PROCEEDINGS OF SCIENCE

Web Data Source Selection for Tourist Culture integration

Song Deng[1][2]
School of Software & Communication Engineering
Jiangxi University of Finance and Economics
Nanchang 330013, China
E-mail: daonicool@sina.com

In order to improve the efficiency of cultural information integration in the field of tourism, we propose a new strategy of data source selection based on the association between celebrities and mark words. Firstly, we acquire the sampling data based on character ontology, collect information in relation to the content of celebrity based on related themes, recognize the cultural theme of sentences based on the user's feedback, and finally extend character based on the project evaluation techniques. After the cultural summary is built, a data source scoring approach has been proposed based on the culture content gain of "single celebrity " and "synthetic character". We use domain datasets to do experiments. The results show that the accuracy of our method is higher and our method could provide effective support for characters and information integration of attractions.

CENet2015
12-13 September 2015
Shanghai, China

[1]Speaker

[2]Our research was supported by the Natural Science Foundation of China under grant No: 61462037 and the Natural Science Foundation of Jiangxi under grant No: 20142BAB217014.

PoS(CENet2015)020

1. Introduction

Currently, the existing information integration systems of tourism process tourist information as general information [1], they don't mine the cultural information of attractions, but this information is valuable. If a celebrity has travel, living and missionary experience in terms of an attraction, it can often greatly enhance the culture connotation of this attraction; thus it is necessary to establish a Web data integration system based on the culture mining to further use these resources. Data integration primarily include several stages, such as Web data source selection, entity data fusion and association mining. If we integrate information of all data sources to get the culture-related information in the field of tourism, it apparently does not work because there are a lot of Web data sources available in each field. The approach above makes data integration too costly and difficult to ensure the quality of data. For these reasons, Web data source selection is a key issue of tourist culture integration[2]. People usually want to get the results of culture integration by integrating a very small amount of Web data source, in this sense, the data source selection techniques are important.

The culture and tourist value of attractions need the following two requirements: 1) the attractions have not much culture content and need to be further enriched; 2) the cultural information of attractions related to someone important is less and need to be further enriched. The cultural themes of attractions primarily cover 9 aspects: travel, religion, literature, education, politics, life, economics, architecture and scientific research.

In order to effectively select data sources, we start by organizing the humanistic content of each data source in accordance with celebrities. To further refine the humanistic information corresponding to each celebrity, we select the mark words (emotional words, words in special symbols and the last word of attraction name) to manifest the cultural content and finally select data sources based on a gain model. Experimental results show that the proposed tourist cultural information integration-oriented Web data source selection has high accuracy and plays a greater role in improving the value of entity information integration.

2. Related Work

The Web data sources are growing rapidly with the popularity of Internet, in order to help users to effectively use these data sources, Web data integrated technology has attracted more and more attention. As the key technology of Web data integration, the data source selection has always been the research hot spot[2]. Currency, the research include following aspects:

1) Data Source Selection based on Data Quality and Money Cost

Balakrishnan used "recommended map" based on similarity of tuples to build summaries of data source and select a data source[3] . Dong balanced the quality and money cost of data, selected top-N data sources based on the theory of marginal idea[4,5] . As to the scientific issues of quality dimension selection in the process of data source selection, Deng proposed a selected idea based on user feedback[6], which featured high accuracy and and small computation. In addition, the quality of data source will change with time. Thus, Rekatsinas studied the problem of dynamic data source selection[7].

2) Data Source Selection Based on Semantic

Fan built the sampling summary of data source based on keywords-fields property related[8]. The accuracy of this approach is higher for the strong structured fields while the accuracy is lower for the weak structured fields. Wan studied the problem of mixed-type keywords-oriented data source selection [9] , built a hierarchical summary of data source based on subject words association and the attributes histogram. The above summary contained rich subject and the semantic information. Wang has made some achievements in structured data source selection based on ontology mapping data source form attributes[10].

