Journal of Computational Information Systems 11: 7 (2015) 2663–2672
Available at http://www.Jofcis.com

A New Method for Text Summary Extract Base on Event Network ⋆

Junhui YANG [1,2,*], Zongtian LIU [1], Wei LIU [1], Xiaoying SHU [1]

[1] *School of Computer Engineering and Science, Shanghai University, Shanghai 200072, China*

[2] *School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China*

Abstract

Text can be partitioned and expressed by events and let events as the basic semantic unit to establish event ontology. Event network direct diagram constructed according to the relationship between events, which can better describe semantic information of the text and the importance degree of relationship between events. PAGERANK algorithm is employed to calculate and rank the event importance degree corresponding to each node in network diagram to output the original statement corresponding to events as abstract according to the time sequence of events. The experimental results show that extraction method of automatic summary based on the event network has achieved better performance.

Keywords: Text Representation; Event Ontology; Event-network; PAGERANK

1 Introduction

The essence of automatic summary technology is information concentration and mining, one of the important signs to comprehend natural language, and the research on automatic summary technology may contribute to comprehend natural language text and acquire knowledge model. Automatic summary technology began to rise from the 50's of the twentieth Century, initially supported by statistics, which generated summary for articles relying on the words frequency, location and other information, mainly applied to the more formal technical documents [1].

Text event network refers to text can be partitioned and expressed based on events as the basic semantic unit to establish event ontology and constructs event network according to the relationship between events, using directed graph model to reflect the structure relationship between the text events, which could well express the semantic information of text, make up for the deficiencies of the frequency characterization. Then PAGERANK adopted to measure the degree of events corresponding to each node, showing the importance degree of relationship between events and

⋆Project supported by the National Nature Science Foundation of China (No. 61273328, 61305053, 31260292) and Shanghai Natural Science Foundation of China (No. 12ZR1410900).

*Corresponding author.
Email address: tkzy@shu.edu.cn,jwcjhy@126.com (Junhui YANG).

1553–9105 / Copyright © 2015 Binary Information Press
DOI: 10.12733/jcis13975
April 1, 2015

sorting, output the original statement corresponding to events as abstract according to the time sequence of events.

The content of this paper is organized as following. After the introduction is given by Section 1, some related works are presented in Section 2. In Section 3, on the basis of defining some basic concepts and related theorems of event and event network, then introduce how to establish event network and gived the algorithm of reasoning. And the experiments and evaluate are proposed by Section 4. Finally, we give conclusions and discuss future works in Section 5.

2 Relation Works

In recent years, with the in-depth research on automatic summary, researchers have proposed different methods. Lin et al proposed another summary method [2]. They assumed that various features used for summary extraction are interrelated, and they use decision trees rather than Bias classification model to score sentences to extract sentences with the highest scores as summary.

Jiang et al adopt identifying combinations of words and paragraphs clustering to achieve Chinese automatic summary [3]. First, calculate their weights according to the frequency, part of speech, location and the length of words or phrases, and thereby calculate the weight of sentences; then poly adjacent paragraphs to the same class or different classes according to the similarity; finally, select summary sentences to constitute summary according to the weight of the sentences within the class.

Zhang proposed a method for automatic summary extraction based on sentence clustering [4]. First, cluster sentences in text according to semantic distance, then calculate weights for each sentence based on multi-feature fusion method, finally extract sentences to constitute summary with certain rules.

M. Chandra proposed a kind of automatic summary based on statistical methods, used kmixture probability model to establish the value of feature words, identify the semantic relations between words, and extract sentences corresponding to feature words and eventually determine summarization by ranking the weights [5].

Erkan proposed text processing based on graph, by which text can be divided into sentence sets to construct the graph of the vertex by sentences, calculate sentence salience by graph, and then extract sentences according to the saliency [6].

Zwaan from America Florida University equated each sentence as an "event" [7]. At the ACE meeting, "event" was described as an action occurred or states changed [8]. Some scholars began to recognize that event as the basic unit of knowledge, can better reflect the motility essence of objective world knowledge, suitable for the research of text automatic summarization and events ontology.

Liu proposed a multi document summarization based on event term semantic graph clustering method, regarding the verbs and gerunds as events terms, then cluster the event terms, and extract sentences that contains event terms to constitute summarization [9].

Han proposed the multi-document automatic summarization method of Netnews based on event extraction, discriminated the events or non-events in text through the binary classifier; then changed from Physical partitioning in paragraphs or sentences to content logic partition in events through clustering, and finally generated summarization through the extraction, ranking and polish of main events [10].

