One should find a subset $\mathbf{X}$ which can minimize the total predictive variance. In our case, we apply sequential optimization of Kernel Transductive Experimental Design [4].

3. EXPERIMENTAL RESULTS

We select 100 category names widely spread over different topics from eBay and Amazon to form the set of seed terms. They're popular among advertisers and customers. For each characteristic document generation, top 400 Google search-hits are acquired. Top 400 weighted terms in each seed's characteristic document are selected as candidate terms.

Experimental results are averaged over 20 benchmark seed terms. Ground truth labeling of relevant/irrelevant was provided by 5 human evaluators who are familiar with related materials, and it was performed beforehand after we generated candidate terms, whose initial ranking is regarded as **Baseline0**. For the sake of user satisfaction, we select just 10 samples from the set of each seed's 400 candidates for user labeling. Given these 10 labeled samples, we equivalently apply kernel regression with Gaussian kernels to predict the relevance scores of unlabeled candidates. The following methods are compared: (1) **Pseudo-Relevance Feedback**: we select 5 samples as positive ones from the top of the initial ranking and 5 samples as negative ones from the end. It is regarded as **Baseline1** since no user labeling is required; (2) **Random Sampling**: we randomly select samples and repeat 100 rounds to get an average result. (3) **TED** is employed to select samples. μ is fixed as 0.0001.

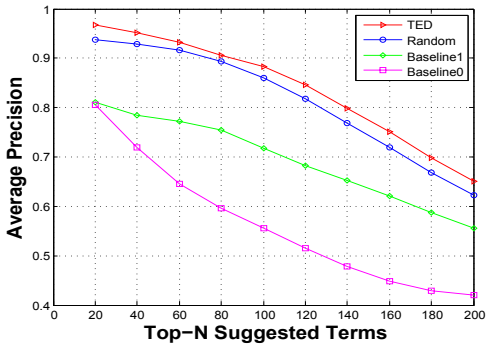

Figure 1: Average Precision Curves

Figure 1 shows the average precision curves for the relevance of top 200 suggested terms after re-ranking, where precision is defined as the ratio of relevant terms generated to the total number of terms generated. As illustrated, both random sampling and TED significantly outperform pseudo-relevance feedback since additional user relevance feedback is used. In this case, TED consistently performs the best, with 88.3% average precision achieved for top 100 suggestions, and 65.1% for top 200 suggestions. TED achieves 2.2% performance improvement for top 100 suggestions, and 2.7% for top 200 suggestions, comparing to random sampling. This is because active learning selects the training samples which can minimize the total predictive variance.

4. CONCLUSION AND FUTURE WORK

We have shown the effectiveness of our relevance feedback based interactive model in generating relevant terms. Experimental results highlight the advantage of the active learning algorithm in selecting the most informative samples for user labeling. Furthermore, our approach achieves a good tradeoff between the performance and user satisfaction. In practice, our keyword generation framework can be extended to other applications such as term clustering. Currently, only single word terms are considered. We are going to explore the use of phrases in the future work.

5. REFERENCES

[1] V. Abhishek and K. Hosanagar. Keyword generation for search engine advertising using semantic similarity between terms. In *ICEC'07*, 2007.

[2] Y. Chen, G.-R. Xue, and Y. Yu. Advertising keyword suggestion based on concept hierarchy. In *WSDM'08*, 2008.

[3] A. Joshi and R. Motwani. Keyword generation for search engine advertising. In *ICDM'06*, 2006.

[4] K. Yu, J. Bi, and V. Tresp. Active learning via transductive experimental design. In *ICML'06*, 2006.