3) Data Source Selection based on the Specific Needs

There are so many factors that would affect the accuracy of data source selection, such as sampling quality, document sorting quality. Aiming at those problems, Markov alleviated the uncertainty of the above data source selected method based on a combination of evaluating

strategies[11]. In practice, different users may have different integrated requirements; therefore, evidences should be selected according to the characteristics of user's demand. For the above problem, Hong proposed a data source selection strategy based on the diversity results [12].

Nevertheless, the data source selection based on the quality and cost did not consider the query relevance; the data source selection based on keywords semantic only considered the association between keywords. In this sense, none of them can server the purpose of tourism culture integration. The key of culture integration is mining stories about celebrity and attractions, thus we build data source summary based on celebrity and remark words; in addition, we designed a data source selection policy based on a single celebrity and "synthetic celebrity", which could support the retrieval of humanities information effectively.

3. Construction of Web Data Source Summary

For Web data source selection, we need to build a summary of data source that takes up small storage space and retention of critical information for each Web data source. For this purpose, we use the sampling method to build the summary of the data source based on character-characteristics related, as shown in Fig. 1. As the celebrities are the core value of tourism culture, we use the names as sample words to obtain the summary information.

In order to minimize the storage space of the summary and enhance the represent activeness of the sample celebrity, firstly, we need to build a sample celebrity ontology. It can be done by the following steps: 1) use the crawler to get the Web page information of famous attractions from the Baidu.com encyclopedia; 2) use the segmentation software(ICTCLAS 2013) to get names[2]; 3) according to the cultural theme fields, the names are divided into three collections in accordance with the frequency of occurrence and according to four properties, namely , dynasty, achievements, official or not, position in each field randomly tagging F celebrities from the three collections.

In order to expediently use a wide cultural information of multiple data sources, the content features of cultural theme in the summary of data source need to be further refined. The content remark words of cultural theme mainly include: 1) phrases in special symbols; 2) emotional words; and 3) the last word of attraction name. For example, double quotation marks in the data source contained in the title mark "Watching the Lu Mountain Falls" or "Fishing Islands", are obviously corresponding to key elements of the theme; such emotional words as "Widely known " and "Masterwork" made content of poem clearly. To further compress the storage space of summary, only the first two words of special symbol were retained in the feature words. The last word of attractions' name was preserved, such as "mountain" and "City", which can improve accuracy of integration of "attractions" + "name", for example "Mountain Lu Li Bai".

Due to limit of the information of sampling summary, some celebrities' information cannot be reflected directly in the summary, so we speculate that related celebrities information based on similar celebrities extend thestrategy, for instance, in the summary missing tourist culture information of Tang poet "Lu Lun", we can discriminate the most similar characters in the summary according to the similar properties (dynasty, achieved field, official or not and position in each field) between him and other poets.

To build the summary above, firstly, we need to identify contents associated with celebrities in data source, then distinguish cultural themes relevant to celebrities, finally realize the celebrity extended strategy and enrich the summary information.

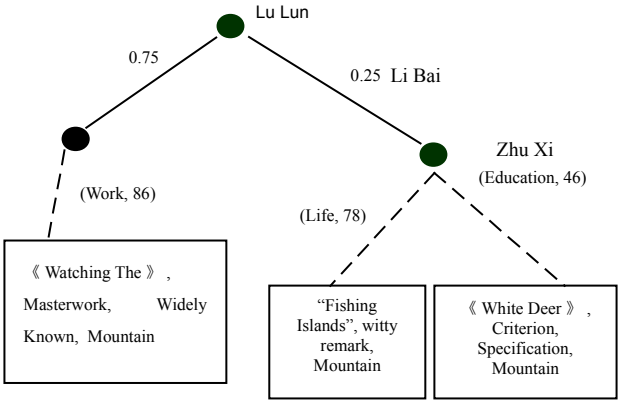

Figure 1: Associated Summary Based on Celebrities and Remark Words

3.1 Measurement of Culture Content-Length corresponding to Celebrities

To obtain the celebrity-related sentences, finding the sentence with the names of celebrities is not enough because some front and rear sentences are inseparable from the original sentences, though they don't contain names. If we need to find all the celebrity-related content, we should consider two following aspects: 1) whether two sentences could be associated by conjunctions, pronouns, and names; 2) whether the keywords involved in two sentences are similar (because if keywords of before and after sentences are the same, then two sentences could talk about the same person). The keywords can be scoreded according to Equation (3.1).