SS Ge proposed a clustering algorithm based on $SNMF$ and consider the relations between sentences to sort the statement by weighted graph model, finally form summarization [11].

Peter Thwaites proposed the construction of the chain of events graph (CEG) as an auxiliary graphical model to analyze and represent the causal Bayesian network of the events [12].

Zhong proposed a sorting method for text event importance based on event relation graph to describe the degree of relevance between events in text collection by constructing the event influence factor matrix; expound identification method for the important events in a set of documents [13].

In conclusion, although some scholars have proposed the event as the unit to outline the content of a document, even obtain event term semantic relationship through the external semantic resource to achieve automatic summarization of document, but degree of correlation between the events were considered as the main factor to determine the importance degree of the sentence. Although some articles mentioned the paragraph or sentence as a unit event, calculate the distance from the unit event to the center event to determine the extraction of main events, but all of them do not consider the role of the relationship between the events happened.

3 Model of Extraction Adstract Based on Event Network

3.1 Definitions on event and event network

Definition 1. (Event) Occurring in a particular time and environment with some characters involved, it refers to something which shows some movement features [14]. Event e can be defined as a 6-tuple formalize $E = \langle A, O, T, V, P, L\rangle$ in which the elements are called event elements, representing object, movement, time, environment, assertion, Language Performance.

(Event Class) Event Class means a set of events with common characteristics, defined as:

$$EC = (E, C_1, C_2, \cdots, C_6). \tag{1}$$

Where E is event set, called extension of event class. C_i called connotation of the event class. It denotes the common characteristics set of certain event factor (factor i). C_{im} denotes one of the common characteristics of event factors i.

Definition 2. (**Event Ontology**) An event ontology is a formal, explicit specification of a shared event systematic model that exist objectively, denoted as EO. The logic structure of event ontology can be defined as a 3-tuple. $EO := \{EC_S, R, Rules\}$. Where EC_S is the set of all events, R indicates all taxonomic relationships and non-taxonomic relationships between events. Taxonomic relationships could construct a hierarchy of event classes while non-taxonomic relationships could label relationship types. As is shown below:

Definition 3. Events non-taxonomic relationships

- **Composite Relationship**
  If event e could be decomposed to several events $e_i(i \in [1, n])$ with smaller granularity. There exists composite relationship between them, and events $e_i(i \in [1, n])$ was the part of e. denoted as $R_I(e_i : e_1, e_2)$.

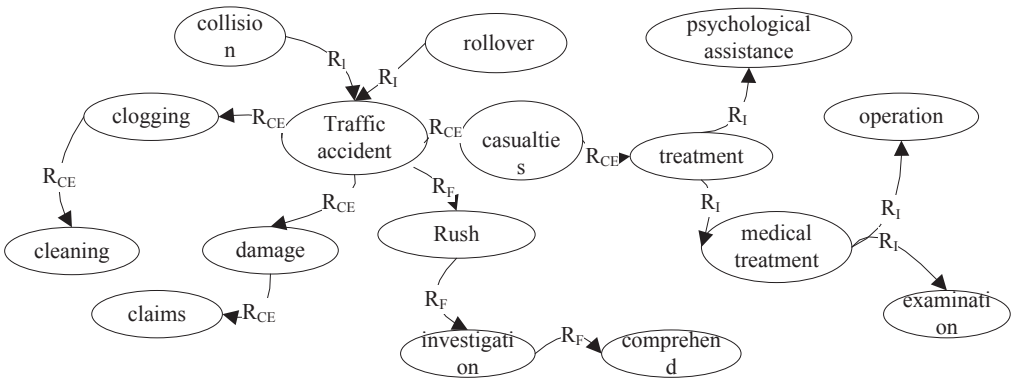

Fig. 1: Emergency event ontology model (part)

- **Causal Relationship**
  If an event e_1 causes e_2 to happen, and the probability of occurrence was larger than a given threshold, namely, there has causal relationship between events, denoted as $R_{CE}(e_1, e_2)$.

- **Follow Relationship**
  Within a certain length of time, event e_2 may follow with event e_1 at a larger specified probability threshold, there has a follow relationship between and, denoted as $R_F(e_1, e_2)$.

- **Accompany Relationship**
  If e_1 concur with e_2 in a certain period of time, and the occurrence probability is larger than a specified threshold, there has an accompany relationship between and, denoted as $R_C(e_1, e_2)$.

