Sentences Extraction from Digital Publication for Domain-Specific Knowledge Service

Mao Ye[1,2,3], Lifeng Jin[2], Zhi Tang[1,2], and Jianbo Xu[2]

[1] Peking University, Beijing, China
[2] State Key Laboratory of Digital Publishing Technology
(Peking University Founder Group Co. LTD.), Beijing, China
[3]Postdoctoral Workstation of the Zhongguancun Haidian Science Park, Beijing, China
xjtuyemao@163.com, {lifeng.jin,zhi.tang,jianbo.xu}@founder.com

Abstract. Digital publication resources contain a lot of useful and authoritative information which is normally organized in small sections such as paragraphs, book sections or chapters. It is important to use the information from digital publication resources for knowledge service. In this paper, concepts in a domain are obtained from encyclopedia. Sections are extracted from e-books and then indexed for searching. The related sections for the important concepts are then found by using full text search technique. SVM is used to classify the related sections and the semantic information is computed for the concept. The sentences are then extracted by dynamically extending the adjacent sentences into sentence group. With the method, the sentences extracted are continuous and the length of the sentences would approximate to a specified length statistically. The method is effective for domain-specific knowledge service.

Keywords: knowledge service, sentence extraction, digital publication.

1 Introduction

Knowledge service [1][2] is a high value-added service which manages knowledge from a variety of resources by searching, organization, analyzing and restructuring. It is an advanced stage of information service to solve users' problems [3]. Digital publication contains a lot of useful and authoritative information which is normally organized in sections. The technique to use the information of the digital publication resources for domain-specific knowledge service is important and useful. Our project is to build a domain-specific knowledge service with the digital publication as source of information. The concepts in the domain are extracted from encyclopedia. Encyclopedia is a kind of digital publication which contains a summary of information from either all domains of knowledge or a particular domain of knowledge in the format of articles or entries. Sentences are extracted from e-books for the important concepts. When users learn the concept, they can review the sentences related with the concept for brief information and read the sections or e-books from which the sentences are extracted for detailed information. To build such a knowledge service system, one important step is to extract sentences for concepts from the digital publication re-

C. Stephanidis (Ed.): HCII 2014 Posters, Part I, CCIS 434, pp. 274–279, 2014.

sources. Some methods to extract sentences or passage are presented in the references [4][5][6][7][8]. However, the sentences obtained by these methods are not continuous which may be difficult to understand. In addition, they don't follow an expected length statistically, which may be not easy for displaying or reading. Furthermore, the sentences extracted don't corresponding to a specific domain, which may not fit the requirement for a domain-specific knowledge service.

2 Sentences Extraction from Digital Publication Resources

After the concepts are extracted from the encyclopedia, sentences will be extracted from the e-books for the important concepts. It is needed that the sentences should be continuous and follow an expected length. The length of the sentence is counted by the character number in the sentence. Let $O=\{o_1,o_2,\cdots,o_n\}$ be the concepts set extracted from the encyclopedia and $X=\{x_i,i=1,\cdots,n\}$ be the label set of the concepts in the encyclopedia where x_i is the label of the concept o_i. The main process of our approach for extracting sentences is as Figure 1. Firstly, the sections are extracted from the e-books to get the set $D=\{d_i,i=1,2,\cdots,z\}$, where d_i is a section. The size of the set D is z which may be very large because there are a lot of sections extracted from the e-books. Secondly, all the sections in the set D are indexed with full text indexing technique. The dictionary used for word segmentation in this step consists of all labels X. To index the sections is necessary for us to find the related sections quickly because the number of the sections is large. Thirdly, the set $D^{'}\subseteq D$ is obtained from the indexed full text library for a concept o_i by searching with its label x_i. $D^{'}$ can be a reduced set if the number of the matched sections still be large. Apply support vector machine to classify the sections in $D^{'}$ to get a class label c_i for each $d_i^{'}\in D^{'}$. Support vector machine is an effective method for classification problem [9][10][11]. The domain related sections R are then selected according to its class label. For the concept o_i, obtain the context of the concept in R through the sliding window method. They are the sentences adjacent to the ones in which the label of the concept o_i is displayed. All these sentences are segmented into words and stop words are removed from the result set. Compute the semantic information of the concept and get a vector $W=\{w_1,w_2,...,w_q\}$ which is represented by words' weight, where q is the dimension of the vector. The weight w_i is computed by $\frac{t_i}{\sum t_i}$ where t_i is the occurrence frequency of the i^{th} word. Finally the sentences are extracted dynamically by the vector W. The sentences ex-

tracted are regarded as the related sentences of the concept o_i. They will be combined with other attributes and values of the concepts for domain-specific knowledge service.

