http://ansinet.com/itj

ITJ

ISSN 1812-5638

# INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 9 (1): 174-178, 2010
ISSN 1812-5638


# Ranking Events Based on Event Relation Graph for a Single Document

Zhaoman Zhong and Zongtian Liu
School of Computer Engineering and Science, Shanghai University, Shanghai, 200072, China

**Abstract:** We present a new event ranking algorithm that applies to a single document. The rules of identifying events are defined first and then the event relation graph is constructed to represent the single document. Based on the adjacency matrix of the event relation graph, the method of ranking events for a single document is elaborated. Present experimental results show that the algorithm of ranking events can achieve satisfied ranking results with an average Recall of 74.5% and an average Kendall' T of 0.374 for 40 documents.

**Key words:** Event, event relation, event relation graph, event ranking, link analysis

## INTRODUCTION

In recent years, the notion of an event has been applied to many research areas such as computational linguistics, artificial intelligence, information retrieval, information extraction, automatic summarization, natural language processing and so on. Pustejovsky (2000) proposed the definition of an event centering on verbs and their attributes from semantics and he believe that the basic semantic unit of an event is a verb. TDT (Topic Detection and Tracking) evaluating conference sponsored by DARPA defines an event as a 'narrowly defined topic for search'. Some researchers applied clustering algorithms to detect events (Yang *et al*., 1998). Li *et al*. (2005) employed a probabilistic model for retrospective news event detection. Information extraction is another hot topic and event extraction is one of three main tasks. Qindong *et al*. (2009) proposed the algorithm of hot topic detection based on feature association. The evaluation conference of ACE (Automatic Content Extraction) defines an event as a special process including participants, which is usually described as the change of a status. The evaluation conference of MUC before ACE has the task of scenarios template and it is focused on event extraction.

The basic work for further processing text is to compute event weight and to rank events from a document in some domains. For example, some important events can be extracted as the answers for question answering systems; the system can choose the sentences including important events as the unit for automatic text summarization; the system can select some important events as the candidate objects of query expansion in information retrieval; and the important events can be recommended to construct the ontology, providing the referenced answers for evaluating the ontology, etc.

The study of ranking events from documents is mainly present in automatic summarization. Zhou (2007) proposed the method of automatic summarization based on event and FCA, which computed the importance of events using the algorithm of PageRank, but did not consider the weight of event relations. Li *et al*. (2006) and Vanderwende *et al*. (2004) believed that the event relation is connected by event elements. If different events contain the same element, these different events have associative relations. They applied this kind of event relations to construct event map for a document and computed event importance using PageRank ranking algorithm. There are two problems about the existing methods computing event importance. First, it is a hard work to extract elements for every event and the result of extracting event elements is non-ideal so far. Second, the associative strength of events is different and it is not accurate to depict event association only judging whether two events have relation.

We believe that the event relation is related by many elements and different elements associate events in different forms. The relation of events associated by the action element is cause-effect relation in many cases. For example, after the earthquake happened, people soon reflect the events such as death, succor, rebuild and so on; and hearing the topic elect, the events such as lecture, debate and vote will call to mind. The relations of events associated by other elements such as locations, objects and time, are identity relation, inclusion relation or subject-object relation. For a single document, the event association in the same sentence is the strongest, the event association in different sentence of the same paragraph is stronger and the event association in different paragraph of the same document is weaker.

**Corresponding Author:** Zhaoman Zhong, School of Computer Engineering and Science, Shanghai University, Shanghai, 200072, China

## EVENTS AND THEIR ASSOCIATIVE STRENGTH

### Event identifying

**Definition 1:** An event refers to one thing happening at a specific time and location, involving a number of actors and showing action characteristics (Zhong *et al*., 2009). We use e to formally denote an event, which can be defined as a five-tuple: e = <A, S, O, T, L>.

The elements of an event five-tuple are called event elements, which represent actions (A), subjects (S), objects (O), time (T) and locations (L), respectively. The action element, called the event trigger, is the key element, it is described by verbs or gerunds and it is the basis for identifying and extracting events.

For different application domains, the number of extracting event elements is different. Judging whether a document include an event or not, we only need to judge whether the action element of an event exists in a document. For event-based text summarization, we need to extract event elements as much as possible in order to full describe an event.

The granularity of events depends on different areas. We identify events using the notion of atomic events introduced by Filatova and Hatzivassiloglou (2003) and namely, the event can be a phrase, clause and sentence. If there are event triggers in the text, we think that the document contains this kind of events.