Definition 4. (**Event elements fill**) Refers to the six elements of any event need to display in the text.

Considering that the Chinese text description events often omit part of sentence composition, thus some event elements could not be clearly expressed. For example, "A semi-trailer collided with a minibus waiting for clearance and burst into flames". Three event trigger words "collided", "fire", "flame" involved, but the text only describe the event trigger essential factors of "collision", other essential factors of events trigger "fire" and "flame" did not describe the six elements, only describe the action essential factor. Therefore, it need fill up completion all event elements, could be compare relations between events and calculation similarity of events. The event elements fill algorithm describes.

Step 1 Word segmentation and tagging available elements of events in text, at the same time determine which events elements need to fill.

Step 2 Select the most near event elements as candidate event elements which events existence of non taxonomic relations with need to fill elements event.

Step 3 Judging whether to the event which had filled event elements satisfy the non taxonomic relation with the approach event. If satisfied, then could determine the candidate event elements for the event elements. Otherwise, select next existence non taxonomic relationships event elements as candidate event elements, until event which had fill elements could satisfied non taxonomic relations with the previously defined event.

Definition 5 (**Event similarity)** It indicates the degree of the similarity of the events, usually expresses by the value of [0, 1] interval.

Suppose an event set contains two events e_i and e_j, event similarity can be calculated according to the similarity corresponding to event elements, denoted as:

$$SIM(e_1, e_2) = \sum_{k=1}^{6} w_k s(e_{ik}, e_{jk}), k = (o, a, t, v, p, l) \tag{2}$$

Where $SIM(e_i, e_j)$ refers to the similarity between e_i and e_j, e_{ik} indicates the k elements of event e_i and e_{jk} indicates the k elements of event e_j, w_k indicates the weight of the events for calculation event similarity, denoted as $\sum w_k = 1$, the weight of the event similarity between e_i and e_j is [0, 1]. obviously, the similarity between e_i and e_j is [0, 1], that is to say, two events could be exactly the same, otherwise they may have nothing to do with each other. The "0" (less than a certain defined threshold) expresses no similarity between two events whereas "1" expresses two events exactly the same, namely, the same event.

To calculation of similarity elements of events need combine with synonyms, if they were synonyms (casualties and injuries), then its similarity is defined as "1", if there exists a relationship (fire and flaming) it consider as similar and the similarity is "0.5". According to status of event essential factor in the text, it defined w_1=0.5, w_2=0.3, $w_{3,4,5,6}$=0.1. Through the experiment, when the event similarity $SIM(e_i, e_j) \geq 0.7$, could think of (e_i,e_i) is similar.

Definition 6 (**Event Network)** Event network refers to a collection of directed graphs containing a series of event nodes and edges, nodes represent events, edge represents the relationship between events. Unidirectional edges represent the relationship between events; bidirectional edges represent the events of high similarity between events. Formalization denoted as:

$$GRE = [Events, Ls, W] \tag{3}$$

$$Events = [e_1, e_2, \cdots, e_n] \tag{4}$$

$$Ls : (e_1, e_2, l(e_1, e_2)), (e_1, e_3, l(e_1, e_3)), \cdots, (e_i, e_j, l(e_i, e_j)) \tag{5}$$

$$W = l(e_i, e_j), W \in [0, 1] \tag{6}$$

Where GRE indicates event network.

Events indicate events set, every e_i of the nodes set $N = e_1, e_2, \cdots, e_n$ expresses one of the event essential factor and n stands for the numbers of nodes in the graph.

Ls indicate relationship sets between events, in the directed edges $E = \{\cdots, l_{ij}, \cdots\}$, every direct edge $l_{ij}(i, j = 1, 2, \cdots, n, i \neq j)$ represent the kinds of relationships of neighboring nodes maps events.

W indicates interlinking degree between unit events and other events, using interval between [0, 1] to represent the value of $l(e_i, e_j)$, and $\sum_{k=1}^{n} = l(e_i, e_j)$.

3.2 Construct event network

To build a relationship between events, event and event elements would be first extracted from the labeled text D, relationship between events was to be extracted according to constructed event

ontology, then achieved text set $E = \{e_1, e_2, \cdots, e_i, e_j, \cdots, e_n\}$; On this basis event network was built, construction steps were as follows.

Step 1 Initialize the set of nodes $N_d = \{\}$, directed edge set $E(d) = \{\}$.

