Personalized Internet Advertisement Recommendation Service Based on Keyword Similarity

[1] Dipika Deshmukh, [2]Dr. D.R. Ingle
[1] Student1, [2] Professor
[1, 2] Department of Computer Engineering, Bharati Vidyapeeth College of Engineering, Kharghar, Navi , Mumbai
[1] dipika.deshmukh6@gmail.com, [2] dringleus@gmail.com

***Abstract*— online advertising is major source of revenue for today's online world. With rapid development of E-commerce audiences put higher requirements on personalized Internet advertisements.This study focus on developing a famework for personalized Internet advertisement recommendation service. Personalized advertisement aims to most suitable advertisement for anonymous users on website. All advertisements are categorized by commercial categories provided by yahoo. Then keywords are extracted from each advertisement and the most correlated keywords to each category are identified through Term Frequency and Inverted Domain Frequency (TF-IDF) analysis. Thus, the ontology of the advertisement is built. Normalized Google Distance (NGD) relationships between keywords are computed to derive the characteristic vector of each advertisement. Besides, based on user's responses to some advertisements, the user profile, which describes a user's preferences for advertisements, is established through logistic regression. Finally, for a new advertisement, a recommendation value is computed by using the characteristic vector and the user profile. The value is used to determine whether this advertisement should be recommended to the user or not. A prototype website for verifying the proposed schemes was developed.**

***Keywords*— Internet advertisement, Recommendation, Term Frequency and Inverted Domain Frequency, Logistic Regression Normalized Google Distance.**

I. INTRODUCTION

Personalization of online advertising is a great challenge while the market is moving and adapting to the realities of the Internet. Many existing approaches to advertisement recommendation are based on demographic targeting or on information gained directly from the user. Internet is constantly innovated and matured, user and usage information can be recorded and preserved. To recommend an advertisement to a user, first analyze the contents of the advertisement and categorize it. One can collect and analyze this information to find a particular user's interests and hobbies. This helps the user find the information that is most relevant to him/her and therefore increases the effectiveness of the advertisement. The interactive nature of the Internet, the behavior of the user and the products that he/she selects can be used as a basis for the recommendation. This paper presents an integrated approach to analyzing advertisements and recommending them according to user's interests. It is briefly described as follows.

Keyword identification is the first step to identify keywords. The CKIP Chinese word segmentation system [1] can separate paragraphs into the smallest units of terms and individual phrases. It is used to extract the necessary keywords needed for categorization. Term Frequency and Inverse Document Frequency (TF-IDF) values [2] are used to determine what category the keywords are correlated to. Thus, the advertisement ontology can be established through categorization, keyword extraction and correlation determination,. Then, by using the NGD (Normalized Google Distance) values [3], relationships between different keywords are identified, which can be used to describe the characteristic vector of an advertisement. Based on user's responses to some advertisements, user profile, which describes a user's preferences for advertisements, is established through logistic regression. Finally, by using the characteristic vector of an advertisement and the user profile, a recommendation reference value for this advertisement is computed. Thus,this advertisement is recommended to the user or not is determined according to this reference value.

In this paper related work is describes in Section II. The theory of the operations is given in the Section III. In Section IV is Conclusion.

II. RELATED WORK

A. Personalized Advertisement Recommendation

In the world of e-commerce, users are interested to get personalized content on sites they often visit. Advertisement on sites is a way of generating the revenue for publisher and supports availability of free contents on internet. Internet advertising generally performs advertisement recommendation by applying pre-specified rules or conditions. For example, if user pays some fee to a web server, for advertisements, this advertisement will pop up when a user visits this web server. However, the user may have no interests in the recommended advertisements. That is, the user will not click the advertisements or even to read them, and hence the effectiveness of the advertising will be poor. Therefore, to increase the effectiveness of personalized advertisement, we recommends the advertisements meeting users' preferences, is a important issue for internet advertising. Another approaches to this issue recommend related commercial advertisements to a user according to his or her purchase records or even browsing records [4][5][6]. However, such an approach can only be applied to specific kind of advertisements recommendation. To recommend a general advertisement, a more flexible and automatic methodology should be developed. To solve this problem, this study proposed a framework for performing personalized internet advertisement recommendation based on keyword similarity and the advertisements that the user has interest can be determined.

