Keyword Extraction for Contextual Advertisement

Xiaoyuan Wu
eBay Research Labs
No.88 KeYuan Rd.
Shanghai, China
xiaowu@ebay.com

Alvaro Bolivar
eBay Research Labs
2145 Hamilton Avenue
San Jose, CA 95125
abolivar@ebay.com

ABSTRACT
As the largest online marketplace, eBay strives to promote its inventory throughout the Web via different types of online advertisement. Contextually relevant links to eBay assets on third party sites is one example of such advertisement avenues. Keyword extraction is the task at the core of any contextual advertisement system. In this paper, we explore a machine learning approach to this problem. The proposed solution uses linear and logistic regression models learnt from human labeled data, combined with document, text and eBay specific features. In addition, we propose a solution to identify the prevalent category of eBay items in order to solve the problem of keyword ambiguity.

Categories and Subject Descriptors
H.3.1 [Content Analysis and Indexing]: Abstracting methods

General Terms: Algorithms, experimentation

Keywords: Keyword extraction, contextual advertisement

1. INTRODUCTION
Ebay, the World's Online Marketplace, enables trade on a local, national and international basis. Given the size of the market and the need to maintain a balance between supply and demand, driving potential buyers to the site is one of eBay's most important priorities. Contextual advertising of relevant eBay items (ads) is one important strategy to achieve such purpose. In the future, eBay items could be made available everywhere on the Web with an instant purchase option and without the buyer having to visit the eBay site. For example, a user who is browsing an "iPod" related web-page could find "iPod" related items and purchase them directly from that page.

In this paper, we will focus on the core technology of any contextual advertisement system, specifically, the keyword extraction algorithm. The more relevant the keywords extracted are, the more accurate the ads provided will be, and in turn, increased click-through-rate as well as revenue potential. In the first part of our work, we explore a significant set of features which are potentially helpful to determine the importance of keywords. In addition, regression models learnt from previous training data are applied to combine the features into a single keyword score. The basic idea is similar to the method presented in [3]; however, in our work, we investigate additional HTML features. We also take advantage of eBay proprietary data, such as query logs, keyword's item frequency entropy across categories, numbers of categories matched by keywords, among others and use them as additional features. In the second part of our work, we aim to resolve the problem of keyword ambiguity i.e. a keyword may have multiple intents. As a result, even if we extract the right keywords, the ads may not be relevant. A novel method proposed by [1] intends to classify Web pages and keywords to the same taxonomy, and use the proximity of the ad and page classes in the ranking formula. Unfortunately, we do not have the resource to maintain a large taxonomy of Web pages; however, we have a hierarchical category tree of items. Therefore, instead of classifying Web pages and keywords, we provide an approach to take advantage of the contextual keywords extracted from the same page to select a proper category to display ads.

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

2. KEYWORD RANKING
After HTML clean-up and tokenization processes, the next step is to rank the resulting keywords and phrases by their relevance score. Experiments with a large set of features were executed. These features can be divided into two groups, features related to the content of the source web-page and features related to eBay's view of such keywords. The final goal is to model relevance scores obtained from human labeled data through linear and logistic regression models to combine these features and obtain a keyword ranking score.

2.1 Features from Web Page
Table 1 lists a set of features related to Web page, which are potentially useful to rank keywords.

Table 1. Features from Web page

Term frequency (TF)	Phrase length
Title	Meta keywords/Meta description
Capitalization	Term's Position
H1/H2	Positive/Negative font attributes
Internal/External anchor text	

2.2 Features from eBay
Query Log: A large item-based search engine is available on eBay sites. Intuitively, the more times a query is used, the higher probability the query is a good keyword or phrase.

Entropy (Leaf Category): eBay maintains a large single-parent category tree, which buyers and sellers use to browse and list their items. If the items matched by a term are distributed over many leaf categories, we deem the term as not informative enough. The higher a term's leaf-level category entropy is, the higher the likelihood of this term to be irrelevant.

Entropy (Root Category): We found that only using leaf category entropy to determine a term's importance may be deceptive. Important terms (e.g., iPod) may be very popular in eBay, and they may appear in many leaf categories. As a result, their entropy is very high; however, those leaf categories may

belong to a single root category. Hence, we calculate root-category entropy using root-category level item counts.

Number of Categories: We calculate the number of categories in which a term appears. Similar to the entropy calculation, we obtain the number matching leaf categories and root categories.

Number of Items: The number of items matched by a term.