After using Standford POS Tagger to tag text, which using Penn Treebank Tag Set, the steps of identifying event trigger are as follows:

- **Step 1:** Extracting words tagged by VB, VBD, VBG, VBN, VBP and VBZ as the candidate set E of event triggers
- **Step 2:** Deleting link verbs, modal verbs and auxiliary verbs from set E
- **Step 3:** Deleting the kind of verbs used to express people's perception or message's source from set E such as say, think, report, know, announce, wish, point out, etc. We have collected 48 words belonging to this kind of verbs to create list 1
- **Step 4:** Deleting the kind of verbs which are not notional from set E such as has, last, continue, change, show, come, go, include, etc. 73 words have been collected to construct list 2
- **Step 5:** Deleting the kind of verbs which are not notional and often appear with other verbs or nouns from set E such as generate, cause, occur, try, lead, claim, etc. In this situation, other verbs or nouns are event triggers and they should be added to set E. 51 words have been collected to create list 2

We create three kinds of Lists of special verbs in the process of identifying event triggers. This is similar to the collection of stop words.

**Event associative strength:** The associative strength between events is different. For example, hearing the topic elect, the events such as lecture, debate and vote will call to mind. But a few people will think of events such as operation and succor. The associative strength of event $e_i$ and event $e_j$ is denoted by $\lambda_i$, $0 \le \lambda_{i,j} \le 1$.

The association of events existing in the same sentence is the strongest, the association of events existing in different sentence of the same paragraph is stronger and the association of events existing in different paragraph of the same document is weaker. We can assign different weight for the above three kinds of event relations in order to embody different associative strength.

In this study, we choose reasonable weight through experiments for three kinds of event relations. The rules of quantifying associative strength are as follows:

- **Rule 1:** If event $e_i$ and event $e_j$ exist in the same sentence, $\lambda_{i,j} = 0.8$
- **Rule 2:** If event $e_i$ and event $e_j$ exist in different sentence of the same paragraph, $\lambda_{i,j} = 0.5$
- **Rule 3:** If event $e_i$ and event $e_j$ exist in different paragraph of the same document, $\lambda_{i,j} = 0.2$

If the relations of two events accord with two or three rules at the same time, we take the max value as their associative strength.

## EVENT RELATION GRAPH AND AN EXAMPLE

### Event relation graph

**Definition 2:** An event relation graph can be defined as a pair (V, R), where V is a set of events and R is a set of event relations. The event relation graph describes the events and their relations of a document

**Definition 3:** The matrix $W = (\lambda_{ij})$ is called the adjacency matrix of the event relation graph, where n is the number of events in a document, $1 \le i, j \le n$ and $\lambda_{ij}$ is the associative strength between $e_i$ and $e_j$

**An example of the event relation graph:** Table 1 lists the segment of a document downloaded from the Internet. The left column of Table 1 is original text and the right of it is tagged text. The segment includes two paragraphs. Each paragraph all contains three sentences.

According to the rules of identifying events introduced in section 1.1, 9 event triggers are identified

Table 1: The segment of a document

| Original text | Tagged text |
|---|---|
| According to Chinese state officials, the quake caused 69,181 known deaths including 68,636 in Sichuan province; 18,498 people missed, and 374,171 injured. But these figures may further increase as more reports come in. This estimate includes 158 earthquake relief workers who were lled in landslides as they tried to repair roads. | According/IN to/TO Chinese/JJ state/NN officials,/NN the/DT quake/NN caused/VBD 69,181/CD known/VBN deaths/NNS including/VBG 68,636/CD in/IN Sichuan/NNP province;/NNP 18,498/CD people/NNS missed,/VBN and/CC 374,171/CD injured./VBN But/CC these/DT figures/NNS may/MD further/RB increase/VB as/IN more/JJR reports/NNS come/VBP in./RB This/DT estimate/NN includes/VBZ 158/CD earthquake/NN relief/NN workers/NNS who/WP were/VBD killed/VBN in/IN landslides/NNS as/IN they/PRP tried/VBD to/TO repair/VB roads./NNP |
| One rescue team reported only 2,300 survived from Yingxiu, out of a total population of about 9,000. 3,000 to 5,000 people were killed in Beichuan county, Sichuan province alone; in the same location, 10,000 people were injured and 80% of the buildings were destroyed. Eight schools were toppled in Dujiangyan. | One/CD rescue/NN team/NN reported/VBD only/RB 2,300/CD survived/VBD from/IN Yingxiu,/NNP out/IN of/IN a/DT total/JJ population/NN of/IN about/IN 9,000./CD 3,000/CD to/TO 5,000/CD people/NNS were/VBD killed/VBN in/IN Beichuan/NNP county,/NNP Sichuan/NNP province/NN alone;/NN in/IN the/DT same/JJ location,/NN 10,000/CD people/NNS were/VBD injured/VBN and/CC 80%/CD of/IN the/DT buildings/NNS were/VBD destroyed./VBN Eight/CD schools/NNS were/VBD toppled/VBN in/IN Dujiangyan./NNP |