$$\text{TD}*IDF(w_i, E) = \frac{freq(w_i, E)}{NumWords(E)} * \log_2 \frac{N_C}{count(w_i)} \tag{3.1}$$

where E is the celebrity name, N_C is the number of documents in the collection C for a travel document, $count(w_i)$ is the number of containing words and w_i in the C, $freq(w_i, E)$ is the number of w_i of sentences contain E in the C, $NumWords(E)$ is the number of words in the collection of sentences contain E.

On the basis of these ideas, the method of content-length measurement of corresponding celebrity is designed. Firstly, identify documents that contain celebrity E in the data source S, and then get paragraphs that contain E in the specific document, compute cultural content-length in the specific paragraph, finally obtain the total content-length of celebrity E in the data source through accumulating and compute the content-length of celebrity in each paragraph according to Algorithm 1.

Algorithm 1: measure the content-length of celebrities in the paragraph

Input: paragraph D_E with E, entity name E, the content of the document R;
Output: E content-length of related theme;
$Result = \emptyset$;
for each $Sentence \in D_E$
if $\exists$ $Sentence_{begin}$ contain E
$Result = Result \cup J_{begin}$;
break;
else
return;
while(1)
if J_{next} (the next sentence of J_{begin}) contain pronouns, E, conjunctions
$Result = Result \cup J_{next}$;
$J_{next} = (J_{next})_{next}$;

continue;

if according to Eq.(1) gets Top-N keywords groups C_1, C_2 of J_{next} and J_{prior}, $term1_i$ and $term2_i$ are synonymous, $term1_i \in C_1$, $term2_i \in C_2$.

$Result = Result \cup (J_{next})_{next}$;

$J_{next} = (J_{next})_{next}$;

continue;

break;

while(1)

if according to Eq.(1) gets Top-N keywords group K_1, K_2 of J_{begin} and J_{prior} (the prior sentence of J_{begin}), $term1_i$ and $term1_i$ are synonymous, $term1_i \in K_1$, $term2_i \in K_2$.

$Result = J_{prior} \cup Result$;

$J_{prior} = (J_{prior})_{prior}$;

continue;

break;

return *Length*(Result)

3.2 Cultural Themes Recognition of Sentence in the field of Tourism

The sentences of the tourism document that describe celebrities and the related cultural themes of attractions are tending to be short and refined. In order to get higher recognition of the theme, we have designed a method for distinguishing the cultural theme of sentence based on user feedback.

In the process of cultural theme recognition, firstly, artificially mark Num sentences which cultural theme is C_i. Users need to select and sort the characteristic words which may present the cultural theme in each sentence based on their own experience and knowledge. One characteristic word *T* was selected frequently into the collection C_i by the user and ranked near the top; at meanwhile, this word was selected seldom in other collections of characteristic words and ranked lower, indicating that the accuracy of selection would be higher.