3. CATEGORY SELECTION

The output of the keyword ranking process is a ranked list of keywords for a given web-page. However, many keywords are ambiguous. For example, the keyword "css" may be extracted from a page describing web-page development; however, the ads displayed on the page might refer to a "Sony CSS-PHA Cybershot Station". The problem is that there are several matching categories for the term "css" on eBay site, but only a subset refers to the proper context. Therefore, our purpose is to select a proper category for each keyword according to the context of the web-page; in particular, we hope to determine a category for each keyword with the help of other keywords from the same page. For example, if "css" is together with other keywords such as "javascript" and "html" etc, it is more likely that we need to get ads from "computer" category, rather than from the "camera" category.

Inventory information and daily user activities are recorded by the eBay logging system in order to capture supply data (item counts) and demand data (user activities) for each keyword. By the supply data, we could get a vector of categories in which the keyword *i* appears, and by the demand data, we also could get a vector of categories in which users click through to view/bid/buy items by querying the keyword *i*. Hence, by the combination of supply data and demand data, for each keyword, we could obtain a vector of candidate categories. The target is leaf categories, where each category has a value which is calculated by the combination of item counts and buyer activity (view/bid/buy) in that category.

Given any Web page, the scores for all matching leaf categories are rolled up to root categories, and the top *N* root categories by voting are selected. For example, the leaf category vectors for "css", "html" and "javascript" roll up to the root category "computer" instead of "camera". The idea here is somewhat similar to collaborative filtering [2]. Finally, the most representative leaf category within the top *N* root categories is selected for each keyword.

4. EXPERIMENTAL EVALUATION

4.1 Experiment Setup

The training/testing set is made up of 800 Web pages. This set of pages was randomly selected from a large pool of eBay partner Web sites. The dataset was further divided in six types of Web pages. Table 2 shows the detailed break down.

Table 2. The distribution of Web pages.

News	Portal & Homepage	Blog	Forum	Social Network	Product Review	Total
138	96	279	115	62	96	800

Each page-keyword pair was judged by five annotators on a 1-4 scale. Moreover, precision of top *N* keywords (P@*N*) is used to evaluate the performance.

4.2 Features Selection

After a bunch of experiments, such as t-test for linear regression, z-test for logistic regression, leave-one-out method and multicollinearity analysis, we select TF, length, title, number of root categories, Meta keywords, Meta description, query log, root entropy, position and H1 as final features to rank keywords.

4.3 System Performance Comparison

The final system is benchmarked against a set of external keyword extraction systems, namely: Yahoo, Inxight, MediaRiver and KEA. Figure 1, shows that our system (eBay KES) outperforms all other systems.

4.4 Performance on Different Types of Pages

We analyzed the performance of the system on different page types, the results are presented in Figure 2. We could conclude that the more content-targeted types of pages, the easier to extract keywords. This results call for customizing the keyword extraction system based on the web-page type.

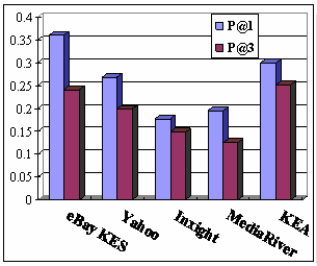

Figure 1. Comparison of keyword extraction systems.

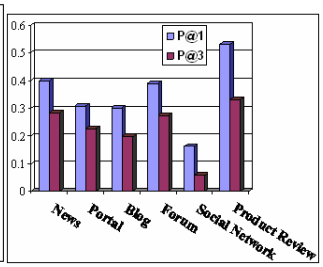

Figure 2. Performance comparison on different types of pages.

5. CONCLUSIONS

The eBay contextual advertisement platform has been created to automatically associate contextually relevant eBay assets to web-pages. This study explored a machine learning approach and described in detail the system at features for ranking keywords, as well as category selection to avoid keyword ambiguity. Our experimental result verifies the effectiveness of the keyword extraction system. Because we only annotated page-keyword pairs, we could not report the performance of category selection algorithm in this paper and we will consider it as future work.

6. REFERENCES

[1] A. Broder, M. Fontoura, V. Josifovski and L. Riedel. A Semantic Approach to Contextual Advertising. In SIGIR 2007, pages 559–566, Amsterdam, July 2007.

[2] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom and J. Riedl. GroupLens: An open architecture for collaborative filtering of netnews, Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work, 175-186.

[3] W. Yih, J. Goodman, and V. R. Carvalho. Finding advertising keywords on Web pages. In WWW'06, pages 213–222, New York, NY, 2006. ACM.