A Survey on Automatic Summarization

Sicui Wang, Weijiang Li*, Feng Wang, Hui Deng
Computer Appliance Key Lab of Yunnan Province
Kunming University of Science and Technology
Kunming 650051, China
e-mail: wangsicui@cnlab.net, liweijiang@cnlab.net, wangfeng@cnlab.net, denghui@cnlab.net

Abstract—**With the increasing popularity of Internet and the diversity of information obtaining technologies, the amount of quickly growing information has gone beyond our imaginations. Many techniques were presented to help users to find the desired information from large data set quickly and accurately, automatic summarization is an effective approach. In this paper, after careful investigation, existing automatic summarization techniques are classified as five categories: automatic extraction, understanding-based automatic summarization, information extraction, automatic summarization based on discourse and automatic summarization based on user-query. The history of automatic summarization is outlined. The principles of the five categories methods are respectively described in detail. In the end, the five categories methods are compared and the future work is discussed.**

Keywords-automatic summarization; history; principles; comparison

I. INTRODUCTION

Nowadays, with the increasing popularity of Internet and the diversity of information obtaining technologies, large amounts of information are emerging continuously. Facing with so large data set, users are very difficult to find the desired information quickly and accurately. It is helpful for users to retrieve the summary of original article instead of original article, with the purpose of finding the desired information efficiently and accurately.

The automatic summarization is that the computer automatically extracts summary from the original article, and in the ideal case, the summary can describe the main content of article accurately and comprehensively, and the language of the summary is coherent and smooth.

In this paper, after careful investigation, we classify existing automatic summarization techniques into five categories: automatic extraction, understanding-based automatic summarization, information extraction, automatic summarization based on discourse and automatic summarization based on user-query. Then the history of automatic summarization is outlined, and the principles of five categories methods are respectively described in detail. Finally, we compare the five categories methods and discuss the future work.

* Corresponding Author: Weijiang Li, Email: liweijiang@cnlab.net

II. HISTORY OF AUTOMATIC SUMMARIZATION

Automatic summarization was firstly proposed by Luhn in 1958[1].Then, Edmundson and Wyllys[2], Edmundson[3], Rush[4], Paice[5] used formal characteristics of terms as the key to extract summary.

Subsequently, researchers began to consider using syntactic features and semantic features of article to study automatic summarization. Schank and Abelson[6], Fum, Guida and Tasso[7], Jacobs and Rau[8] separately used script analysis, first-order predicate logic inference and framework to express the structure and meaning of the article, then analyzed and inferred to acquire the summary.

In addition, researchers tried to find other approaches to study automatic summarization. With the developing of machine learning, cognitive psychology and linguistics, automatic summarization has entered a new pluralism era. Bonzi and Liddy proposed humanoid approach[9], Kenji Ono et al. researched automatic summarization according to rhetorical structure[10, 11], E.F.skoroxod'ko extracted summary based on relevant sentences network[12], Kupiec, Pedersen and Chen proposed the corpus-based approach to calculate the weight of sentence[13], Barzilay and Elhada extracted summary according to lexical chains[14], Maeda extracted summary according to pragmatic analysis[15], Marcu used rhetorical structure tree to extract summary[16]. Chali, Matwin and Szpakowicz used query expansion to extract summary [17].

In China, until 1958, researchers began to study automatic summarization. From the late 80s of the 20th century, some universities and research institutes began to study the automatic summarization, and made great contributions. Mo, Wang and Xu developed OA Chinese automatic summarization system [18, 19]. Li and Xu designed the EAAS system under the guidance of Ma. Wang developed the MATAS automatic summarization system of military field, HIT-863 I automatic summarization system, HIT-97 I English automatic summarization system and HIT-863 II automatic summarization system [20]. Zheng, Huang and Wu developed the automatic text summarization system [21]. Zhong developed Glance system for computer viruses, News system for news coverage, and Ladies system for the learning algorithm field of neural network [22].

III. AUTOMATIC SUMMARIZATION METHODS

A. Automatic Extraction

1) The Principles of the Automatic Extraction

The basic idea of automatic extraction is that article is a linear sequence of sentences, and sentence is a linear sequence of terms [23]. Automatic extraction has four steps:

- Calculating the weight of term.
- Calculating the weight of sentence.
- Sequencing sentences in descending order by their weights, then defining threshold value, the sentence whose weight is above the threshold value is selected as summary sentence.
- Outputting all the summary sentences according to their appearance order in the original article.

Various studies have led to the proposal of the following criteria of measuring sentence significance for effective summary generation [20]: Frequency of Term is abbreviated as F, Title is abbreviated as T, Location is abbreviated as L, Syntactic Structure is abbreviated as S, Cue Terms is abbreviated as C[1, 2], and Indicative Phrases is abbreviated as I [5].