When keyword identification is done, this paper uses the following equation to calculate the cultural theme score corresponding to each word:

$$Score(T,c) = recommend(T,c) - reject(T,\bar{c}) \tag{3.2}$$

where $recommend(T,c)$ is the theme score of *T* corresponding to *c*, $reject(T,\bar{c})$ is the theme score of *T* corresponding to non-*c*.

$$recommend(T,c) = \sum_{i=1}^{k} \left(\frac{(Num(T,c,S_i) + 1 - R(T,c,S_i))}{(Num(T,c,S_i))} \right) \tag{3.3}$$

where *k* is the number of sentences of theme, *c* contains characteristic word *T*, $Num(T,c,S_i)$ is the total number of characteristic words in the sentence S_i of theme *c* contains characteristic word *T*, $R(T,c,S_i)$ is the rank of characteristic word *T* in the S_i.

$$reject(T,\bar{c}) = \sum_{i=1}^{k} \left(\frac{(Num(T,\bar{c},S_i) + 1 - R(T,\bar{c},S_i))}{(Num(T,\bar{c},S_i))} \right) \tag{3.4}$$

where *l* is the number of sentences of theme, non-*c* contains characteristic word *T*, $Num(T,\bar{c},S_i)$ is the total number of characteristic words in the sentence S_i of theme, non-*c* contains characteristic word *T*, and $R(T,\bar{c},S_i)$ is the rank of characteristic word *T* in the S_i.

Match characteristic words in each sentence; according to Equation (3.2), calculate the score of characteristic words in every single sentence of each theme. Thus the theme which have the highest score is the cultural theme of corresponding sentence.

During the feedback training, data are limited and there will be some characteristic words difficult to get. For this problem, we use the following strategies: 1) extend based on synonym; 2) extend based on the exist characteristic words.

As to the idea of extension based on the existing characteristic words, if there is more than one word of discriminating words existing in the collection of the existing characteristic words,

it would be more likely the extended words amd its score of theme can be evaluated according to the score of theme of characteristic words corresponding to each word in the extended words. With such strategy, we calculate the score of theme of the terms by using the following equation:

$$Score_{extent}(J,c)=\frac{1}{J}\sum_{p=1}^{J}(\min_{Y\supseteq J_p, Y\in U}(Score_{one}(Y,J_p,c))) \quad (3.5)$$

where $Score_{extent}(J,c)$ is the score of the extended words, J corresponding to theme *c*, $|J|$ is the number of single word in the extended words J, J_p is the p[th] word of J, U is the collection of the marked characteristic words which contains J_p, Y is the characteristic words of the U which contain J_p, $(Score_{one}(Y,J_p,c))$ is the score of theme *c* corresponding to the characteristic words Y which contain J_p, using Equation (3.6) to score.

$$Score_{one}(Y,J_p,c)=\frac{1}{|Y|}Score(Y,c) \quad (3.6)$$

4. Strategy of Web Data Source Selection

When the summaries of data sources have been built, we can select Top-N data sources based on content gain. For the requirements of culture integration, we should calculate the cultural content gain which each data source provide according to Equation (4.1). Firstly, we select the data source which have the maximum information gain, and then select the next data source order by information gain in the selected collection of data sources.

$$\Delta(S_i,p)=\sum_{j=1}^{|O|}(L(S_i,p,z_j))\frac{|T(S_i,p,z_j)-T_d(p,z_j)|}{U} \quad (4.1)$$

where $\Delta(S_i,p)$ is the cultural content gain of celebrity p provided by the data source S_i, O is the theme of containing the last word of the attraction's name corresponding to the collection of characteristic words and related to the celebrity p in the data source S_i (for example, if the attraction is "Lushan", we need to find the cultural theme of containing "shan" and corresponding to celebrity in the data source S_i), z_j is the specific theme, $L(S_i,p,z_j)$ is the length of content of the theme z_j relating to the celebrity p in the data source S_i, $T(S_i,p,z_j)$ is the collection of characteristic words of the theme z_j relating to the celebrity p in the data source S_i, $T_d(p,z_j)$ is the collection of characteristic words of the theme z_j relating to the celebrity p in all the selected data sources, $U=\min(|T(S_i,p,z_j)|,V)$, which V is the maximum number of characteristic words of theme which cumulative scores corresponding to the reserved celebrity near the top in summary.

