A Keyword Extraction Based Model for Web Advertisement

Ning Zhou[1], Jiaxin Wu[1,2], Shaolong Zhang[2]
[1] Research Center of Information Resources, Wuhan University, Wuhan, 430072, China
[2] School of Information Management, Wuhan University, Wuhan, 430072, China
logofish@126.com

Abstract. In this paper, a keyword extraction based model is proposed to deal with web advertisement. In our model, we take web advertisement as an information retrieval problem. Web page and advertisement are firstly represented with a simple data structure which will be the source file for keyword extraction based on χ^2-measure for single document. Later we get two vectors to make a retrieval process with a specific similarity function. This model is suitable for common cases of web advertisement. It supports the web page selection in view of advertisement as well as the advertisement selection for specific web page.

1 Introduction

With the rapid growth of internet, web becomes a significant medium of our daily life. More and more people are accustomed to read news, publish their blogs, search information on internet. Internet is turning to be a virtual society where global user exchange and share information as well as service.

Where there is a medium, where there will be advertisement. So web is growing to be a vital place to post advertisements for each company which could be traditional company as well as innovative company. It will be meaningful to construct the model for the post of advisement on web pages. This model must make： 1) An advertisement is post on a web page which latent customer of advisement product will browse. 2) A web page should better include most relevant advertisement to maximize the probability for the user to click. 3) There should be a quantity to show which site or page is the best one to post ads.

There are some obvious problems on web advertisement: 1) the advertisement is not related with the content of web pages, and user has no interest to click.2) Too

much unrelated ads make the browsing process to be disappointing. 3) Ads are posted on a large quantity of pages but bring few clicks. 4) Company can't evaluate which site or page is the most beneficial one to post ads until the end of advertisement period.

Many papers pay attention to the web advertisement, but most of them do that in a qualitative way. They give some principles in how to choose and organize the posting strategy, but seldom bring quantitative method to evaluate the ads posting. This paper pays more attention to the data structure of web page and advertisement. Later, based on the data structure, we give a quantitative evaluation of similarity between web page and advertisement which will make it easier to choose the ads for web page and in return, to select web pages for ads.

This paper is organized in 4 sections. Section 2 discusses several recent approaches in web advertisement. Section 3 describes our model in detail. Section 4 gives our experiment result and the work to enhance our model in future.

2 Related Work

Online advertising continues to be a significant source of income for many Internet-based organizations. Banner advertisement is important advertising style for famous websites. Ali[4] extends the problem of scheduling banner advertisement to a more realistic setting, where the customer is allowed to specify a set of acceptable display frequency, the Lagrangian decomposition-based solution was presented to provide good schedules in a reasonable period of time.

Thawani[5] proposed a system for the selection and presentation of advertisements based on detected program events and profiles contained in the home information system. In this system, a comprehensive event prediction and event analysis is performed based on which relevant ads are either selected for transport stream insertion or for caching purposes.

Vincent[3] describes a new advertising agent based on user information. In this agent, the user's interests are discovered by the Order Pattern Mining algorithm and represented in user's profiles with the Gaussian curve transformation. User's profile is used to implement an effective and efficient advertisement mechanism.

Because user's profile is hard to collect, especially user doesn't want to or can't express themselves clearly about their interests to large quantity of products and advertisements, so we put the profile aside. Also the keywords advertisements still a very tidy and meaningful style which is used as the main advertisement style of many search engines, such as Google, Baidu, Yahoo. So we take the advertising as the problem of information retrieval problem of the match between advertisement and web page, and in the end, we could get the similarity as well as a ranked list which is very suitable to the selection of advertisement.

3 Keyword Extraction Based Model

3.1 Data Structure of web Page and Advertisement

In this model, only two kinds of sources are considered, web page and advertisement. We form web page as 5 fields: classID, url, title, meta keywords, text.

classID stands for the category id of given web page in a predefined class catalog which must be carefully designed to cover most of pages; URL is the url of web page such as http://sports.sina.com.cn/; Title is the text in the title field of html source code; Meta keywords is the meta data field in head field of html source code; text is the main text which contains the main content of web page. A html web page should be parsed to get these four field. Then this five construct like the table-1 below.

Table 1. Web page data structure.

classID	sports
url	http://sports.sina.com.cn/cba/2006-12-24/21172659112.shtml
title	I like music and pop star.
Meta keywords	Sports, Basketball, Sina, YiJianlian, CBA
text	Yi Jian Lian gets 23 points; he is a talent basketball player.

