Leveraging Wikipedia concept and category information to enhance contextual advertising

Wu, Zongda; Xu, Guandong; Pan, Rong; Zhang, Yanchun; Hu, Zhiwen; Lu, Jianfeng

Published in:
Proceedings of the 20th ACM international conference on Information and knowledge management

Publication date:
2011

Document Version
Accepted author manuscript, peer reviewed version

Citation for published version (APA):
Wu, Z., Xu, G., Pan, R., Zhang, Y., Hu, Z., & Lu, J. (2011). Leveraging Wikipedia concept and category information to enhance contextual advertising. In B. Berendt, A. D. Vries, W. Fan, C. Macdonald, I. Ounis, & I. Ruthven (Eds.), *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 2105-2108). Association for Computing Machinery. http://doi.acm.org/10.1145/2063576.2063901

Leveraging Wikipedia Concept and Category Information to Enhance Contextual Advertising

Zongda Wu
Wenzhou University, China
zongda1983@163.com

Guandong Xu
Victoria University, Australia
guandong.xu@vu.edu.au

Rong Pan
Aalborg University, Denmark
rpan@cs.aau.dk

Yanchun Zhang
Victoria University, Australia
yanchun.zhang@vu.edu.au

Zhiwen Hu
Wenzhou University, China
sunneyhu@gmail.com

Jianfeng Lu
Zhejiang Normal University
lujianfeng@zjnu.cn

ABSTRACT

As a prevalent type of Web advertising, contextual advertising refers to the placement of the most relevant ads into a Web page, so as to increase the number of ad-clicks. However, some problems of homonymy and polysemy, low intersection of keywords etc., can lead to the selection of irrelevant ads for a page. In this paper, we present a new contextual advertising approach to overcome the problems, which uses Wikipedia concept and category information to enrich the content representation of an ad (or a page). First, we map each ad and page into a keyword vector, a concept vector and a category vector. Next, we select the relevant ads for a given page based on a similarity metric that combines the above three feature vectors together. Last, we evaluate our approach by using real ads, pages, as well as a great number of concepts and categories of Wikipedia. Experimental results show that our approach can improve the precision of ads-selection effectively.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Theory

1. INTRODUCTION

PwC[1] predicts that Web advertising will become the 2nd largest advertising medium in America after television within the next 4 years, and its spending will increase from 24.2 billion dollars in 2009 to 34.4 billion dollars in 2014. As an important type of Web advertising, contextual advertising aims to embed the most relevant ads into a page, so as to increase the number of ad-clicks. Most contextual advertising approaches were based on the keyword matching, which estimated the ad relevance based on the co-occurrence of the same keywords between pages and ads [1, 2]. However, as pointed out in [3, 4], the keyword matching may lead to the problems such as homonymy and polysemy, low intersection of keywords, and context mismatch, consequently, degrading its effectiveness.

To solve these problems, in the area of text classification, a new approach called Wikipedia matching was proposed in [5, 6, 7], which uses Wikipedia, the largest knowledge base, as a reference model to enrich the content representation of text documents, so as to improve the accuracy of similarity computation among documents. It was then applied into contextual advertising in [4], and has been proved its effectiveness on solving the problems encountered in the keyword matching. However, this approach may lead to the following problems that decrease its effectiveness in practical contextual advertising. (1) Its limited coverage over semantic concepts. To enhance performance, it only chooses a small part of articles from Wikipedia as a reference model. As a result, for many pages not properly characterized by the reference articles, it is impossible to find out the articles that share the same topics with the pages, thus, leading to the returning of irrelevant ads for the pages. (2) Its time-consuming performance. To solve the limited coverage over semantics, a very straightforward way is to choose a sufficient number of reference articles from Wikipedia. However, this will result in a seriously decreased performance, because the time spending of fulltext matching between all the reference articles and the ads (or pages) is very high.

