Mining Search-Phrase Definitions from Item Descriptions

Hung V. Nguyen and Hasan Davulcu

Department of Computer Science and Engineering, Arizona State University
Tempe, AZ 85287, USA
hung,hdavulcu@asu.edu

***Abstract*— In this paper, we develop a model for representing term dependence based on Markov Random Fields and present an approach based on Markov Chain Monte Carlo technique for generating phrase definitions. This approach can use a small corpus of keyword matching and a random sample of other product descriptions for an advertiser's search-phrase to effectively mine and rank alternative but highly relevant *search-phrase definitions*. These definitions, which are search-phrases themselves, can then be provided as alternative phrases to an advertiser.**

I. Introduction

The World Wide Web has made a dramatic transition from its early beginnings as a distributed repository of browsable information into a dominant medium for conducting e-commerce. In particular, it has become a mainstream advertising medium for retail goods with online advertising reaching $16 billion in revenue in 2006. Because of this immense commercial power of the Web, the number of vendors, both large and small, who are setting up online presence and subscribing to search advertising continues to proliferate.

In search advertising, vendors subscribe to triplets of the form $< searchphrase, product - url, bid >$. For example, a shoe store may subscribe to advertise its "NIKE Airmax 180" product as "running shoes" by specifying the triplet $< runningshoes, NIKEAirmax180, \$.50 >$. This triplet indicates that whenever a web search phrase mentions "running shoes" or a web site contains "running shoes" related information, this vendor would like to list its "NIKE Airmax 180" product and agrees to pay 50 cents per click.

Nevertheless, vendors can not anticipate all possible ways in which to advertise for their products and shoppers can not guess all possible ways to search and find a product. Many times, user's search query may not be a perfect description of their information needs. Even when the information is somewhat well described, a search engine or information retrieval system may not be able to retrieve documents matching the query as stated. We call this phenomenon is "semantic gap" between search phrases and item's information. For example, if a web page mentions "stable lightweight shoes", a smart algorithm should be able to detect the close relationship between "stable lightweight shoes" and "running shoes".

Our work attempts to bridge this gap, in the context of highly descriptive and data rich product information, such as "Nike stable lightweight shoes ..." and two or three word long popular search phrases such as "running shoes".

In order to bridge this gap, we need to substitute original search phrase, such as "running shoes" with a definition that contains other parameters corresponding to technical characteristics of matching product descriptions.

Previous studies on mining definitions have focused on query substitution [1], advertising-page ranking by using genetic programming [2], or syntactic improvements such as spelling changes, synonym substitutions, taxonomy based generalizations and specialization of search-phrases [3]. A different approach adopted in recent studies [4], [5], [6] is to utilize frequent item-set mining algorithms to identify alternative definitions. However, such algorithms generate large numbers of frequent item-sets as possible definitions and we observe the need for more robust filtering algorithms that can use both keyword-matching and a sample of product descriptions to rank the candidate definitions and enlist the most promising definition phrases.

In this paper, we develop a model for representing the dependence among phrases and present a robust automated algorithm that can use a small corpus of keyword matching and a random sample of other product descriptions for a popular search-phrase to effectively mine and rank alternative but highly relevant *search-phrase definitions*.

Basically, we reduce the problem significantly by modeling the joint distribution as a product of conditional distributions, modeled as a Markov Random Field. We observe that, in textual product descriptions, the critical statistical relationship among terms is the co-occurrence. This kind of correlation is undirected in nature. Moreover, it is hard to determine whether some term causes the presence of another term in a description in general. Hence, we propose an undirected graphical model that can be succinctly used to model the dependence among objects.

Hence we propose an undirected graphical model that can be succinctly used to model the dependence among objects and it is also suitable for modeling and reasoning with the term dependencies within highly descriptive data-rich product descriptions.

In our approach, an undirected graphical model is a graph $G(V, E)$ in which V is the set of random variables, E is the set of edges connecting pairs of random variables. In our context, the set of random variables corresponds to the set of terms. The rule in the form of $target\ phrase \leftarrow term_1 \wedge term_2 ... \wedge term_n$ is said to be a *Phrase Definition rule*

if given G and a real number τ, $(0 < \tau \leq 1)$, the conditional probability $P(target\ phrase|term_1,\ term_2, ...,\ term_n) \geq \tau$. The *target phrase* is the phrase to be defined using other terms on the right hand side of the rule.

Our experimental results illustrate that our technique yields high recall (above 90% on average) as well as high precision (above 84 % on average).

Our contributions are: (1) Define a model for estimating conditional probabilities via sampling true joint distribution among phrases, (2) Develop an algorithm that utilizes these conditional probabilities to find relevant definitions of search phrases that in turn can serve as alternative highly relevant definitions.

II. Problem Description

In this section, we intuitively motivate the use of probabilistic inference techniques to solve our problem. Given a search phrase, the descriptive information about that phrase is an important clue for efficiently finding the relevant items (item descriptions in catalog databases) that match the search phrase. We adopt the standard statistical IR assumption that inter-term relevance is reflected in the co-occurrence statistics over the corpus and use it to generate definition rules in the above format.