Advertisement is divided into 5 fields too: classID, title, keywords, URL, description. Where classID has the same meaning with web page's classID; Title is the title of advertisement; Keywords is related to the company or product; URL stands for the web site of company or the activity site of marketing; description gives a detail description to the product, company or related matter. An advertisement example is in the table-2 below.

Table 2. Advertisement data structure.

classID	literature
url	www.firstbook.org
title	Queen of the Scene
keywords	book, firstbook, literature
description	Grammy-winner Queen Latifah writes, a book for kids.

The structure of web page and advertisement are simple enough to abstract both of them. After getting the structured data source, next process will be the word segmentation and keyword extraction.

3.2 Word Segmentation and Keyword Extraction

In web page structure, text field holds the main meaning. According to vector space model[4], each web page can be seen as a document, text must be segmented as many weighted keywords which all together hold the semantics of a document. Chinese is a kind of language in which there is no separator character (in English,

space is separator) to separate the keywords in a sentence. So it is more complex to segment text of Chinese into keywords. We use Hailiang Technology's word segmentation software to implement Chinese word segmentation.

After segmentation of text, we will get a bundle of keywords, and each keyword is called a term in a document. Then we must make each term weighted to assure some term which approximate the semantics of document have larger weight, and vice versa. In traditional way, the weight of given term is calculated in equation-1, called tf-idf scheme[6] after all the documents are processed.

$$w_{i,j} = tf * idf = \frac{freq_{i,j}}{Max_l freq_{l,j}} * \log\left(\frac{N}{n_i}\right) \quad (1)$$

Where tf stands for term frequency; $freq_{i,j}$ is the raw frequency of term k_i in the document d_j; maxl $freq_{l,j}$ is the maximum value of the term frequency over all terms which are mentioned in the text of the document d_j. N is the total numbers of documents, n_i is the numbers of document which contains $term_i$.

But when we take web page as the document, the total number of web pages is unstable. At the same time, the text of a web page with a definite URL is also variable, so, td-idf scheme is not suitable here. We apply a single document keyword extraction method presented by [1], which could get weighted keywords based on single document using word co-occurrence statistical information. Because this method is English oriented, we tune it to fit Chinese by using Chinese word-segmentation tool mentioned above.

1. Word segmentation

In the text of web pages, not all the words are meaningful although they maybe appear frequently such as pronoun, conjunction and some frequently used verb. They are treated as stop words which is stored in a stop words list. Noun and verb is the units to segment in a sentence. After all the noun and verb are extracted, the construction of word co-occurrence matrix will be done.

2. Word co-occurrence matrix

In order to construct word co-occurrence matrix, the top N terms with high frequency are extracted by ranking of term frequency in descent order. Then we define the meaning of co-occurrence as: if $term_i$ and $term_j$ appear in a same unit which is predefined, then they co-occur once, and $freq_{i,j}$ should be added one. Here we make the unit as a sentence which is separated by Chinese punctuation such as '。', '！', '？'.

Our purpose to construct word co-occurrence matrix is to find the most representative terms for a web page. It is time consuming to calculate all possible term pairs. So we just compute the co-occurrence frequency between each terms and top N high frequency terms. It is obvious that the matrix is symmetrical, so $freq_{i,j}$ is equal to $freq_{j,i}$.

3. Calculation of $\chi'2$ value

Based on co-occurrence matrix, the χ'^2 value is calculated for each term. The higher of the χ'^2 value, the more important of the term to represent the semantics of document. χ'^2 value can be calculated in equation-2.

$$\chi'^2(\omega) = \sum_{g \in G} \frac{\left(freq(\omega, g) - n_w p_g\right)^2}{n_w p_g} \quad (2)$$

$$\chi'^2(\omega) = \chi^2(\omega) - \max_{g \in G}\left\{\frac{\left(freq(\omega, g) - n_w p_g\right)^2}{n_w p_g}\right\}$$

G as the set of top N frequent terms; g is a member of G; ω as the current term; p_g as (the sum of the total number of terms in sentences where g appears) divided by (the total number of terms in the document); n_w as the total number of terms in sentences where w appears. freq (ω, g) as the co-occurrence frequency between term pair ω and g. We take $\chi'^2(\omega)$ as the measure of term ω. The higher value of $\chi'^2(\omega)$, the bigger weight of ω should be assigned.