In order to better balance effectiveness and efficiency, we in this paper present a new contextual advertising approach by combining Wikipedia knowledge with the keyword matching, which considers two aspects of similarity between ads and pages, where the keyword-based similarity captures the textual commonness, while the Wikipedia-based similarity measures the relatedness from the semantic perspectives of concepts and categories. Its process consists of the following three steps. First, we choose a sufficient number of articles from Wikipedia, to cover as many concepts as possible. Next, we map each ad (as well as each page) into a keyword

[1] PricewaterhouseCoopers - http://www.pwc.com

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...$10.00.

vector, a concept vector and a category vector. Last, combining the three feature vectors together, we make the top-N ads selection. The experimental results show that our approach can improve the accuracy of ads selection effectively. And, due to avoiding time-consuming fulltext matching between all the reference articles and the pages (or ads), our approach also obtains a good running performance.

2. RELATED WORK

The keyword matching approach estimated the ad relevance by analyzing the co-occurrence of the same keywords within ads and within pages. One of the recent results on applying the keyword matching into contextual advertising was presented in [2], whose main idea was to use a technique called "Impedance Coupling Strategy" to augment a page with additional keywords from other similar pages, so as to overcome the problem of low intersection of keywords between pages and ads. Under the assumption that an ad can be seen as a noisy translation of a page, it was proposed in [8] to select the ads that provide the best translation for a page. In [9], it was proposed to leverage sentiment detection to improve contextual advertising. In [10], the authors proposed to use lexical graphs created from web corpora as a means of computing improved content similarity metrics between ads and pages. However, as pointed out in [4, 5, 6], the main drawback of the keyword matching approach is that it may lead to the problems of homonymy and polysemy etc., resulting in degrading the relevance of selected ads to their pages.

For solving the problems of homonymy and polysemy etc., the Wikipedia matching was proposed, whose main idea is to leverage the Wikipedia as an intermediate reference model to enhance the semantic representation of text documents and thus improve the precision of similarity measure among documents. In [4], a solution to contextual advertising was proposed by using the Wikipedia matching. In this solution, a group of reference articles is first chosen from Wikipedia. Next, through using the articles as the intermediate reference model on which the ads and the page is re-expressed as feature vectors, the ads that exhibit more relevance to a targeted page are determined, and a ranking function to select the most relevant ads to the page is constructed. In [5, 6, 7], a similar method was proposed aiming to textual document clustering. It was proposed in [11] to use Wikipedia to understand a user's query intent, without the need to collect large quantities of examples to train an intent classifier. However, the main drawback of the traditional Wikipedia matching is due to the problems that we have mentioned in Section 1 (i.e., the problems of limited coverage of semantic concepts and time-consuming performance) can dramatically degrade the relevance of selected ads with their pages.

3. METHODOLOGY

Figure 1 shows the framework of our approach to improving contextual advertising using the Wikipedia knowledge, where each ad or page is mapped into a keyword vector, a concept vector and a category vector, and then the three vectors are combined together to measure the similarity between a page and an ad. This process is similar to that used in [6] for clustering, but has a different way for constructing concept vector and category vector. We below introduce how to construct such two types of feature vectors.

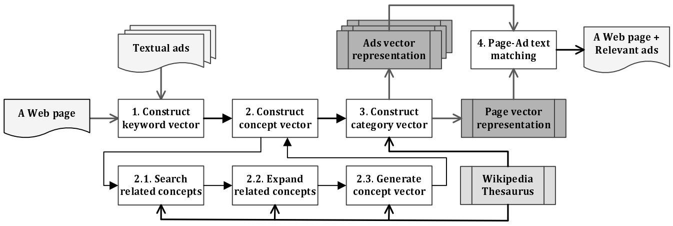
Textual ads
Ads vector representation
4. Page-Ad text matching
A Web page + Relevant ads
A Web page
1. Construct keyword vector
2. Construct concept vector
3. Construct category vector
Page vector representation
2.1. Search related concepts
2.2. Expand related concepts
2.3. Generate concept vector
Wikipedia Thesaurus

Figure 1: Contextual advertising framework

3.1 Constructing Concept Vector

As shown in Figure 1, the process of generating a concept vector includes the following three steps: (1) search related concepts appearing in an ad (or a page); (2) add concepts semantically related to the previous ones; and (3) generate a concept vector based on the frequency values of the concepts that are semantically related to the ad. We below introduce these steps, respectively.