Step 2 Unit events in the text D event set $E(D) = \{e_1, e_2, \cdots, e_i, e_j, \cdots, e_n\}$ map to the event network diagram nodes in turn, getting the node set $N_d = \{n_1, n_2, \cdots, n_i, n_j, \cdots, n_k\}$.

Step 3 Select one of set nodes n_i as any node of event network, and seek the associated nodes with n_i from node set N_d in turn. If they exit causality or follow relationships between the two node then add a direct edge $\overrightarrow{e_i e_j}$ between n_i and n_j. If they exit accompany relationship then add a bidirectional edge $\overleftarrow{e_i e_j}$ and $\overrightarrow{e_i e_j}$ between n_i and n_j.

Step 4 Select one node n_i in the set of the node N_d, in proper order ergodic other node n_j in set of N_d, then calculated event elements similarity, if the similarity is greater than or equal to the threshold (here the threshold set to 0.7), then add two oppositely directed edges $\overleftarrow{e_i e_j}$ and $\overrightarrow{e_i e_j}$ between n_i and n_j.

Step 5 According to [Step 3] and [Step 4] could obtain direct graph set $E(D) = \{\cdots, e_{ij}, \cdots\}$ and event network digraph of text D.

An example on event network listed below to show the relationship between events, text extract from http://www.zaobao.com/realtime/china/story20140325-324877, "An accident took place in Qianjiang section of Baomao Expressway". The text contains 4 paragraphs, 8 sentences and 28 events. Using the CEC annotation tool for event annotation of text D, and our method adopted to construct event network directed graph for text D, including 28 nodes, 88 directed edges. Text can be visualized through NetDraw, each node can be identified using events and action elements; The relationship between events would be described via lines with arrows, as shown in Fig. 2.

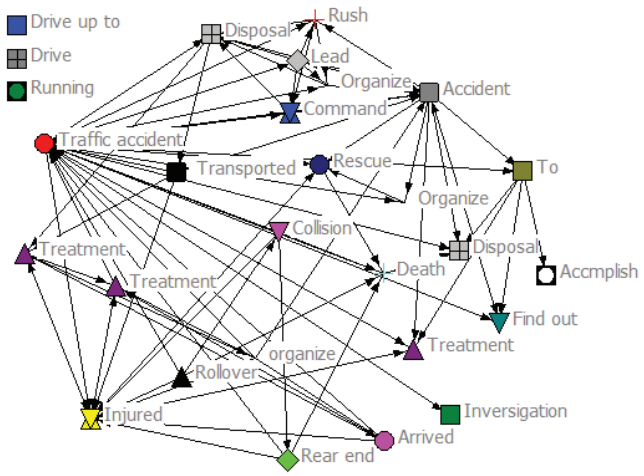

Fig. 2: Event network directed graph

3.3 Formation abstract based on event topic extraction

In order to obtain the text abstract, it needs sorting the event importance of the text then concatenate all text events, ultimately generating abstract. Because the text description of the event tends to describe the same event with similar language, therefore in the acquisition of event importance, not only construct the event network for association events in text, but also the high similarity event, and calculated the importance of each node in the event network. Then select the highest importance events as event theme events. The abstract of event network representation model as shown in Fig. 3.

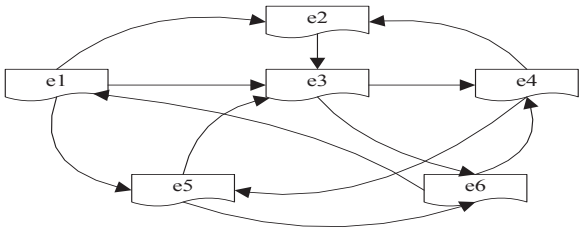

Fig. 3: Model of events network

As could be seen from Fig. 3 there were one or more directed edges connected each other between events, if a node have K sides chain out, N sides chain in. that is to say, current events may affect K other events, meanwhile it may be affected by N other events. So we could represent the relations between events through a matrix between chain in edges and chain out edges.

$$W = \begin{vmatrix} l(e_1,e_1) & l(e_1,e_2) & \cdots & l(e_1,e_n) \\ \vdots & \vdots & l(e_i,e_j) & \vdots \\ l(e_{n-1},e_1) & l(e_{n-1},e_2) & \cdots & l(e_{n-1},e_n) \\ l(e_n,e_1) & l(e_n,e_2) & \cdots & l(e_n,e_n) \end{vmatrix} \tag{7}$$

Where $l(e_i,e_j)$ indicates one of edge node e_i point to node e_j direction. If the edge exit then $\sum_{i=1}^{n} l(e_i,e_j) = 1$ otherwise $l(e_i,e_j) = 0$.