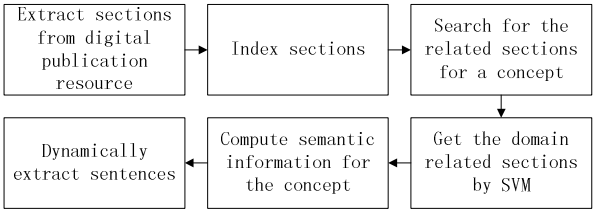

Fig. 1. The main process for extracting sentences

Let E be the expected sentence length to extract from the sections. Let $f(s)$ be the function to compute the length of the sentence s by counting the character number in s. Let m be the initial number of sentences in a sentence group. We will extend the sentence group towards two directions left and right. Let α be the parameter to control the number limit of the sentences which can be extended to the sentence group from a direction. Let $\theta > 1$ be the factor to control the dynamical threshold. Let $W = \{w_1, w_2, ..., w_q\}$ be words' weight for a concept. The detail description to dynamically extracting sentences from a section $r_i \in R$ by W is summarized as the following steps.

Step 1: Split the section r_i to get an ordered list $L_1 = s_1 s_2 ... s_n$ of sentences, where s_i is the i^{th} sentence in L_1.

Step 2: Compute the average weight $\overline{w} = \left(\sum_i w_i \right) / q$.

Step 3: Combine the sentences in the list L_1 to get a sentence group list L_2. Each sentence group $g_i = s_i s_{i+1} ... s_{i+m-1}$ initially consists of m sentences from L_1 in order.

Step 4: Get a sentence group g_i from L_2 and set $k = 1$. Set the weight of g_i by summing up w_i in W for words which the sentence group g_i contains.

Step 5: If $k > \alpha$ or there is no sentence s in L_1 which is on the left side of s_i, go to the next step. Otherwise, compute the weight of the sentence s by summing up w_i in W for words which the sentence s contains. Set the

variable $\beta = E/\left(f(g_i) + \frac{f(s)}{2} \right)$. If $\beta < 1$, then set $\beta = \beta / \theta$. If $\beta > 1$, then set $\beta = \beta * \theta$. If the weight of the sentence s is less than $\overline{w} / \beta$, then go to the next step. Otherwise, append s to the left side of g_i. Accumulate the weight of the sentence s to the weight of g_i. Set $k = k + 1$ and continue to perform this step.

Step 6: Reset $k = 1$.

Step 7: This step is similar to the step 5. The difference is that this step is to append the right side's sentences to the sentence group g_i. If $k > \alpha$ or there is no sentence s in L_1 which is on the right side of s_i, go to the next step. Otherwise, compute the weight of the sentence s by summing up w_i in W for words which the sentence s contains. Set the variable $\beta = E/\left(f(g_i) + \frac{f(s)}{2} \right)$. If $\beta < 1$, then set $\beta = \beta / \theta$. If $\beta > 1$, then set $\beta = \beta * \theta$. If the weight of the sentence s is less than $\overline{w} / \beta$, then go to the next step. Otherwise, append s to the right side of g_i. Accumulate the weight of the sentence s to the weight of g_i. Set $k = k + 1$ and continue to perform this step.

Step 8: Add the sentence group g_i generated to the map M. Go to the step 4 until all sentence groups in L_2 are processed.

Step 9: Sort all the sentence groups in the map M by the weight density in descending order which is computed by $h_i / f(g_i)$, where h_i is the weight of g_i. Select the top N sentences which are the most related sentences for the concept.

3 A Case Study

The concept adopted in the case is the emperor of QinShiHuang (秦始皇) in Chinese history. The aim is to extract the continuous sentences approaching the expected length statistically for the concept in the domain of history. More than 40 thousand e-books are prepared to extract sections which are then indexed in the full text library. The domain-related sections are obtained by searching from the library with the label "秦始皇" and classifying them by SVM. One section obtained is "…秦始皇像 陕西临潼秦始皇陵 战国末年，从诸侯割据向全国统一的趋势已日益明显…公元前238年，他亲理国事，平定嫪毐的叛乱，免除吕不韦的相职，令其徙处蜀郡…终

于建立了中国历史上第一个统一的、多民族的、专制主义中央集权制国家秦朝。…秦二世胡亥即位后，对人民的剥削和压迫变本加厉…不久，秦朝灭亡。”. The length of the section is 1700 characters. The text of the section describes the emperor Qinshihuang and the events and persons related to him in detail. Apostrophe is used in the above text for simplification. Words and their respective weights computed from the data are listed in the table 1. The field “Words” shows the words which are related with the concept Qinshihuang and the field “Weight” shows the relatedness between them. It is shown from the table that the concept Qinshihuang is described by a group of words and weights.