(1) Searching Related Concepts. This step aims to determine a set of related concepts for an ad (or a page) and to count their frequency values in the ad. First, we scan a given ad to find all the Wikipedia titles that appear in the ad. Such titles are called related titles, and the concepts associated with them are called related concepts. We search related titles using a similar method mentioned in [7]. Once we determine a set of titles related to an ad, actually, we determine a set of concepts related to the ad. Next, we compute the frequency of each related concept to the ad. Let $\mathbf{count}(\boldsymbol{t}, \boldsymbol{a})$ be the frequency of a related title $\boldsymbol{t}$ appearing in a page $\boldsymbol{a}$, which is determined above. Let $\mathbf{cots}(\boldsymbol{t})$ be a set of concepts associated with the title $\boldsymbol{t}$, i.e., each of which satisfying that at least one of its titles is identical to $\boldsymbol{t}$. Then, the frequency of occurrences of any concept $\boldsymbol{c}$ related to $\boldsymbol{t}$ in $\boldsymbol{a}$ is computed as follows: if $\boldsymbol{t}$ is not a title of the concept $\boldsymbol{c}$, then $\mathbf{count}(\boldsymbol{c}, \boldsymbol{t}, \boldsymbol{a}) = 0$; and otherwise,

$$\mathbf{count}(\boldsymbol{c}, \boldsymbol{t}, \boldsymbol{a}) = \left(\frac{\mathbf{sim^k}(\boldsymbol{c}, \boldsymbol{a})}{\sum\limits_{c' \in \mathbf{cots}(\boldsymbol{t})} \mathbf{sim^k}(c', \boldsymbol{a})} \right) \cdot \mathbf{count}(\boldsymbol{t}, \boldsymbol{a})$$

where $\mathbf{sim^k}(\boldsymbol{c}, \boldsymbol{a})$ denotes the similarity between $\boldsymbol{a}$ and $\boldsymbol{c}$ computed using the traditional keyword matching.

In Wikipedia, a concept may have several titles [11], and thus to count the actual frequency of occurrences of a concept in a given ad, we need to sum up the frequency value of the concept related to each one of its titles appearing in the ad. Let $\mathbf{titles}(\boldsymbol{a})$ be all the titles appearing in an ad $\boldsymbol{a}$ (i.e., the related titles of $\boldsymbol{a}$). Then, the frequency of any concept $\boldsymbol{c}$ appearing in $\boldsymbol{a}$ is computed as follows:

$$\mathbf{count}(\boldsymbol{c}, \boldsymbol{a}) = \sum\nolimits_{\boldsymbol{t} \in \mathbf{titles}(\boldsymbol{a})} \mathbf{count}(\boldsymbol{c}, \boldsymbol{t}, \boldsymbol{a})$$

Now, we determine a set $\mathbf{C}(\boldsymbol{a}) = \{\mathbf{count}(\boldsymbol{c_j}, \boldsymbol{a})\}_{j=1}^{N_a^c}$, consisting of frequency of each related concept in an ad $\boldsymbol{a}$, where N_a^c is the number of related concepts in $\boldsymbol{a}$. Similarly, we determine a set $\mathbf{C}(\boldsymbol{p}) = \{\mathbf{count}(\boldsymbol{c_j}, \boldsymbol{p})\}_{j=1}^{N_p^c}$ for a page $\boldsymbol{p}$.

(2) Expanding Related Concepts. There are fewer keywords in an ad $\boldsymbol{a}$ than a generic document, due to its limited size, so $\mathbf{C}(\boldsymbol{a})$ is of a smaller size, resulting in the decreased precision of similarity computation between pages and ads. However, two Wikipedia concepts are hyperlinked to each other, generally, denoting that the two concepts are semantically related to each other [11]. So, this step aims to expand and enrich $\mathbf{C}(\boldsymbol{a})$, based on the hyperlinks within concepts. Actually, the expansion is a process of breadth-first graph traversal that starts from the concepts associated with $\mathbf{C}(\boldsymbol{a})$. First, for each concept $\boldsymbol{c}$ in $\mathbf{C}(\boldsymbol{a})$, we obtain all the concepts hyperlinking to $\boldsymbol{c}$, noted as $\mathbf{cots}(\boldsymbol{c})$. Second, we calculate the frequency value of each concept $\boldsymbol{e}$ in $\mathbf{cots}(\boldsymbol{c})$:

$$\mathbf{count}(\boldsymbol{e},\boldsymbol{a}) = \left(\frac{\mathbf{num^k}(\boldsymbol{e},\boldsymbol{c})}{\sum\limits_{e' \in \mathbf{cots}(\boldsymbol{c})} \mathbf{num^k}(e',\boldsymbol{c})} \right) \cdot \mathbf{count}(\boldsymbol{c},\boldsymbol{a})$$

where $\mathbf{num^k}(\boldsymbol{e},\boldsymbol{c})$ denotes the number of hyperlinks between the concepts $\boldsymbol{e}$ and $\boldsymbol{c}$.