2) Sentence Weighting

In the automatic extraction, calculating the weight of sentence generally uses the vector space model based on term statistics. In the vector space model, the sentence S_i is formalized as $S_i=<W_1,W_2,W_3,\ldots,W_n>$, each dimension of this vector indicates the weight of the term w_i: $W_i=<F(w_i),T(w_i),L(w_i),S(w_i),C(w_i),I(w_i)>$. The weight of term w_i is calculated as follows [23],

$$\begin{aligned} Score(w_i) = {} & x_1 * F(w_i) + x_2 * T(w_i) + x_3 * L(w_i) \\ & + x_4 * S(w_i) + x_5 * C(w_i) + x_6 * I(w_i) \end{aligned} \quad (1)$$

where x_1, x_2, x_3, x_4, x_5 and x_6 are adjustment coefficients, different articles have different adjustment coefficients.

The weight of sentence S_i is calculated as follows [23].

$$Score(S_i) = \sum_{j=1}^{n} Score(w_j) \quad (2)$$

B. Understanding-based Automatic Summarization

Understanding-based automatic summarization utilizes linguistics knowledge to acquire language structure, and more importantly uses domain knowledge to judge and infer to acquire the meaning expression of article, eventually generates the summary form meaning expression. It has four research topics: script, concept dependency structure, frame and first-order predicate. They all can be divided into four steps [20]:

- Parsing. Using linguistics knowledge in dictionary to parse sentences, then generating syntax tree.
- Semantic Analysis. Using semantic knowledge of the Knowledge to convert the syntax tree into semantic expression based on logic and meaning.
- Pragmatic Analysis and Information Extraction. According to domain knowledge pre-stored in the Knowledge, reasoning in context, and then storing the extracted main content into an information table.
- Summary generation. Converting the content in the information table to a complete and coherent summary, and then outputting.

C. Information Extraction

Summary framework is the core of the information extraction. Information extraction can be divided into two phases: the selection phase and the generation phase. Summary framework proposes the content which would be extracted from the original article in form of empty slot.

In the selection phase, summary framework is filled with relevant phrases or sentences which are extracted from the original article. In the generation phase, the content in the summary framework is converted to summary.

D. Automatic Summarization Based on Discourse

Discourse is an organic structure. Different parts of discourse bear different functions, and have complex relationships among them. Automatic summarization based on discourse attempts to analyze the structural features of discourse to identify the main content of the article. Currently, automatic summarization based on discourse has five main research topics: rhetorical structure analysis, pragmatic analysis, lexical chain, relationship map and latent semantic analysis.

1) Rhetorical Structure Analysis

Rhetorical structure theory was proposed by Mann and Thompson, and its core is the rhetorical relationship. The rhetorical relationship is a relationship which connects two non-overlapping sets (the Nucleus set and the Satellite set).

The Nucleus set is different from the Satellite set [24]. The Nucleus set performances the intention of author, but the Satellite set assists readers to understand and proves the Nucleus set. And the understanding of the Nucleus set is independent on the Satellite set, but the understanding of the Satellite set is dependent on the Nucleus set.

This method gets the rhetorical relationships between sentences according to conjunctions and predictions, then establishes rhetorical relationship analysis tree. According to the significance of rhetorical relationship, we can extract the main content of article [11].

2) Pragmatic Analysis

Pragmatic analysis is only applied to scientific articles. Scientific article has very strict format. Different parts of the scientific article assume different pragmatic functions. Thereby, through analyzing the pragmatic function, the main content of the article can be identified to compose summary.

3) Lexical Chain

Lexical chain is a term-set. It is consisted of all the terms with the same concept in an article, so every lexical chain represents a concept described in the article.

This method firstly extracts noun from article, then acquires the concept of term through the HowNet[25], finally composes all the terms with the same concept to a lexical chain. After establishment of lexical chain, the strongest lexical chains are defined and sorted in descending order. Because all the terms in a lexical chain reflect the same concept, we select a typical term from each strongest lexical chain to express the theme of the lexical chain, eventually select sentences which contain the typical terms as summary sentences.

4) Relationship Map

Relationship map is a network. It treats every sub-unit (paragraph or sentence, or even term) as a node. When there are semantic relations between two nodes, the relationship map links them with an edge. In the network, the number of edges of node is called node degree. The greater the node degree, the more significant the node is in the network.

Shorter article is generally treated as a sentences relationship network, in which the sentences with high node degree are selected as summary sentences. For longer articles, the relationship network among sentences is very large, so we can treat article as a paragraph relationship network.

5) Latent Semantic Analysis

Latent semantic analysis is a knowledge model based on probability statistic. It uses singular decomposition and reducing dimension to abstract inputted knowledge model. The whole process not only highlights the implicit semantics, but also raises the original inputted knowledge model to the higher semantic level [26].