The degree of important events using the classical PAGERANK [15] algorithm sort of nodes chain out or chain in based on event network, the formulae for degree of the each node in the graph:

$$R(e_i) = d \sum_{i \in L(e_i)} R(e_j)/L(e_j) + (1-d)/n, d \in [0,1] \tag{8}$$

Where e_i, e_j, were any node in the event graph, $R(e_i)$ represents the importance of events e_i, $R(e_j)$ express the important of events e_j. $L(e_i)$ Represents a collection of connected line points to e_i (total events of lead to e_i happened), $L(e_j)$ represents the number of a connection line of points to other events e_k and $k \in n$ (the total events of e_j cause e_k). n is the number of nodes in the graph (a number associated with the event), d is a parameter, as an attenuation factor, also known as the damping coefficients [0, 1], usually take $d = 0.85$.

Sorting theme events was essential for generating summary. If the sequence was improper, it would decrease the summary quality and reliability. Therefore the theme events should be sorted

according to the importance and development process of events. It should be sorted according to the development process of events based on the importance degree, if the theme events can not compare time but belong to the same document, it should be sorted according to the sequence order. Finally, gradually remove information contribution minimum sentences until the sum of the rest sentence length reaches target abstract length.

4 Experimental Results and Analysis

4.1 Performance analysis of the experimental data and evaluation

In the experiment, 203 text corpuses were selected at random from the 5 kinds of events of CEC corpus, after the preprocessing such as sentence partition, part of speech tagging and event relation extraction. The numbers of event items of each document set, relationship events and time events statistical information as shown in Table 1 as shown in.

Table 1: The statistical data of CEC corpus

Corpus	*Document(N)*	*Events(N)*	*Sentences(N)*	*Relationships events(N)*
Traffic accident	54	837	265	1614
Earthquake	45	704	292	1208
Fire disaster	31	531	260	962
Bromatoxism	43	191	288	322
Terrorist attack	30	490	249	880
Total	203	2753	1354	4986

At present, the evaluation method for automatic summarization usually adopts the intrinsic evaluation and extrinsic evaluation. Two evaluation methods have their own advantages and disadvantages, the intrinsic evaluation method is simple, easy to handle, but subjectivity too strongwhile extrinsic evaluation method is more objective, proper to evaluate more than one summary system in large scale, but the large consumption of resources, and the evaluation has certain limitations. The essence of automatic summary was information extraction and compression, three evaluating indicators such as the recall rate of R (Recall), the accuracy of P (Precision) and the harmonic mean F (F-Measure) were primarily used to evaluate automatic summary system intrinsically.

4.2 Performance analysis of the experimental data and evaluation

In order to verify the validity of our method for the text automatic summarization, other recent text automatic summarization research methods selected at home and abroad to be compared and evaluated with our methods by experiments.

Method 1 [16]: first, select the sentence as basic extraction unit for event, use binary classifier to distinguish event sentences and non-event sentences; then cluster event sentences, obtained

different sets of events from the same subject document sets, complete the event extraction, and come into being summary.

Method 2 [17]: first, use the sliding window method to extract keywords, construct the vector space and generate undirected graph, then calculate edge weight based on the vector space model, finally, use the weight model of document sentence similarity matrix to model and calculate for document sentence weight, and then gain the document topic sentence according to the compression ratio, complete the event extraction, and come into being summary.

Method 3 [18]: Extracting local properties of a single sentence and global properties between sentences. The local property can be considered as clusters of significant words within each sentence, while the global property can be though of as relations of all sentences in a document. These two properties are combined for ranking and extracting summary sentences. In this experiment, use this method to generate a summary for each experimental corpus at first, compared with method one, method two and method three, and calculate the value of the above three indicators. The results were as shown in Table 2.

Table 2: Experimental results compared with other methods

Method	*Experimental corpus*	*P*	*R*	*F*
1	Selected 100 articles from 1998 "people's Daily".	0.62	0.51	0.56
2	Select 27 news, 20 literature articles and 20 articles	0.57	0.54	0.55
3	Select 10 articles taken from a corpus, the Ziff-Davis.	0.60	0.44	0.50
This method	Randomly selected 100 articles from the CEC corpus.	0.65	0.58	0.61

The contrast shown in Table 2 from different corpus summary generation under different methods could reveal that the method proposed in this paper has achieved a higher recall rate, accuracy of Precision and harmonic mean than other categories (and probably most of the experimental corpus is limited to the emergency categories).