Third, if $\mathbf{count}(\boldsymbol{e},\boldsymbol{a})$ is greater a given parameter μ that is assigned by users to control the depth of graph traversal, then $\mathbf{count}(\boldsymbol{e},\boldsymbol{a})$ would be added into $\mathbf{C}(\boldsymbol{a})$. Such a process is kept on until all the concepts (including the new added ones) associated with $\mathbf{C}(\boldsymbol{a})$ are traversed. Similarly, we can expand $\mathbf{C}(\boldsymbol{p})$ for a page $\boldsymbol{p}$.

(3) Generating Concept Vector. After the above step, $\mathbf{C}(\boldsymbol{a})$ would be expanded with more concepts that are all semantically related to the ad $\boldsymbol{a}$, i.e., obtaining a set of frequency values of concepts for $\boldsymbol{a}$. Similarly to the traditional keyword matching, based on $\mathbf{C}(\boldsymbol{a})$, we can generate a concept vector for $\boldsymbol{a}$, which consists of the *tf-idf* values [12] of all the concepts in $\mathbf{C}(\boldsymbol{a})$. We also generate a concept vector for a page $\boldsymbol{p}$. Last, we compute the concept-based semantic similarity $\mathbf{sim^c}(\boldsymbol{a},\boldsymbol{p})$ between $\boldsymbol{p}$ and $\boldsymbol{a}$, by using the two concept vectors.

3.2 Category Vector Construction

In this subsection, by combining concept vector and the hierarchical relation between concepts and categories or within categories, we describe how to generate a category vector for an ad (or a page), and leverage it further to enrich the semantic representation for pages and ads.

Let $\mathbf{cots}(\boldsymbol{d})$ be all the concepts that belong to a category $\boldsymbol{d}$, and $\mathbf{cats}(\boldsymbol{d})$ all the immediate subcategories that belong to $\boldsymbol{d}$ (i.e., there is a hyperlink from each category in $\mathbf{cats}(\boldsymbol{d})$ to $\boldsymbol{d}$). $\mathbf{cots}(\boldsymbol{d})$ and $\mathbf{cats}(\boldsymbol{d})$ are determined by the hierarchical categorization system in Wikipedia. Then, we define the related frequency of any category $\boldsymbol{d}$ appearing in an ad (or a page) $\boldsymbol{a}$ as follows:

$$\mathbf{count}(\boldsymbol{d},\boldsymbol{a}) = \sum_{c \in \mathbf{cots}(\boldsymbol{d})} \frac{\mathbf{count}(c,\boldsymbol{a})}{\alpha_1} + \sum_{d' \in \mathbf{cats}(\boldsymbol{d})} \frac{\mathbf{count}(d',\boldsymbol{p})}{\alpha_2}$$

where α_1 and α_2 are two attenuation coefficients, used to balance importance of frequency values of categories in different depths.

Now, we determine a set $\mathbf{D}(\boldsymbol{a}) = \{\mathbf{count}(\boldsymbol{d_j},\boldsymbol{a})\}_{j=1}^{N_a^d}$, consisting of frequency of each category related to an ad $\boldsymbol{a}$, where N_a^d is the number of related categories in $\boldsymbol{a}$. The generation of $\mathbf{D}(\boldsymbol{a})$ for an ad $\boldsymbol{a}$ is also a process of breadth-first graph traversal starting from the concepts associated with $\mathbf{C}(\boldsymbol{a})$. However, the traversed graph consists of (1) nodes, concepts and categories, and (2) edges, the hyperlinks within concepts and categories, as well as the hyperlinks within categories. Furthermore, based on $\mathbf{D}(\boldsymbol{a})$, we can generate a category vector for the ad $\boldsymbol{a}$, which consists of the *tf-idf* values of all the categories in $\mathbf{D}(\boldsymbol{a})$. Similarly, we also generate a category vector for a page $\boldsymbol{p}$. Last, we can compute the category-based semantic similarity $\mathbf{sim^d}(\boldsymbol{a},\boldsymbol{p})$ between $\boldsymbol{p}$ and $\boldsymbol{a}$, by using the two category vectors.