Firstly, an article is as a unit to build Word-by-Sentence matrix denoted by A, the element of the A represents the eigenvalues of a term in a sentence. The Word is verb acted as SVO or noun in the article. Then singular decomposition is realized on A and make $A=USV^T$, where S represents semantic space, U represents the notation of keywords in this space, V^T represents the notation of Word in the semantic space. The semantic space is the definition space of every term in article. Through reducing dimension to rebuild $A'=U'S'V'^T$, now we acquire semantic Word-by-Sentence matrix, in which each row vector represents the weight of the keyword in every sentence, and each column vector represents the meaning of the sentence composed by every keyword. Through calculating inner product value of any two column vectors of matrix, we can infer their sentences semantic relevance.

TABLE I. THE COMPARISON OF THE FIVE CATEGORIES AUTOMATIC SUMMARIZATION METHODS

	Automatic extraction	**Understanding-based automatic summarization**	**Information extraction**	**Automatic summarization based on discourse**	**Automatic summarization based on user-query**
Summary generation	*The summary is composed by the sentences of original article.*	*The summary is generated from the meaning expression of article.*	*The summary is generated by the content of summary framework.*	*The summary is composed by the sentences of original article.*	*The summary is composed by the sentences of original article.*
How to select summary sentences from original article	*The basis is the six formal characteristics of term: F, T, L, S, C, I.*	*No.*	*No.*	*The basis is the discourse significance of sentence in article.*	*The basis is the user query and the six formal characteristics of term: F, T, L, S, C, I.*
Used methods	*Statistics method.*	*Parsing, semantic analysis, pragmatic analysis.*	*Statistics method.*	*Rhetorical analysis, pragmatic analysis, lexical chain, relationship map, latent semantic analysis.*	*Statistics method.*
The level of analyzing the article	*Terms of article.*	*Meaning of article.*	*Terms of article.*	*Discourse of article.*	*Terms of article.*
Application areas	*Any areas.*	*Certain area.*	*Certain areas.*	*Any areas.*	*Any areas.*
Application articles	*Any articles.*	*Any articles.*	*Any articles.*	*The articles with clear structure.*	*Any articles.*
Algorithm difficulty	*Low.*	*High.*	*Low.*	*Middle.*	*Low.*

E. Automatic Summarization Based on User-query

Automatic summarization based on user-query was proposed by Jiang, Fan and Chen in 2008. Its core idea is to integrate the sentence weight based on user-query into the sentence weight calculation [27].

The weight of sentence S_i is calculated as follows [27],

$$Score(S_i) = \mu * W(S_i) + (1-\mu) W(q, S_i) \quad (3)$$

μ is a adjustment coefficient. $W(S_i)$ is the weight of sentence which is calculated by formula (2). $W(q,S_i)$ is the sentence weight based on user-query q, it is calculated as follows[27],

$$W(q, S_i) = \frac{C^2(q, S_i)}{len(q)} \quad (4)$$

where $C(q,S_i)$ indicates the number of S_i containing the keywords in the query q, $len(q)$ indicates the number of keywords in the query q.

IV. COMPARISON AND DISCUSSION

In this chapter, the five categories automatic summarization methods are compared in seven aspects in TABLE I. we can find from TABLE I that automatic summarization based on discourse is better than the other four methods.

Understanding-based automatic summarization uses complex natural language understanding and generation technologies to analyze the article, so it can grasp the main content of original article more accurately. However, it requires the computer not only to understand and process the natural language, but also to express and organize all kinds of backgrounds and domain knowledge. These tasks are very difficult to realize, and the application area of it is limited to certain area.

As automatic extraction and automatic summarization based on user-query, the summary of automatic summarization based on discourse is composed by the

sentences extracted from original article. Compared with automatic extraction and automatic summarization based on user-query only analyzing the terms of original article, automatic summarization based on discourse analyzes the discourse of original article, so it can grasp the main content of original article more accurately and comprehensively.

Information extraction also only analyzes the terms of original article. In addition, the summary of information extraction is generated by summary framework, and a summary framework only corresponds to an area, so the application area of information extraction is limited to certain area, and the language of summary is stereotyped.

V. CONCLUSION

The ultimate goal of automatic summarization research is to make the computer read various articles like human beings and generate satisfactory summaries. However, due to the flexible of natural language and the limited capacity of computer processing natural language, the summaries generated by existing automatic summarization techniques are unable to meet users' need.

We intend to integrate automatic summarization based on user-query into automatic summarization based on discourse to extract the summary from original article. In theory, the quality of the summary should be more adapted to the original article and more helpful for users to find the desired information.

ACKNOWLEDGMENT

This work is in part supported by the Natural Science Foundation of Yunnan Province of China under award number 2009ZC032M. We would like to thank all the authors of the papers which we surveyed. We thank the members of Yunnan Compute Technology Application Key Lab in Kunming University of Science and Technology for their support and assistance. We are also grateful to anonymous reviewers for your insights.