B. Ontology

Ontology is the core technology originates from philology. Ontology is defined generally as a formal, explicit specification of a shared conceptualization. In other words ontology is a shared and common understanding of some domain that can be communicated across people and computers. [7] Ontology can be shared and reused among different applications. Ontology can be regarded as a dictionary that consists of words and relations. The website directory service was used as ontology to identify user’s browsing behaviors on Internet so as to discover user preferences for providing effective personalized services.

C. Logistic Regression

Regression is a useful technique to explain relationships between variables and to predict a quantitative response. Linear regression is commonly used for predictive analysis. Most regression examples use continuous dependent variables. However, in contrast, logistic regression is appropriate regression analysis to conduct when the dependent variable is binary or dichotomous. Application of logistic regression is used to build indicator for prediction [9]. In this study, since the final recommendation is dichotomous, i.e., yes or no. The details will be elaborated in Section III.

D. NGD (Normalized Google Distance)

NGD is a semantic similarity measure between two terms using the number of hits returned by Google search engine as searching the terms. Words with the same or similar meanings tend to be close in terms of Google distance, while words with dissimilar meanings tend to be farther apart. In this study, the similarity of keywords is identified by utilizing the NGD values so as to determine the characteristic vector of an advertisement. Further details will be provided in Section III

III. THEORY OF OPERATIONS

A. Architecture of the Proposed Recommendation Service

This paper presents an integrated personalized recommendation service based on keywords extracted from real advertisements. The objective is to display and recommend advertisement/commercial to draw user’s attention and interests. The architecture and flow of the system is shown in Figure 1:

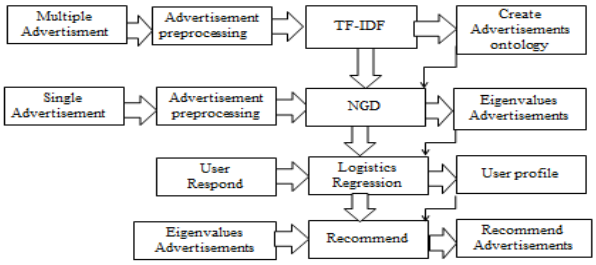

Multiple Advertisment
Advertisement preprocessing
TF-IDF
Create Advertisements ontology
Single Advertisement
Advertisement preprocessing
NGD
Eigenvalues Advertisements
User Respond
Logistics Regression
User profile
Eigenvalues Advertisements
Recommend
Recommend Advertisements

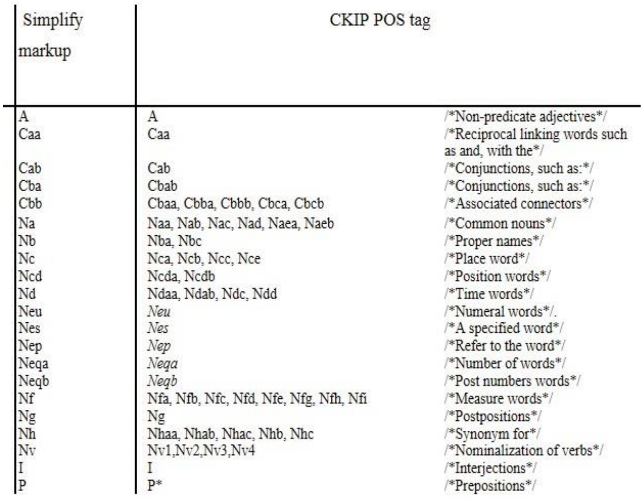
Fig.1. System Architecture

B. Implementation Details

The proposed personalized advertisement recommendation service consists of four main steps:

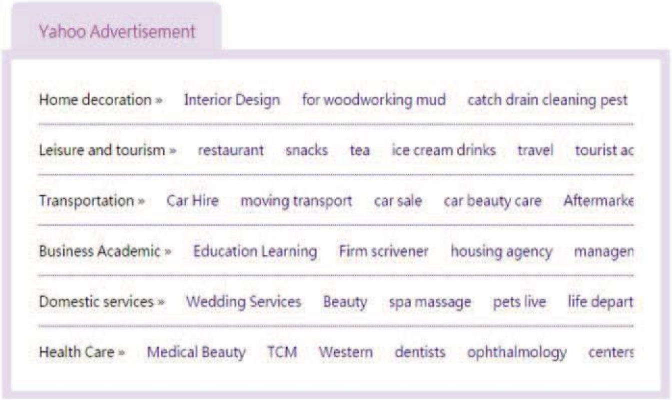
Step 1: Building Advertisement Ontology

At first, we need to find the keywords in the content of an advertisement. The CKIP Chinese word segmentation system [1] is used to separate all terms. All terms have their characteristics as shown in Figure 2. For nouns, there is Na (normal nouns), Nb (special nouns), Nc (place nouns), and Nd (time nouns). Using YAHOO [11] category classifications, as shown in Figure 3, some advertisement samples are selected from each category. These advertisement samples are passed through CKIP Chinese word segmentation system to obtain all Na (normal nouns), Nb (special nouns), Nc (place nouns), and Nd (time nouns). Then, TF-IDF (term frequency–inverse document frequency) is used to filter out the keywords for further analysis.



Simplify markup	CKIP POS tag	
A	A	/*Non-predicate adjectives*/
Caa	Caa	/*Reciprocal linking words such as and, with the*/
Cab	Cab	/*Conjunctions, such as:*/
Cba	Cbab	/*Conjunctions, such as:*/
Cbb	Cbaa, Cbba, Cbbb, Cbca, Cbcb	/*Associated connectors*/
Na	Naa, Nab, Nac, Nad, Naea, Naeb	/*Common nouns*/
Nb	Nba, Nbc	/*Proper names*/
Nc	Nca, Ncb, Ncc, Nce	/*Place word*/
Ncd	Ncda, Ncdb	/*Position words*/
Nd	Ndaa, Ndab, Ndc, Ndd	/*Time words*/
Neu	*Neu*	/*Numeral words*/.
Nes	*Nes*	/*A specified word*/
Nep	*Nep*	/*Refer to the word*/
Neqa	*Neqa*	/*Number of words*/
Neqb	*Neqb*	/*Post numbers words*/
Nf	Nfa, Nfb, Nfc, Nfd, Nfe, Nfg, Nfh, Nfi	/*Measure words*/
Ng	Ng	/*Postpositions*/
Nh	Nhaa, Nhab, Nhac, Nhb, Nhc	/*Synonym for*/
Nv	Nv1,Nv2,Nv3,Nv4	/*Nominalization of verbs*/
I	I	/*Interjections*/
P	P*	/*Prepositions*/

Fig.2.Term classification



Yahoo Advertisement

Home decoration » Interior Design for woodworking mud catch drain cleaning pest

Leisure and tourism » restaurant snacks tea ice cream drinks travel tourist ac

Transportation » Car Hire moving transport car sale car beauty care Aftermarke

Business Academic » Education Learning Firm scrivener housing agency managen

Domestic services » Wedding Services Beauty spa massage pets live life depart

Health Care » Medical Beauty TCM Western dentists ophthalmology centers

Fig.3.Advertisement classification

Term Frequency (TF) and Inverse Document Frequency(IDF) are common techniques for data indexing searches and data mining. When a term shows up frequently (high TF) in a document and seldom in others (high IDF), the term is a

good candidate as keyword differentiator. The TermFrequency (TF) of a term ti is formulated as.