5. Experimental Results and Analysis

There isn't a collection of standard test in the field of Web data source selection, therefore, we selected about 40 real Web data sources, such as Baidu Encyclopedia, 360 Encyclopedia, Tencent tourism, and Sohu tourism, Sina tourism, NetEase tourism, and China tourist Web, China attractions Web, LvMaMa tourism, Xinxin tourism, Tuniu.com, tourist yellow pages, JiuYou Web, tourist Web of regions, etc. To build a summary of data source need to identify the related sentences of celebrity firstly, we then computed the score of theme for each word according to Equation (2).

We used Algorithm 1get all the related sentences of one celebrity in a document, Fig. 2 shows the relationship between the number of keywords and accuracy in Algorithm 1. Using conjunctions, pronouns and names can easily find the sentences before and after one celebrity;

in this sense, the key is to find the sentences before and after based on keywords. Two adjacent sentences may contain more than one keyword, we only need to find the same keyword to determine their relationship. Because if two adjacent sentences contain the same keyword, the probability that they are associated with the same celebrity is high. In practice, we could set a matching window and only match Top-N keywords of each sentence to find similar keywords.

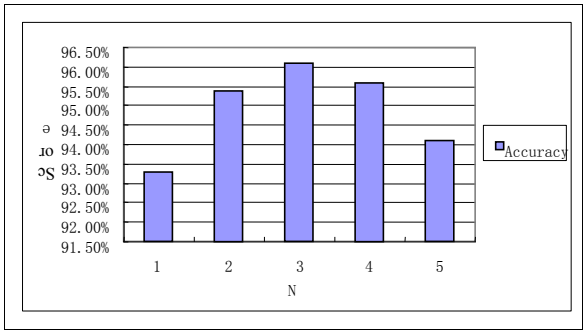

Figure 2: Relationship between *N* and Accuracy

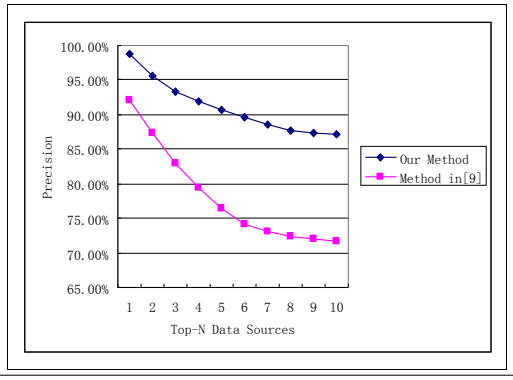

Figure 3: Comparison of Accuracy of Data Soure Selection Methods for Cultural Information Integration

As can be seen from Fig. 2, the accuracy increases when *N* is from one to three, and the accuracy exceeds 96% when *N* is three; but if *N*'s value continues to expand, the accuracy will decline because when *N* is too small, the matched range of keywords of two collections corresponding to the adjacent two sentences is too small; when *N* is larger, the matched range of keywords of two collections is larger, and the matched keywords can not represent the main topic of the two sentences; therefore, we just need set *N*=3.

Due to the entity-oriented integration of data source selection, the criteria of evaluation based on the relevance cannot be taken. In order to effectively evaluate the accuracy of data source selection, we calculated the accuracy of data source selection on the basis of $C1/C2$, where $C1$ is the number of non-repeating words of the selected Top-N data source, $C2$ is the number of non-repeating words of the real Top-N data sources. We chose the average accuracy of data source selection corresponding to 1300 celebrities and 300 attractions culture integrations as test results. To evaluate the effectiveness of method of this paper proposed, we chose the hybrid method [9] as a comparison method featuring sound performance and considering the topic and semantic information. In the process of experiment, when the celebrity ontology was built, each celebrity collection was divided into three parts by frequency and F's value was 30 of each part. Fig. 3 shows the results of two methods for the culture integration.