REFERENCES

[1] H. P. Luhn, "The automatic creation of literature abstracts," IBM J. Res. Dev., vol. 2, no. 2, pp. 159-165, 1958.

[2] H. P. Edmundson, and R. E. Wyllys, "Automatic abstracting andindexing\—survey and recommendations," Commun. ACM, vol. 4, no. 5, pp. 226-234, 1961.

[3] H. P. Edmundson, "New Methods in Automatic Extracting," J. ACM, vol. 16, no. 2, pp. 264-285, 1969.

[4] J. J. Pollock, and A. Zamora, "Automatic Abstracting Research at Chemical Abstracts Service," Journal of Chemical Information and Computer Sciences, vol. 15, no. 4, pp. 226-232, 1975.

[5] C. D. Paice, "The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases," in Proceedings of the 3rd annual ACM conference on Research and development in information retrieval, Cambridge, England, 1981.

[6] R. C. Schank, and R. P. Abelson, Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures: Lawrence Erlbaum Associates, Hillsdale, New Jerse 1977.

[7] D. Fum, G. Guida, and C. Tasso, "Forward and backward reasoning in automatic abstracting," in Proceedings of the 9th conference on Computational linguistics - Volume 1, Prague, Czechoslovakia, 1982.

[8] P. S. Jacobs, and L. F. Rau, "SCISOR: extracting information from on-line news," Commun. ACM, vol. 33, no. 11, pp. 88-97, 1990.

[9] S. Bonzi, and E. Liddy, "The use of anaphoric resolution for document description in information retrieval," Inf. Process. Manage., vol. 25, no. 4, pp. 429-441, 1989.

[10] S. Miike, E. Itoh, K. Ono et al., "A full-text retrieval system with a dynamic abstract generation function," in Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, Dublin, Ireland, 1994.

[11] K. Ono, K. Sumita, and S. Miike, "Abstract generation based on rhetorical structure extraction," in Proceedings of the 15th conference on Computational linguistics - Volume 1, Kyoto, Japan, 1994.

[12] B. A. Mathis, and J. E. Rush, "Abstracting," Encyclopedia of Computer and Tehcnology, vol. 1, pp. 102-142, 1975.

[13] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer," in Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, Washington, United States, 1995.

[14] R. Barzilay, and M. Elhada, "Using Lexical Chains for Text Summarization Proceedings of Intelligent Scalable Text Summarization Workshop." pp. 10-17.

[15] T. Maeda, "An approach toward functional text structure analysis of scientific and technical documents " Information Processing&Management, vol. 17, no. 6, pp. 329-339, 1981.

[16] D. Marcu, The Theory and Practice of Discourse Parsing and Summarization: MIT Press, 2000.

[17] Y. Chali, S. Matwin, and S. Szpakowicz, "Query-Biased Text Summarization as a Question-Answering Technique," AmericanA ssociation for Artificial Intelligence, pp. 52-56, 1999.

[18] Y. Mo, and Y. Wang, "Automatic Abstraction of Chinese Document," New Technology of Library and Information Service, vol. 3, pp. 10-12, 1993.

[19] H. Xu, and Y. Wang, "OA Automatic Abstracting System on Chinese Documents," Journal of The China Society For Scientific and Technical Information, vol. 16, no. 2, pp. 128-132, 1997.

[20] K. Wang, and T. Liu, "Four Kinds of Main Methods of Automatic Abstracting," Journal of The China Society For Scientific and Technical Information, vol. 18, no. 1, pp. 10-19, 1999.

[21] Y. Zheng, S. Huang, and L. Wu, "Research and Implementation of Automatic Multi-Document Summarization System," Journal of Computer Research and Development, vol. 40, no. 11, pp. 1606-1611, 2003.

[22] L. Li, X. Guo, and Y. Zhong, "An Understanding-based Chinese Automatic Abstract System in Special Field," Journal of Computer Research and Development, vol. 37, no. 4, pp. 6-10, 2000.

[23] C. Tan, and Y. Chen, "Literature Review of Automatic Summarization Methods," Journal of the China Society for Scientific andTechnical Information, vol. 27, no. 1, pp. 62-68, 2008.

[24] Draft, "Rhetorical structure," June, 2006.

[25] Z. Dong, and Q. Dong, http://www.keenage.com.

[26] J. Steinberger, "Using Latent Semantic Analysis in Text Summarization and Summary Evaluation," Proc.ISIM 04, 2004.

[27] X. Jiang, X. Fan, and K. Chen, "Research on Chinese automatic summarization based on user-query," Computer Engineering and Applications, vol. 44, no. 5, pp. 48-50, 2008.