5 Conclusions and Future Works

This paper aims at narrative documents for emergency containing a large number of events. According to the mutual relationship between events, event relationship digraph network is established. Event relationship network with digraph could not only comprehend the development trend of events clearly, but also could show the numbers of events related to other events corresponding to the specified node according to the numbers of nodes correlation of digraph network, nodes relative importance degree was calculated by classical PAGERANK algorithm to represent the importance degree of each sub event in events, and sort for the event of development based on importance ranking of events. Finally, gradually remove information contribution minimum sentences until the sum of the rest sentence length reaches target abstract length. The experiment showed that this method can better generalize the main contents of the text.

There are still some works to be solved in future, such as without considering the distinction importance degree between event node chain out and chain into, only considering the number of the node chain into (out). And automatically constructing event ontology and formally representing event ontology and so on; these problems require further study in the future to be resolved.

References

[1] X. Hu, Summary of Automatic Text Summarization Techniques, Journal of intelligence, vol. 29, pp. 144-147, 2010.

[2] C. Y. Lin, Training a Selection Function for Extraction, the Eighth ACM Conference, In Proceedings of the eighth international conference on Information and knowledge management, pp. 55-62, 1999.

[3] C. J. Jiang, H. Peng, Q. L. Ma, etal.Automatic Summarization for Chinese Text Based on Combined Words Recognition and Paragraph Clustering, Proceedings of the 2010 Third International Symposium on Intelligent Information Technology and Security Informatics (IITSI '10), pp. 591-594, 2010.

[4] P. Y. Zhang, C. H. Li, Automatic text summarization based on sentences clustering and extraction, Proceedings of the 2009 2nd International Conference on Computer Science and Information Technology (ICCSIT 2009), pp. 167-170, 2009.

[5] M. Chandra, V. Gupta, S. K. Paul, A Statistical approach for Automatic Text Summarization by Extraction, 2011 International Conference on Communication Systems and Network Technologies, pp. 268-271, 2011.

[6] G. Erkan, DR Radev LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Arti?cial Intelligence Research, pp. 457-479, 2004.

[7] R. A., Radvansky, G. A. Situation models in language comprehension and memory. Psychological bulletin, vol. 123, pp. 162-185, 1998.

[8] ACE (Automatic Contente Extraction) Chinese Annotation Guidelines for Events. National Institute of Standards and Technology. 2005.

[9] M. F. LIU, W. j. LI, Multi-Document Summarization Based on Event Term Semantic Relation Graph Clustering, journal of Chinese information processing, vol. 24, pp. 77-84, 2010.

[10] Y. F. HAN. Web News Multi-document Summarization Based on Event. Extraction journal of Chinese information processing, vol. 26, pp. 58-66, 2012.

[11] Ge. S. S, Z. Zhang, H. He. Weighted Graph Model Based Sentence Clustering and Ranking for Document Summarization 4th International Conference on Interaction Sciences (ICIS), pp. 90-95, 2011.

[12] P. Thwaites. Causal identifiability via Chain Event Graphs, Artificial Intelligence, vol. 195, pp. 291-315. 2013.

[13] Z. M. Zhong, Z. T. Liu, Ranking Events based on Event Relation Graph for a Single Document, Information Technology Journal, vol. 9, pp. 174-178, 2010.

[14] Z. T. Liu, M. L. Huang, W. Zhou, Z. M. Zhong, J. F. Fu, J. F. Shan and H. L. Zhi, Research on event-oriented ontology model, Computer Science, vol. 36, pp. 189-192, 2009.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd, The Pagerank citation ranking: Bringing order to the web, Technical report, Stanford University, 1998.

[16] X. Y. Jing, Automatic Summarization Algorithm Based 0n Kevword Extraction, Computer Engineering, vol. 38, pp. 183-186, 2012.

[17] B. Ge, F. F. Li, F. Li, W. D. Xiao, Subject Sentence Extraction Based on Undirected Graph Construction. Computer Science, vol. 38, pp. 181-185, 2011.

[18] C. Jaruskulchai, C. Kruengkrai Generic, text summarization using local and global properties of sentences, Proceedings of the IEEE/WIC International Conference on Web Intelligence. Piscataway, USA: IEEE Press, pp. 201-206, 2003.