As can be seen from Fig. 3, the approach of this paper proposed have a greater advantage than the method of word frequency. When the Top-1 data source is selected, the accuracy exceeds 98%, over 6% than the method of word frequency; when the number of selected data source is increased, the advantages get to further expansion and select the Top-10 data sources, the accuracy exceeds 87%, compared to 71.6% of the method of word frequency because 1) the method of word frequency does not take into account the dependent and repetitive of the related contents of celebrity between data sources; 2) the method of word frequency will miss many contents which people names do not appear in it, but really related to the celebrity; 3) the method of word frequency is difficult to extend celebrity if a summary does not contain the celebrity. From Fig. 3, we can also see the two methods that when the number of the selected data sources increase, the accuracy decrease, and the space decrease gradually in that when the number of the selected data sources reached a certain number, the celebrity-related content would be abundant, the difficulty of determining innovative data has also increased and the ratio of innovative data that subsequent data source can provide would be very small.

6. Conclusion

To enhance the efficiency of culture integration in the field of tourism, a strategy of data source selection has been proposed based on celebrity-summary of characteristic words and the effectiveness of this strategy has been confirmed by experiments. In our future work, we will research a new data source selection strategy for applying the entity relation to better meet the demand of mining culture.

References

[1] X. L. Yang, W. Y. Gan, L. H. Yang,Y. L. Wang, Y. G. Fu. *On Web Services-based tourism information integration technology*[J]. Journal of Fuzhou Universiy, 41(2),178-181(2013). (In Chinese)

[2] C. X. Wan, S. Deng, X. P. Liu, G. Q. Liao, D. X. Liu, T. J. Jiang. *Web data source selection technologies*[J]. Journal of Software, 24(4), 781-797(2013). (In Chinese)

[3] R. Balakrishnan, S. Kambhampati. *SourceRank: Relevance and trust assessment for deep Web sources based on inter-source agreement*[C]. Proceedings of the 20th Int'l Conf. on World Wide Web (WWW 2011), New York, ACM. pp, 227-236(2011).

[4] X. L. Dong , B. Saha, D. Srivastava. *Less is more: selecting sources wisely for integration*[C]. Proceedings of the 39th Int'l Conf. on Very Large Data Bases (VLDB 2013), San Francisco, Morgan Kaufmann Publishers, pp,37-48(2013).

[5] T. Rekatsinas, X. L. Dong. *Finding quality in quantity: the challenge of discovering valuable sources for integration*[C]. Proceedings of the 7th Biennial Conference on Innovative Data Systems Research (CIDR'15), New York, ACM, pp, 1-7(2015).

[6] S. Deng, C. X. Wan, X. P. Liu, G. Q. Liao. *Selection of deep Web data sources based on user feedback*[J]. Journal of Chinese Computer Systems, 33(11),2367-2371(2012). (In Chinese)

[7] T. Rekatsinas, X. L. Dong. *Characterizing and selecting fresh data sources*[C]. Proceedings of the 2014 ACM SIGMOD Int'l Conference on Management of Data(SIGMOD 2014), New York, ACM, pp, 919-930(2014).

[8] J. Fan, L. Z. Zhou. *Keyword-based deep web database selection* [J]. Chinese Journal of Computers, 34(40), 1797-1804(2011). (In Chinese)

[9] C. X. Wan, S. Deng, D. X. Liu, T. J. Jiang, X. P. Liu. *Non-cooperative structured deep Web selection based on hybrid type keyword retrieval* [J]. Journal of Computer Research and Development, 51(4), 905-917(2014). (In Chinese)

[10] Y. Wang, W. Zuo, F. He, X. Wang, A. Zhang. *Ontology-assisted deep Web source selection*[J]. Computer Science for Environmental Engineering and EcoInformatics, 159(2), 66-71(2011).

[11] I. Markov, L. Azzopardi, F. Crestani. *Reducing the uncertainty in resource selection*[C]. Proceedings of the 35th European Conf. on IR Research (ECIR 2013), Heidelberg, Springer-Verlag, pp, 507-519(2013).

[12] D. Hong, L. Si, *Search result diversification in resource selection for federated search*[C]. Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13), New York, ACM, pp, 613-622(2013).