4. Normalization of $\chi'2(\omega)$

$\chi'^2(\omega)$ value's range is not between 0 and 1. It should be normalized to [0, 1] in equation-3. Then the weight will be assigned to each term as term-weight.

$$weight_i = \frac{\chi'^2(i)}{\max_l \chi'^2(l)} \quad (3)$$

Table-3 is an example of keyword extraction results from a web page, which is about the car. After calculating and Normalizing $\chi'^2(\omega)$ value, top 6 terms which have higher $\chi'^2(\omega)$ value are listed. It shows that a term which have high frequency doesn't necessarily have high $\chi'^2(\omega)$ value.

Table 3. Top 6 keywords after keyword extraction.

term	frequency	$\chi'^2(\omega)$	normalized $\chi'^2(\omega)$
automobile	17	519.22	1
company	11	347.54	0.67
market	8	305.52	0.59
trade	12	222.56	0.43
car	9	206.75	0.40
brand	4	174.92	0.34

5. Similarity

After the extraction of weight for each term in web page, we could extract the weight of terms in the description field of advertisement in the same way.

Both webpage and advertisement can be transformed to vector space model.

1) Keywords vector (KV)

Web page's meta keywords field can be represented as meta keywords vector, by inserting the keywords of title and keywords of high importance in the text field of page into meta keywords vector we will get the keywords vector (KV). Each keyword in KV is assigned in the way below:

Calculate the frequency of $term_i$ in meta keywords and title, then get the maximum frequency to normalize the weight in equation-4.

$$weight_i = \frac{freq_i}{\max_l freq_l} \quad (4)$$

Extracting 10 terms with high weight from text field after the normalization of $\chi'^2(\omega)$, then insert it into KV. If existing terms in KV contains the same one of 10 terms, then calculate the weight of the same term in equation-5 below.

$$weight_i = \max(weight_{vi}, weight_{ti}) \quad (5)$$

Where $weight_{vi}$ is the weight of $term_i$ in current KV, $weight_{ti}$ is the normalized weight $\chi'^2(\omega)$ of $term_i$ in text field.

Advertisement' KV can be made by combine the keywords fields, title field and top 5 terms of high normalized $\chi'^2(\omega)$ weight into KV, the calculation formulary of the weight to each term in vector is the same with web page's KV.

2) Content vector (CV)

As described above, we could evaluate the weight for terms in text field of web page as well as description field of advertisement. All of these weighted terms can be transformed to a content vector (CV) which represents the content of text field and description field. Because we extracted the top N terms from text field in to KV, they must be excluded from the CV. Certainly, the highest term-weight will not be 1 anymore, so the normalized $\chi'^2(\omega)$ weight should be calculated again in equation (6) below.

$$weight_i = \frac{weight_i}{\max_l weight_l} \quad (6)$$

$weight_i$ is the normalized χ'^2 value just mentioned above.

With KV and CV, we could define the similarity function between web page and advertisement as equation-7 below:

$$Sim(p_i, a_j) = \alpha * SimV(k_i, k_j)) + \beta * SimV(c_i, c_j) + \gamma * SimV(k_i, c_j) + \delta * SimV(c_i, k_j)$$
$$\alpha + \beta + \gamma + \delta = 1 \quad (7)$$
$$SimV(\vec{v}_i, \vec{v}_j) = \frac{\vec{v}_i \bullet \vec{v}_j}{|\vec{v}_i| \times |\vec{v}_j|}$$

Where p_i is web page i ; a_j is advertisement j; k_i is p_i's KV; c_i is p_i's CV; k_j is a_j's KV; c_j is a_j's CV. α, β, γ, δ is tuning constant, KV contains terms which represent the page or advertisement better than CV, so α, β, γ, δ should reflect the importance of different similarity function and the sum of α, β, γ and δ is 1. We define α=0.45, β=0.15, γ=0.2, δ=0.2; SimV( ) is the cosin function[2] of vector space model.

With the similarity function Sim( ), we could rank the advertisements for each web page to select the top n ads to post. In return, we could also rank the web pages for each advertisement to select the top n web pages to be post.

4 Evaluation and Future Work

To demonstrate our model, we designed the experiment in several steps.

1) We get 300 web pages from three famous Chinese website: Sohu, Sina, 163. We get 300 advertisements from Google's result lists which are supported by Google AdWords[7].

2) Every web page and advertisement is constructed in a XML format as source file.