Item	Number
General pages in dataset	50
Ambiguous pages in dataset	27
Textual ads in dataset	10,244
Wikipedia articles (Wikipedia concepts)	260,000
Wikipedia categories	12,000

Table 1: Dataset characteristics

Item	Explanation
K	solely based on keyword vectors
C	solely based on concept vector
D	solely based on category vector
KC	based on keyword vector and concept vector
KD	based on keyword vector and category vector
CD	based on concept vector and category vector
KCD	based on keyword vector, concept vector and category vector

Table 2: 7 contextual advertising strategies based on different combinations of feature vectors

3.3 Similarity Computation

Now, each ad (or page) has been represented as three feature vectors: a keyword vector, a concept vector and a category vector. So, when measuring similarity between an ad and a page, we combine the similarity values calculated using the three feature vectors. For a given page $\boldsymbol{p}$ and an ad $\boldsymbol{a}$, the similarity between $\boldsymbol{p}$ and $\boldsymbol{a}$ can be computed as follows:

$$\mathbf{sim}(\boldsymbol{a},\boldsymbol{p}) = \alpha \cdot \mathbf{sim^k}(\boldsymbol{a},\boldsymbol{p}) + \eta \cdot \mathbf{sim^c}(\boldsymbol{a},\boldsymbol{p}) + \zeta \cdot \mathbf{sim^d}(\boldsymbol{a},\boldsymbol{p})$$

where the coefficients α, η and ζ indicate the importance of concept vector and category vector in measuring the semantic similarity between the page and the ad, which also can be used to balance the keyword matching and the Wikipedia matching. And $\alpha + \eta + \zeta = 1$.

4. EXPERIMENTS

We evaluated experimentally our approach using a dataset that contains 50 generic pages, 27 ambiguous pages and 10, 224 ads, more detailed characteristics of which are shown as Table 1. For each page, we collected human judgment scores that describe the relevance of ads selected by each of the candidate strategies (see Table 2). The human judgment

Strategy	K	C	D	KC	KD	CD	KCD
Time (ms)	45	290	765	293	775	758	780

Table 3: Running performance for the 7 strategies

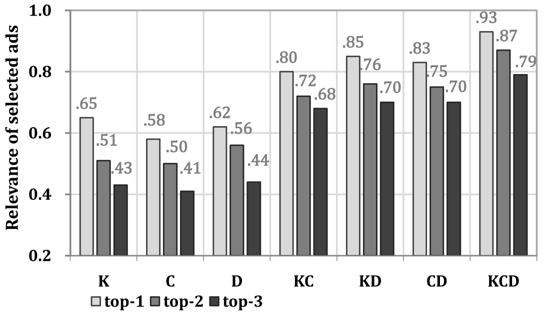

Figure 2: Average relevance for the ads selected by the 7 candidate strategies for general pages

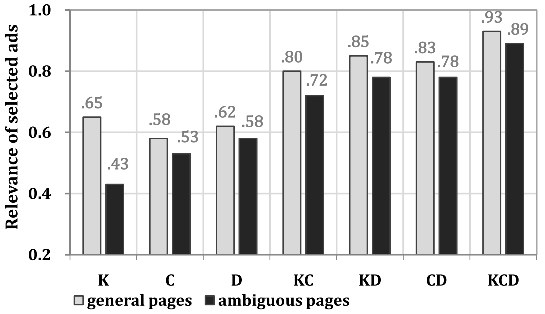

Figure 3: Average relevance for the ads selected by the 7 strategies for general or ambiguous pages

scores for the relevance of embedded ads to a page were determined by using a similar method in [5, 6], and were completed by at least two human assessors in a scale between 0.0 to 1.0.

