Design and Implementation of News-Oriented Automatic Summarization System Based on Chinese RSS

Jie Wang[1], Jie Ma[1,*], and Yingjun Li[2]

[1] College of Software, Nankai University, Tianjin, China
[2] Collage of Information Technology Science, Nankai University, Tianjin, China
majie1765@nankai.edu.cn, {nkwangjie,3handsome}@gmail.com

Abstract. Automatic summarization is an important research branch of natural language processing. The automatic summarization should provide information to users from different point of views for better understanding. Aiming at the characteristics of the news, an automatic summarization system is constructed from two aspects: keywords and key sentences. Then, the location factor is added to optimize the keywords extraction algorithm. Meanwhile, the key sentences extraction algorithm is improved through introducing keywords factors. On this basis, in allusion to the existing problems of RSS, this paper builds a user-interest model. Finally, after the verification in terms of the feasibility and the effectiveness, the result shows it is effective to improve the accuracy and the user experience of the RSS feeds

Keywords: Keywords extraction, Key sentences extraction, RSS feeds.

1 Introduction

As a way to share content among sites, RSS has been widely used [1].While huge RSS feeds bring us rich resources, they also produce difficulties to get effective information. How to locate the news that users are interested in and learn main content of them has become a crucial problem.

Under this background, we put forward an idea: we first extract key information from two aspects: keywords and key sentences, and use user-interest model to extract the abstracts which confirm to the user's interests. Meanwhile, we design and implement a personalized automatic summarization system to provide users high-quality services, which has the function of keywords and key sentences automatic extraction. Finally, we do some experiments in terms of the feasibility and effectiveness, and analyze the recall rate and accuracy of the results.

2 Related Work

RSS provides a quick and easy way for the Internet information sharing. RSS information aggregation and customization have obtained certain achievements.

* Corresponding author.

G. Zhou et al. (Eds.): NLPCC 2013, CCIS 400, pp. 378–385, 2013.

Mu, L. [2] implements a personalized information service system of science and technology based on RSS. In this paper, specific user-interest model is established, which can effectively improve the accuracy.

Since 1958 Luhn porposed the concept of automatic summarization [3], many scholars have achieved fruitful work. Kruengkrai, C. et al.[4] used some key information to determine the weight of the sentences, but did not consider the effect of keywords. Meanwhile, the domestic has multiple experiment systems, Lanke system[1] includes an automatic summarization subsystem, which is based on the relevance of the words to calculate. This paper achieves better effect through optimizing the keywords and key sentences algorithm.

3 System Design

3.1 Principle of the Summarization System

The research shows that news writing has obvious features, known as the "5W1H": when, where, who, What, Why and How. Therefore, provided the abstract covers these aspects, it can be considered as meeting requirements.

3.2 Overall Architectural Design of the Summarization System

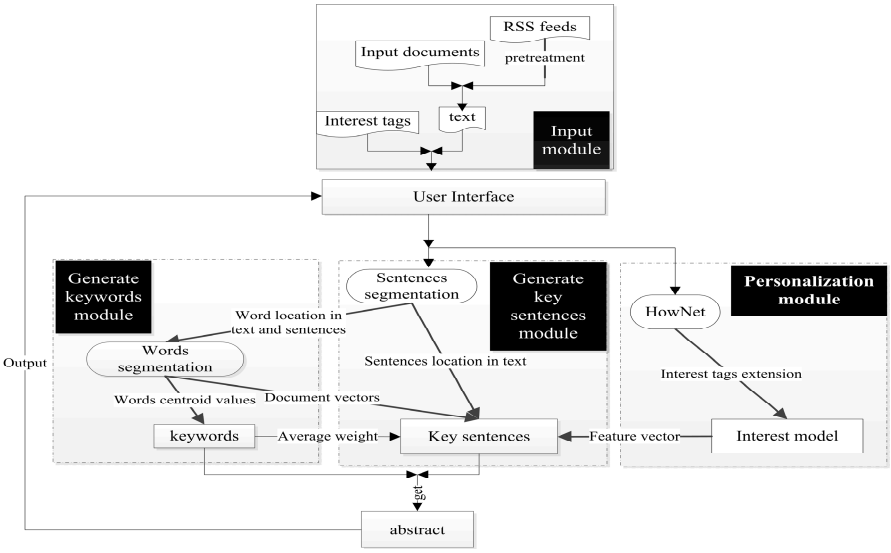

Fig. 1. System frame diagram

[1] http://www.languagetech.cn/class_demo.aspx

4 Core Algorithm

4.1 Optimized Keywords Extraction Algorithm

High frequency words often have a large ratio to become keywords [5,6].In order to improve the accuracy of extraction, we need to add the location of the keywords. Implement method is shown in figure 2:

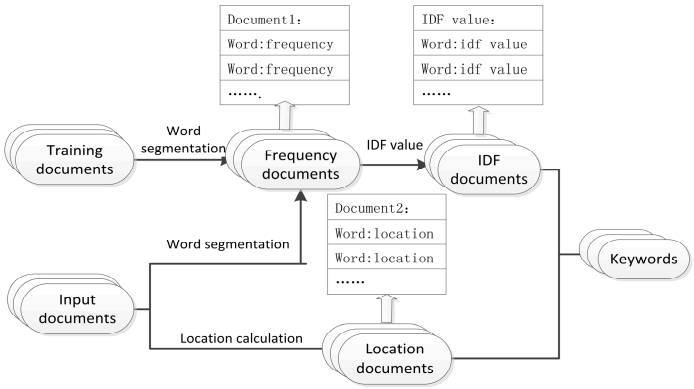

Fig. 2. Keywords extraction module

TF-IDF is a statistical method, as shown in formula (1)(2).

$$TF(w_i) = f(w_i, D) \quad , \tag{1}$$

$$IDF\ (w_i) = \log \frac{|D|}{|\{j : w_i \in d_j\}|} \quad , \tag{2}$$

(1): the occurrence number of *Wi* in the document set *D*.(2) |*D*|:total document numbers,$|\{j: wi \epsilon dj\}|$:the number of document which contains the word *Wi*.

In the training stage, more than 1000 documents have been saved, they are all hot news selected from major portals, and are obtained through web crawler. The larger base is used to calculate the *IDF* to further improve the accuracy.

Our system adds location when calculating the weight. As shown in formula (5)

$$SCORE(S_i) = w_c C_i + w_p P_i \tag{3}$$

$$Ci = \sum_{w \subset Si} TF\ (w) \times IDF\ (w) \tag{4}$$

$$P_i = \left(\sum_{\alpha \subset \beta} \lambda_\alpha \right) \Big/ \beta \tag{5}$$

(4):the centroid value of words,(5):the location of word *Si* in article, $\beta$is the total occurrence numbers of word S_i, $\lambda\alpha$is the weighted values, when the sentence is the title or summary, $\lambda\alpha$=2; when the sentence is the head or tail of the paragraph, $\lambda\alpha$=1.5;in other cases ,$\lambda\alpha$=1. w_c w_p are two constants ,w_c =w_p=1.

4.2 Optimized Key Sentences Extraction Algorithm

In this system, each document is divided into a collection of entries by using the *VSM* model. The implementation is shown in figure 3:

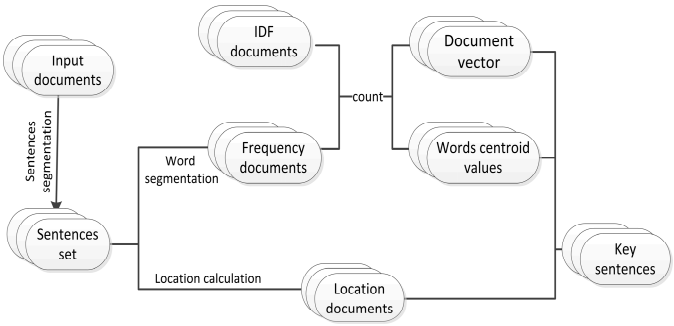

Fig. 3. Key sentences extraction module

The definition of centroid value of document is formula (6):

$$d_i = (Vw_0, Vw_1, ..., Vw_n) \tag{6}$$

$$Vw_i = TF(w_i) * IDF(w_i) \tag{7}$$

Assume S_i is the *i-th* sentence , and the centroid value is formula (8):

$$C_i = \sum_{w \in Si} V_w \bullet f(w, S_i) \tag{8}$$

$f(w,S_i)$ is the frequency of W in the sentence S_i.