We motivate the modeling using an example as follows. The product catalogs of a general web-based vendor such as Amazon or Ebay contain hundreds of thousands of records of all kinds of products ranging from books to video games, apparel or electronics. These products can be listed by many other online sellers and individuals. The descriptions of the products can be very diverging and may not have a uniform structure even for the same kind of products or products from similar categories. In order to match the relevant items to search phrases, the use of the classification technique as a filter does not suffice. For example, although "running shoes" and "dress shoes" are in the same category "men shoes", these two types of shoes can be described by totally different sets of terms. Another example is that "running shoes" and "tennis shoes" are in a narrower category which is "athletic shoes" and yet, their descriptions are still different. Furthermore, each seller or individual can have his or her own way to describe the features as well as the functions and the condition of the products. We may experience this using Ebay. This web site has a "user feedback" solution that offers shoppers to browse different product descriptions with or without specific keywords/terms. But in the scenario of online advertising, an accurate automated technique is a must. Even the set of terms used to describe a specific product is finite, the order and the form of terms can be different in different product descriptions given by different describers. This inconsistency is due to the fact that there are too many ways that a product can be described and this heterogeneity causes difficulties for advertisers. Irrelevant or missing annotation leads to a loss of revenue for advertisers. Let us assume that there exists implicitly an underlying true joint distribution of features that are used to describe what *luxury bedding* is. Different vendors have their own different ways to describe the products that they consider as "luxury bedding". We can view each isolated description as a sample of that distribution. It is important to note that we may not need to make any assumption about this distribution. Finding the true joint distribution helps us compute correct conditional probabilities of the form $p(luxury\ \ bedding|unknown\ \ features)$. These unknown features that we are interested in are the terms that can be used to describe luxury bedding. If these terms are identified, they can be used to annotate relevant web pages or relevant records to increase the chance of getting more hits (of the pages or records). This also helps customers find more relevant product pages given their queries. Given these observations, we need to take into account these hypotheses when we formulate the problem in the next section: 1) Term dependency model does not need to incorporate the information about the order of terms and several context features since the product descriptions, as observed above, are free-style and no context features can entirely capture correctly the product information. We argue that if the model is simple and works, then let it be. Otherwise, we may need to encode more features into the model to increase its efficiency. 2) However, the model may need to take into account the distance among terms within a document or a product description.

In order to explore the above type of descriptions, we model the dependencies among terms in the databases of product descriptions. In the next section, we discuss an undirected graphical model known as *Markov random fields (MRF)* or *Markov network* to model our problem and develop techniques for mining phrase definition rules.

III. Model

A Markov Random Field (MRF) has several components: a set $V = \{1, ..., m\}$ of site v; a neighborhood system $N = \{N_v | v \in V\}$ in which each N_v is a subset of sites in V describing the neighbors of v. These two sets form a undirected graph called G; a field (or set) of random variables is denoted as $\mathbf{X} = \{X_v | v \in V\}$. In our problem, given a set of terms, the set V of vertices in G is the set of the terms. Two terms present two neighbor nodes in the MRF if the co-occurrence of these two terms in the whole collection passes a threshold δ. Each random variable X_v takes a value x_v in some set $L = \{l_1, ..., l_r\}$ of the possible labels. When sampling method is used to recover the joint distribution among terms, for each node in the MRF (i.e. each term), the corresponding random variable can only be 0 or 1 (whether that term appears in a specific sample). Therefore, $L = \{0, 1\}$; a set of potential functions φ_k (also called factors or clique potentials) one for each clique k in G. Note that a clique is a *maximal clique* if it is not contained within any other clique. Each φ_k maps from all possible joint assignments (to the elements to k) to non-negative real numbers. The joint probability of all random variables in the MRF denoted as $Pr(X = \mathbf{x})$ is computed as:

$$Pr(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \prod_{k \in C(G)} \varphi_k(\mathbf{x}_k)$$

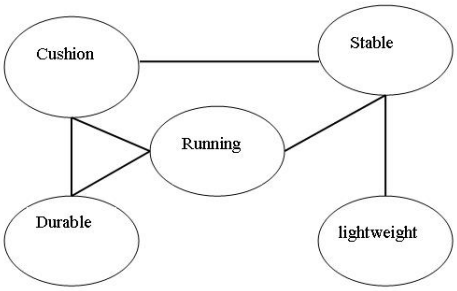

Fig. 1. Sample MRF with {running,cushion,durable}, {running,cushion,stable} are maximal 3-cliques and {lightweight,stable} is a maximal 2-clique

where $C(G)$ is the set of cliques in G, x_k is the values of random variables in kth clique, $Z = \sum_{x \in X} \prod_{c \in C(G)} \varphi(c)$ is the normalized factor of the distribution. For each clique in G, we define potential function φ_k as:

$$\varphi_k = log \left[\frac{df(q_1, ..., q_i)}{|C|} \right]$$

where $df(q_1, ..., q_i)$ denotes the number of times the terms $q_1, ..., q_i$ occur together in the collection and $|C|$ is the number of documents (product descriptions) in the collection. By saying "terms $q_1, ..., q_i$ occur together", it means these terms appear together in a description but they must appear in the same sentence. However, this distance constraint is relaxed for the target phrase. More specifically, the target phrase can be anywhere in the description. Other terms must appear in the same sentence to be considered as co-occurring terms.