Table 1. Words and weights for the concept of Qinshihuang

Word	Weight	Word	Weight	Word	Weight
统一	0.059343900	孟姜女	0.012286522	焚书坑儒	0.009829217
皇帝	0.035876643	蒙恬	0.011795061	割据	0.008477700
秦国	0.023221526	秦朝	0.011426465	法律	0.008109104
秦王	0.014252365	史记	0.011426465	方士	0.007494778
丞相	0.014129500	赵高	0.011303600	本纪	0.007371913
匈奴	0.013023713	中央集权	0.010812139	刘邦	0.007249048
李斯	0.012900848	分封	0.010320678	秦汉	0.007003317
…	…	…	…	…	…

The parameters E, θ, α, m are set to 300, 10, 3, 3 respectively. One sentence group extracted from the section is “中国统一的秦王朝的开国皇帝。名政，秦庄襄王之子，十三岁即王位，三十九岁称帝，在位共三十七年。…公元前 238年，他亲理国事，平定嫪毐的叛乱，免除吕不韦的相职，令其徙处蜀郡；并任用尉缭，李斯等人，部署统一全国的战略和策略。” The length of the sentence group is 208 which consists of 6 continuous sentences. It describes the role of Qinshihuang and his relationship with other persons very briefly in history domain. The length of the top 28 sentence groups are listed in the table 2. The field “Id” is the identifier of the sentence groups extracted and the field “Length” shows their length. The average length of them is about 313 which are very close to the expected length of 300.

Table 2. Sentences extracted and their length for the concept of Qinshihuang

Id	Length	Id	Length	Id	Length	Id	Length
1	208	8	356	15	299	22	232
2	379	9	354	16	341	23	334
3	276	10	333	17	296	24	255
4	341	11	303	18	289	25	366
5	379	12	382	19	177	26	180
6	423	13	344	20	344	27	261
7	455	14	303	21	305	28	247

4 Conclusions

Digital publication resources are important for knowledge service because they contain a lot of useful and authoritative information. A method is proposed in this paper

to use the information in digital publication resources for domain-specific knowledge service. Sections of e-books can be indexed and the ones related with the domain can be obtained for a concept. Sentences are then extracted to associate with the concepts for knowledge service. Compared with the methods in the references [4][5][6][7][8], the sentences extracted by the proposed methods can match the semantic information of the knowledge concept in the domain. The sentences are continuous and the length of the sentences conforms to a specified length statistically. The case shows the effectiveness of the method. The work described here is a first stage study. The next step in this work is to combine ontology information to improve the relatedness of the sentences with the concepts.

Acknowledgment. The work was funded by China Postdoctoral Science Foundation and Beijing Postdoctoral Science Foundation of China.

References

1. Zhang, Q., Zhang, Q., Peng, X.: Research on Knowledge Service System in Open Innovation Environment. In: 2011 International Conference on Management and Service Science (MASS), pp. 1–4. IEEE Press, New York (2011)
2. Li, G., Song, X.: A New Visualization-oriented Knowledge Service Platform. Procedia Engineering 15, 1859–1863 (2011)
3. Wei, X., Ke, Z., Yatao, L.: Research on an Intelligent Knowledge Service System Based on Internet. In: 2nd International Conference on Education Technology and Computer (ICETC), vol. 2, pp. 506–510. IEEE Press, New York (2010)
4. Alguliev, R.M., Aliguliyev, R.M., Mehdiyev, C.A.: Sentence Selection for Generic Document Summarization Using an Adaptive Differential Evolution Algorithm. Swarm and Evolutionary Computation 1(4), 213–222 (2011)
5. Song, X., Huang, J., Zhou, J.-M., Zhang, H.: A Sentence Selection Method of Query-based Chinese Multi-document Summarization. In: Second Asia-Pacific Conference on Computational Intelligence and Industrial Applications, vol. 1, pp. 224–228. IEEE Press, New York (2009)
6. Ko, Y., Seo, J.: An Effective Sentence-extraction Technique Using Contextual Information and Statistical Approaches for Text Summarization. Pattern Recognition Letters 29(9), 1366–1371 (2008)
7. Wang, D., Zhu, S., Li, T., Gong, Y.: Comparative Document Summarization via Discriminative Sentence Selection. ACM Transactions on Knowledge Discovery from Data 7(1), 1–18 (2013)
8. Salton, G., Singhal, A., Mitra, M., Buckley, C.: Automatic Text Structuring and Summarization. Information Processing & Management 33(2), 193–207 (1997)
9. Cortes, C., Vapnik, V.: Support-vector Network. Machine Learning 20, 273–297 (1995)
10. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.: LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research 9, 1871–1874 (2008)
11. Chang, C.-C., Lin, C.-J.: LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology 2(3), 1–27 (2011)