$$\mathrm{TF}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad (1)$$

Where ni,j is the number of occurrences of ti in document dj,the denominator is total occurrences of all terms in document dj. IDF (inverse document frequency) is a popularity measurement of a term. It is the ratio of the total number of documents to the number of documents containing that term. It is formulated as.

$$\mathrm{IDF}_i = \log \frac{|D|}{1+|\{j: t_i \in d_j\}|} \qquad (2)$$

where |D| is the number of documents in the set, and |{j:ti_dj}| is the number of documents containing term ti (i.e., the number of documents with ni,j≠0). Note that if the term is not in the document set, |{j: ti_dj}|=0. So, we have the formulation.

$$\text{TF-IDF}_{i,j}=\mathrm{TF}_{i,j} * \mathrm{IDF}_i \qquad (3)$$

When a term has high document frequency, it will have small IDF value. Thus, even if this term has high TF value in a document, it still has small TF-IDF value. TF-IDF will filter out common words and leave important keywords .After all keywords in each category are identified, then ontology for the advertisement world can be established. An example is shown in Figure 4 Example of Advertisement Ontology.

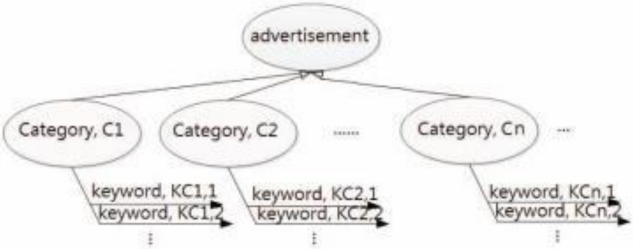


Fig.4.Example of Advertisement Ontology

Step 2: Advertisement Characteristic Vector

After Step 1, assume that the ontology has n different categories,denoted as C1, C2, ⋯, Cn, and each category has m correlated keywords, denoted as KCi,j, 1≤i≤n, 1≤j≤m.For an advertisement A, after keyword extraction, assume that there are s keywords (KAh, 1≤h≤s), the similarity of the keywords KAh and the keywords KCi,j in each category C1, C2, ⋯, Cn will be computed based on NGD (Normalized Google Distance) [3]. The definition is given in the following formula.

$$NGD(w_i, w_j) = \frac{\max\{\log_e(f(w_i)),\log_e(f(w_j))\}-\log_e(f(w_i,w_j))}{\log_e N-\min\{\log_e(f(w_i)),\log_e(f(w_j))\}},$$

where f(wi) and f(wj) are the number of search results of word wi and word wj, respectively; f(wi, wj) is the number of search results of word wi and wj; N is the total number of pages in the Google search engine.

Because NDG value represents the similarity between terms, we can derive the similarity of an advertisement A and category ci by computing the average NGD of all keywords in advertisement A (KAh, 1≤h≤s) and all keywords in category Ci as follows:

$$Sim(A, C_i) = \frac{\sum_{h=1}^{s}\sum_{j=1}^{m} NGD(KA_h, KC_{i,j})}{s*m} \qquad (4)$$

The above values can then be used to define the characteristic value of an advertisement A:

Z= (Sim(A, C1), Sim(A, C2), …, Sim(A, Cn)) (5)

The main objective of this research is to design an integrated advertisement recommendation service based on user's preference. Formula (5) defines a characteristic value of an advertisement for such design.