The first group of experiments aimed to evaluate the execution time of selecting relevant ads for pages. Here, the work of generating feature vectors for all the ads, has been completed in advance (i.e., completed offline). In each ads-selection for a page, we only concern the execution time consumed by (1) generating the feature vectors for the page, and (2) computing the similarity between the page and each ad to choose the most relevant ads. The results are presented in Table 3. The second group of experiments aimed to evaluate the relevance scores of embedded ads to their pages. In our experiments, we invited evaluation assessors to mark score for each ad based on the relevance of the ad to its page, and then averaged the relevance scores given by the evaluation assessors. The results are shown in Figure 2. In the third group of experiments, we have chosen a special dataset that consists of 27 ambiguous pages. In the pages, there are many ambiguous keywords, such as Puma (company versus lion), Rock (person versus music), Driver (software versus car), Game (software versus sports), Window (OS versus glass) and so on. The experimental results of running the seven candidate strategies over the ambiguous pages are shown in Figure 3.

5. CONCLUSIONS

In order to improve contextual advertising, we in this paper presented a new approach by incorporating the Wikipedia concept and category information into the traditional keyword matching to enrich the content representation of pages and ads. We described how to map each ad (or page) into a keyword vector, a concept vector and a category vector, as well as how to combine the three feature vectors together for making the top-N ads selection.

From the experimental results in Section 4, we have the following conclusions. (1) Our approach obtains a satisfactory running performance (the time spending of ads-selection for a page is less than 1000 ms). This is due to avoiding to conduct the time-consuming fulltext matching operation between pages and all the referenced articles, which is used in the previously published Wikipedia contextual advertising approach. (2) Our approach obtains a better effectiveness, i.e., it can well improve the accuracy of ads-selection (the relevance score of embedded ads to their pages is generally greater than 0.8). This is due to that we use the Wikipedia knowledge to enrich the semantic representation of pages and ads, and use them to measure the semantic similarity between pages and ads; while, compared to the surface text information contained in pages and ads, the semantic information has a better stability, i.e., it can reflect out the similarity between pages and ads more accurately.

Funding: This work is supported by the National Science Foundation of the Zhejiang Province of China, under grant Y1100137, the Education Research Program of the Zhejiang Province of China, under grant Y201016197, the Science and Technology Program of the Wenzhou City of China, under grant 2009G0339 and grant S20100055.

6. REFERENCES

[1] A. Lacerda, M. Cristo, M. G. Andre et al. Learning to advertise. In *SIGIR*, 2006.

[2] B. Ribeiro-Neto, M. Cristo, P. B. Golgher et al. Impedance coupling in content-targeted advertising. In *SIGIR*, 2005.

[3] A. Anagnostopoulos, A. Broder, E. Gabrilovich et al. Just-in-time contextual advertising. In *CIKM*, 2007.

[4] A. N. Pak, and C.-W. Chung. A Wikipedia matching approach to contextual advertising. *WWWJ*, 13: 251-274, 2010.

[5] J. Hu, L. J. Fang, Y. Cao, H.-J. Zeng et al. Enhancing text clustering by leveraging Wikipedia semantics. In *SIGIR*, 2008.

[6] X. H. Hu, X. D. Zhang, C. M. Lu et al. Exploiting Wikipedia as external knowledge for document clustering. In *SIGKDD*, 2009.

[7] P. Wang, J. Hu, H.-J. Zeng et al. Using Wikipedia knowledge to improve text classification. *KAIS*, 19: 265-281, 2009.

[8] V. Murdock, M. Ciaramita and V. Plachouras. A noisy-channel approach to contextual advertising. In *SIGKDD* workshops 2007.

[9] T.-K. Fan and C.-H. Chang, "Sentiment-oriented contextual advertising". *KAIS*, 23:321-344, 2010.

[10] S. Papadopoulos, F. Menemenis, Y. Kompatsiaris et al. Lexical graphs for improved contextual ad recommendation. In *ECIR*, 2009.

[11] J. Hu, G. Wang, F. Lochovsky et al. Understanding user's query intent with Wikipedia. In *WWW*, 2009.

[12] H. C. Wu, R. W. P. Luk, K. F. Wong et al. Interpreting TF-IDF term weights as making relevance decisions. ACM TOIS 26, 2008.