3) Implement the word segmentation and construct CV index and KV index for each web page and advertisement.

4) We select 20 web page randomly form 300 source web pages, retrieve with their KV and CV, then find the most similar advertisements according to Sim() function by predefining a threshold θ. then calculate the Precision and Recall. And in

return, we select 20 advertisements randomly from 300 source advertisements. We make the same retrieval process to get the Precision and Recall. Here, Precision is the number of relevant items in retrieval result list divide by the result list's size. Recall is the number of relevant items in retrieval result list divided by all relevant items in source database. The experiment result is shown in Fig. 1 and Fig. 2.

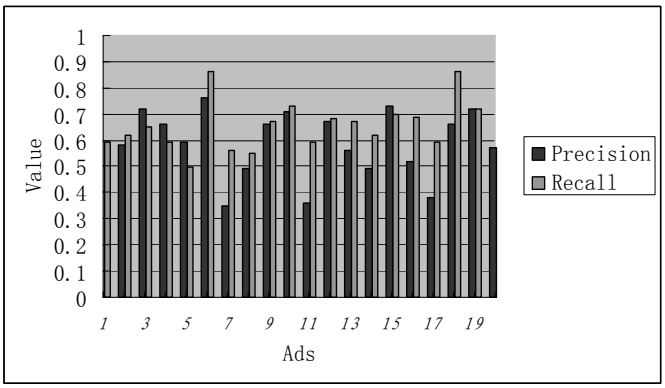

Fig. 1. Precision and Recall of 20 advertisements as the query to retrieve web pages.

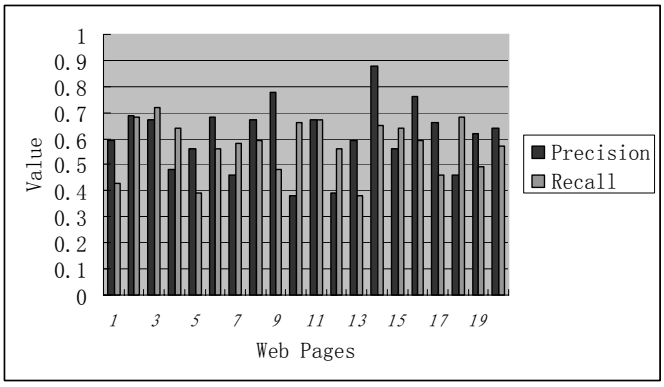

Fig. 2. Precision and Recall of 20 web pages as the query to retrieve advertisements.

In our experiment, the first 5 advertisements and web pages is very relevant, which is suitable to the web advertisement case because there is a limitation for the number of advertisements in one web page.

Our model pays more attention to match the semantics of web page and advertisement by calculating the weight of term in text or description field. But whether customer would like click or how many customers will click greatly rely on the interest of customer and the popularity of that web page. It is reasonable to

assume that if page and advertisement have same semantics, customer who is interested in page has great probability to click the advertisement. In the other hand, the frequency of web page browsed and the percentage of target customers who have interest on advertisement is another very important factor. In our future work we will pays more attention on browsed frequency and target customers distribution to further the study of web advertisement.

Acknowledgement

This research is supported by the AOE Important Project of Philosophy and Social Science. The project number is 05JZD00024. It is also supported by NSFC under Grant 70473068.

References

1. Matsuo Y, Ishizuka M. “Keyword extraction from a single document using word co-occurrence statistical information”. *Int'l Journal on Artificial Intelligence Tools,* 2004, 13 (1):157~169

2. Ricardo Baeza-Yates, *et al. mordern information retrieval*. 1999, ACM press .

3. Ng V, Kwan-Ho Mok. An intelligent agent for Web advertisements. Cooperative Database Systems for Advanced Applications, 23-24 April 2001 Pages:102~109

4. Amiri A, Menon S. Scheduling web banner advertisements with multiple display frequencies Systems, Man and Cybernetics, Part A, IEEE Transactions on Volume 36, Issue 2, March 2006 Pages:245~251

5. Thawani A, Gopalan S. “Event driven semantics based ad selection”. *Multimedia and Expo*, 2004. ICME '04,27-30 June 2004 Pages:1875~1878

6. Salton, G., Buckley, C., “Term weighting approaches in automatic text retrieval.” *Information Processing and Management*, 1988.24(5):513-523.

7. Google AdWards, http://www.google.com/intl/zh-CN/ads/. access date: December 21, 2006