Assume P_i is the location of the *i-th* sentence, λi is the weighted value, when the sentence is the title or summary, λi =2; when the sentence is the head or tail of the paragraph ,λi =1.5; in other cases ,λi=1.

$$P_i = \lambda_i * (n - i + 1) / n \tag{9}$$

Assume F_i is the overlap with the title, the inner product as shown in formula(10):

$$F_i = S_i \bullet S_1 = \sum_{w \in Si \bigcap S1} f(w, S_1) \bullet f(w, S_i) \tag{10}$$

Assume T_i is the average weight of the keywords in the *i-th* sentence:

$$T_i = \sum_{\lambda \in S_i} f(\lambda, S_i) \Big/ Len \tag{11}$$

$f(\lambda,S_i)$ is the occurrences numbers of word λ in sentence S_i, *Len* is the number of keywords that S_i contains. Integrated score of a sentence as shown in formula(12):

$$SCORE(S_i) = (w_c C_i + w_f F_i + w_t T_i) * P_i \tag{12}$$

w_c w_f w_t are three constants. w_c=w_f=w_t=1

Experimental results show that the title and abstract are most important to understand the news, next are the head and tail of a paragraph, so it's necessary to improve the weights of these positions when calculating the score.

4.3 Interested Models and Abstracts Correlation Algorithm

The interest tags that are inputted by a user are saved as a vector, then the user-interest model is built up after the vector is extended by *HowNet*.

Assume m_j is the user-interest model, $g(m_j,L_i)$ is the occurrence numbers of the *j-th* interest tag in document L_i, *Len* is the number of interest tags which L_i contains, α is the weighted value of the interest tag, if it is inputted ,α=1; if it is extended ,α=0.5, so the correlation is as shown in formula(13):

$$SCORE(L_i) = \sum \alpha \cdot g(m_j, L_i) \Big/ Len \tag{13}$$

5 Realization

The system is mainly divided into four modules: input module, keywords generation module, key sentences generation module and personalization module. Keywords generation module is composed of the centroid value of words and the word position; Key sentences generation module is composed of the sentence centroid value, the sentence position, the weights of contained keywords and the overlap with title; Personalization module is composed of the user-interest model.

6 Experiments

6.1 Evaluation Criterion

In 1995, Jones divided the evaluation methods into two categories: internal evaluation method and external evaluation method [7].

Here we use the internal evaluation method. By comparing the key information between automatic generation and artificial extracting, we use the recall ratio, the accuracy ratio and the harmonic value(*F_measure)* as evaluation criteria.

6.2 Experimental Method

This paper from two angles verified the feasibility and validity. 160 articles are selected from the RSS feeds of *ifeng* ,*sina* and *sohu*, which are grouped into eight categories: national ,international, financial, military, education, science , history, and sport. Compression rates were divided into 20%, 25%, 30%, 35%, 40%.

Feasibility Verification

Firstly, according to the compression rate, we manually identified the corresponding number of key information, which was used as the evaluation corpus. Then we compared the key information generated by our system with the evaluation corpus, and calculated the average recall rate and accuracy rate. Meanwhile, the mainstream five-point scoring mechanism was used. Eight students were tested about the eight kinds of news, and were given a mark from three angles: the abstract sets accuracy, the abstract content coverage rate and whether easy to understand.

Effectiveness Verification

Our system was compared with the automatic summarization subsystem provided by Lanke, which did not use the key factors to optimize, and its compression ratio was 25%. In order to verify whether the abstract consist of keywords and key sentences can reach better effect, the evaluation mechanism of Q&A was used. A certain number of problem sets and the corresponding standard answers were provided, reviewers were asked to read three different contents: the full text, key sentences, keywords and key sentences. Then we compared their average response time and the accuracy of their answer.

6.3 Experimental Results and Analysis

Feasibility Verification

Table 1. The *F_measure* of different news under compression ration

M / Category	20%	25%	30%	35%	40%
education	0.592	0.645	0.621	0.613	0.625
financial	0.473	0.515	0.541	0.537	0.565
international	0.532	0.646	0.666	0.699	0.707
national	0.496	0.553	0.596	0.611	0.646
sport	0.665	0.703	0.724	0.717	0.721
science	0.641	0.662	0.675	0.684	0.697
military	0.511	0.532	0.547	0.539	0.556
history	0.438	0.456	0.479	0.459	0.461

After using this system, the ratings of users' are as shown in figure 4:

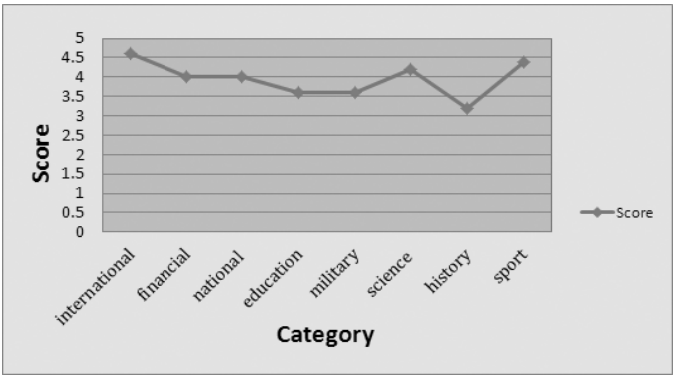

Fig. 4. The scores of all kinds of news

By researching the results in table 1 and finger 4, two groups of independent users gave the same conclusion: the results of international and sport news were better, while history and education news were worse. It is mainly because the extraction algorithm and user-interest model have better effect in international and sport news. Further analysis, we find that for the news emphasizes the timelines such as international news, the title and paragraphs positions are very important. On the contrary, for those timeliness requirements are not that strict or contents are relatively loose, these positions are sometimes not important, and the themes of these articles are often more than one, so when the compression ratio is restricted, it's easy to lose some useful information. Meanwhile, for news having multiple themes, the obtained set of abstracts according to the user-interest model is not often needed by the user. Also this reduces the abstract content coverage rate. So we should consider different strategies for different types of news.

Effectiveness Verification

The difference is given to highlight the results. Results as shown in table 2

Table 2. The comparative result between the two types of system

System / Category	Our system	Lanke system	D-value
Sport	0.703	0.558	0.145
International	0.646	0.412	0.234
History	0.456	0.365	0.091
Education	0.645	0.449	0.196
National	0.553	0.405	0.148
Financial	0.515	0.465	0.050
Science	0.662	0.549	0.113
Military	0.532	0.371	0.161

In table 2, the experimental results show our system has achieved better effect. Especially the international news, with a maximum gap 0.234. Analyzing the Lanke system, we find it just used the method of statistical frequency to extract abstract,

which is the fundamental cause leading to different experimental results, in addition, compared with our system, it lacks of the supplementary of keywords.

Table 3. The Q&A evaluation results

Content	Average time(min)	Accuracy rate
Full text	1.5	93%
Key sentences	0.7	76%
Keywords+Key sentences	1	86%

According to table 3, the advantage and disadvantage of reading full text or key sentences are very obvious, in contrast, the accuracy of reading the keywords and key sentences declines by only 7%, but time reduces by one third. Relatively speaking, the abstract consists of the keywords and key sentences can better achieve the balance between time and accuracy.

7 Conclusion

In this paper, aiming at the existing problems of RSS feeds and the characteristics of the news, we put forward the idea that uses keywords and key sentences as abstract, and select the abstracts which are more relevant to the user-interest model. Experimental results show that our system can effectively improve the quality of automatic summarization in News field.

In the future, the accuracy of the algorithm will be further improved by expanding the data set to correct the weight of the parameters in the formulas.

References

1. Hu, J., Zhang, Z.: Research on personalized information service based on RSS. Computer Applications and Softwar 26(5), 40–42 (2009)
2. Mu, L.: Research on personalized scientific and technological information service system based on really simple syndication. Dalian University of Technology (2008)
3. Luhn, H.P.: The automatic creation of literature abstract. IBM Journal of Research and Development, 159–165 (1958)
4. Kruengkrai, C., Jaruskulchai, C.: Generic test summarization using local and global properties of sentences. In: Proceedings of the IEEE/WIC International Conference on Web Intelligence, pp. 201–206. IEEE, USA (2003)
5. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)
6. Frasconi, P., Soda, G., Vullo, A.: Text categorization for multi-page documents: A hybrid naive Bayes HMM approach. In: ACM/IEEE Joint Conference on Digital Libraries, pp. 11–20. IEEE, USA (2001)
7. Zhang, J., Wang, X., Xu, H.: Survey of automatic summarization evaluation methods. Journal of Chinese Information Processing 22(3), 81–88 (2008)