Step 3: User Profile Establishment

Step 1 defines the advertisement ontology with all related keywords. Step 2 has advertisement characteristic vector defined. The characteristic vector represents the similarity with the advertisement category. In the following step, a user profile is established as follows: First of all, take some sample advertisements and use Step 2 to generate characteristic vectors for all advertisements. Next, the user preferences on these advertisements are collected through some interview process (online or telephone survey). In this system, an on-line survey is used. After registration, a user will pick favorite commercials from a list of advertisements. Through these feedbacks, the system can derive the user profile through logistic regression. The user profile is a set of weights to describe the user preferences on these advertisements. Here at, let us deviate for a while to introduce the Logistic regression. Logistic regression describes the relationship between a binary dependent variable and one or more predictors. In practice, the binary dependent variable can be success or failure, pass or no pass, meet or no meet. With Logistic Regression, the relationship of the dependent variable f(Z) and n independent variables Z1, Z2, …, Zn is defined as follows:

f(Z)=β0+β1Z1+β2Z2+……+βn

The above formula can be used to identify the user profile, which is the set of user's preference weights on advertisement categories. For a user profile U=(β0, β1, …,βn), and advertisement d with characteristic vector Zd=(Zd1,

Zd2, …, Zdn), the logistic function is f(Zd)=β0+β1Zd1+β2Zd2+……+βnZdn. (6)

Through SPSS software, a regression analysis on Formula (6) is done based on user's feedback data to obtain the user profile U=(β0, β1, …, βn), which becomes the feature set of the user.

Step 4: Targeted Advertisement Recommendation

For a new advertisement A, assume that after word segmentation process, s keywords in A, denoted as KAh, 1≤h ≤s, are derived. Let KCi,j represents the ith attribute (i.e.,correlated keyword) in the jth category. With Step 2, through NGD analysis on the keywords KAh and KCi,j, the advertisement characteristic vector Z=(Sim(A, C1), Sim(A,C2), …, Sim(A, Cn)) is obtained. Besides, according to Step3, the user profile U=(β0, β1, …, βn) is derived. Thus, the recommendation value f(Z)=β0+β1Z1+β2Z2+……+βnZn can then be computed. When f(Z)≥0, it implies that the user is interested at the advertisement A, and hence A is recommended to the user. In contrast, when f(Z)<0, the advertisement A is of no interest to the user.

CONCLUSIONS

Internet advertisement is a popular and effective tool for on-line marketing. We proposes a practical method to provide targeted and customized advertising services. A better marketing efficiency is reached through precise matching of advertisement characteristics and a user's profile. This can improve user's shopping experience of online marketing and it also increase customer's spending power to achieve overall economic growth. We also include a self-improving and correction system by taking feedbacks of users.

REFERENCES

[1] M. Shatnawi and N. Mohamed, "Statistical Techniques for Online Personalized Advertising: A survey," 27th Annual ACM Symposium onApplied Computing, pp. 680–687, New York, 2012.

[2] R.-D. Ou, "Using Ontology and Universal User Profile for Personalized Recommendation", Master Thesis (in Chinese), Department of Information Management, Chaoyang University of Technology, 2006.

[3] W. Y. Ma and K. J. Chen, "Design of CKIP Chinese word segmentation system," Chinese and Oriental Languages Information Processing Society, Vol. 14. No. 3. pp. 235-249.

[4] Y. Jang , T. Lee, K. Kim, and W. Lee, "Keyword Management System based on Ontology for Contextual Advertising," 6th International Conference on Advanced Language Processing and Web Information Technology, pp. 440-445, 2007

[5] R. L. Cilibrasi and P. M. Vitanyi, "The Google similarity distance,"IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No.3, pp. 370-383, 2007.

[6] J. S. Hsu and T. C. Hung, "The Use of Ontology in Case-based Reasoning System for Itinerary Recommendation," Journal of Information Management (in Chinese), Vol. 9, pp.31-61, 2004

[7] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information processing & management, Vol. 24, pp.513-523, 1988.

[8] YAHOO Keyword categories, Available on line (Retrived: May 2015):https://tw.ysm.emarketing.yahoo.com/soeasy/

[9] H. C. Huang, M. S. Lin, and H. H. Chen, "Analysis of Intention in Dialogues Using Category Trees and Its Application to Advertisement Recommendation," 3rd International Joint Conference on Natural Language Processing, pp. 625-630, 2008.