As stated in Section I, in order to learn the phrase definition rules $target\ phrase \leftarrow term_1 \wedge term_2 ... \wedge term_n$, we want to find the set of terms $term_1,\ term_2, ..., term_n$ that maximize the probability $P(target\ phrase | term_1,\ term_2, ...,\ term_n)$. More specifically, we find the set of variables $\{X_1, ..., X_n\}$ so that

$$f(X) = Pr(X | X_1, ..., X_n) = Pr(X | A, X_n) \qquad (1)$$

is maximized. Here, X is the random variable in the MRF and presents the target phrase, X_is are the random variables in the MRF and present other terms and A=($X_1, ..., X_{n-1}$) In order to compute the above type of probability, we do the inference in the MRFs by developing an inference algorithm utilizing Markov Chain Monte Carlo technique [7] and derive another algorithm to mine the rules with highest probability values. The mined rules are ranked based on the probability values of corresponding conditional probabilities in the form 1.

IV. Related Works

Ribeiro-Neto et. al. [8] study various strategies to match pages to ads based on extracted keywords. Their approach employs the vector space model to represent ads and pages. The first five strategies proposed in their study match the pages and the ads based on the cosine theta of the angle between the ad vector and the page vector. To identify the important part of the ad, the authors use various ad sections (bid phrase, title, body) as a basis for the ad vector. In a follow-up work [2], the authors use genetic programming paradigm to develop the ranking algorithm for ads. The results show that genetic programming finds matching functions that significantly improve the matching compared to the best method (without page expansion) reported in [8]. In a more recent work by Broder et. al. [9], the authors introduce the class taxonomy to classify ads. This phase acts as a filter before conducting the phrase extraction process and this technique shows a clear improvement. In these previous studies, the accuracy or the relevance of the ads is compared with documents using cosine theta which is different from our study where we find the alternative definitions for the search phrase. The alternative query terms are matched against the product description to judge the relevance by domain experts.

V. Discussion

We performed an extensive evaluation of our rule miner system, which relies on MRF for term dependency and sampling technique to estimate the goodness of the candidate alternative queries (the terms in the right high side of the definition rules). We do not report our results because of space constraints. As a summary of our conclusions, our approach produced rules that yield high recall as well high precision. As expected, our approach also avoids placing irrelevant items in high ranked positions. This criterium is very important in the context of Content Targeted Advertising computing.

VI. Acknowledgment

The first author is funded in part by a grant from the Vietnam Education Foundation (VEF). The opinions, findings, and conclusions stated herein are those of the authors and do not necessarily reflect those of VEF

References

[1] R. Jones, B. Rey, O. Madani, and Greiner, "Generating query substitutionss," in *Proceedings of the 15th WWW Conference*, 2006, pp. 387–396.

[2] A. Lacerda, M. Cristo, M. A. Goncalves, W. Fan, N. Ziviani, and B. Ribeiro-Neto, "Learning to advertise," in *SIGIR '06*. New York, NY, USA: ACM Press, 2006, pp. 549–556.

[3] E. Terra and C. L. Clarke, "Scoring missing terms in information retrieval tasks," in *CIKM '04*. New York, NY, USA: ACM Press, 2004, pp. 50–58.

[4] H. Davulcu, H. V. Nguyen, and V. Ramachandran, "Boosting item findability: Bridging the semantic gap between search phrases and item information," *Enterprise Information Systems VII*, vol. VII, pp. 215–222, 2006.

[5] B. Liu, C. W. Chin, and H. T. Ng, "Mining topic-specific concepts and definitions on the web," in *WWW '03*. New York, NY, USA: ACM Press, 2003, pp. 251–260.

[6] H. V. Nguyen, H. Davulcu, and V. Ramachandran, "Boosting item findability: Bridging the semantic gap between search phrases and item description," *International Journal of Intelligent Information Technologies*, vol. 2, pp. 1–20, 2006.

[7] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan, "An introduction to mcmc for machine learning," *Machine Learning*, vol. 50, pp. 5–43, 2003.

[8] B. Ribeiro-Neto, M. Cristo, P. B. Golgher, and E. S. de Moura, "Impedance coupling in content-targeted advertising," in *SIGIR '05*. New York, NY, USA: ACM Press, 2005, pp. 496–503.

[9] A. Broder, M. Fontoura, V. Josifovski, and L. Riedel, "A semantic approach to contextual advertising," in *SIGIR '07*. New York, NY, USA: ACM Press, 2007, pp. 559–566.