$$W = \begin{array}{c} \begin{matrix} \text{Death} & \text{Missing} & \text{Injures} & \text{Increase} & \text{Killed} & \text{Repair} & \text{Survived} & \text{Destroyed} & \text{Toppled} \end{matrix} \\ \begin{bmatrix} 0.0 & 0.8 & 0.8 & 0.5 & 0.5 & 0.5 & 0.2 & 0.2 & 0.2 \\ 0.8 & 0.0 & 0.8 & 0.5 & 0.5 & 0.5 & 0.2 & 0.2 & 0.2 \\ 0.5 & 0.8 & 0.0 & 0.5 & 0.8 & 0.5 & 0.5 & 0.8 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.0 & 0.5 & 0.5 & 0.2 & 0.2 & 0.2 \\ 0.5 & 0.5 & 0.8 & 0.5 & 0.0 & 0.8 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 0.8 & 0.0 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.5 & 0.2 & 0.5 & 0.2 & 0.0 & 0.5 & 0.5 \\ 0.2 & 0.2 & 0.8 & 0.2 & 0.5 & 0.2 & 0.5 & 0.0 & 0.5 \\ 0.2 & 0.2 & 0.5 & 0.2 & 0.5 & 0.2 & 0.5 & 0.5 & 0.0 \end{bmatrix} \end{array}$$

Fig. 1: The adjacency matrix of event relation graph

from Table 1. The bold words are event triggers in Table 1. Based on the rules of assigning weight elaborated in section 1.2, the adjacency matrix of the event relation graph is shown in Fig. 1. The self associative strength of events in W is zero.

## RANKING EVENTS BASED ON EVENT RELATION GRAPH

PageRank is a link analysis algorithm, originally formulated by Sergey Brin and Larry Page, used by the Google Internet search engine. It assigns a numerical weighting to each element of a hyperlinked set of documents with the purpose of measuring its importance within the set. The algorithm may be applied to any collection of objects with reciprocal quotations and references.

For an event relation graph, if event $e_i$ relates to more events and the associative strength between them is stronger, event $e_i$ is more important. And meanwhile, the events relating to important events are more important. Accordingly, we can apply the algorithm of PageRank to compute event weight.

For the event relation graph, the iterative equation of R ($e_i$) is given below:

$$R(e_i)_k = d \sum_{j \in ES(e_i)} R(e_j)_{k-1} \times \lambda_{i,j} + \frac{1-d}{n} \quad (1)$$

Table 2: Ranking events based on the adjacency matrix W

| Times of iteration | Event weight | | | |
|---|---|---|---|---|
| 41 | (1) injured | 0.199428 | (2) killed | 0.156357 |
| | (3) death | 0.114539 | (4) missing | 0.114539 |
| | (5) repair | 0.098783 | (6) destroyed | 0.085581 |
| | (7) increase | 0.084820 | (8) survived | 0.070793 |
| | (9) toppled | 0.070793 | | |

$\forall e_i \in E, R(e_i)_0 = 1/|E|$; //Initializing weight of each event

//E is the set of event triggers

While $\exists e_i \in E, (|R(e_i)_k - R(e_i)_{k-1}| > \varepsilon)$ //$\varepsilon$ is the precision of convergence

{

for each $e_i \in E$:

$$R(e_i)_k = R(e_i)_{k-1} + d \times \sum_{j \in ES(e_i)} R(e_j)_{k-1} \times \lambda_{i,j} + \frac{1-d}{n};$$

$s = 1/\sum_{e_i \in E} R(e_i)_k$; // standardizing factor

for each $e_i \in E$:

$R(e_i)_k = s \times R(e_i)_k$; //standardizing

}

Fig. 2: The algorithm of ranking events

where, R ($e_i$)k denotes the kth iteration of R ($e_i$). d is a damping coefficient, ranging from 0-1, usually d = 0.85. ES ($e_i$) is the event set in which each event has relation with $e_i$. N is the number of events in event relation graph. The algorithm of employing Eq. 1 to rank events is shown in Fig. 2.

According to the adjacency matrix W, we use the algorithm of PageRank to rank events. The precision of iterative convergence is 1E-9 and d = 0.85. The results are shown in Table 2. We save six figures after the decimal point for outputting the event weight.

## EXPERIMENTS AND EVALUATION

The goals of the experimental study are: (1) to compare the ranking effectiveness and (2) to compare the convergent performance.

We collect 40 documents including four kinds of topics: Wenchuan eqrthquake, Obama election, H1N1 and Pakistan Taliban Terrorist.

We apply the pooling technology to determine the important events and their orders in a document. The detailed steps are:

- **Step 1:** Five undergraduates read 40 documents and subjectively list the top-5 events for each document
- **Step 2:** Using 25 events to get the union set as the pool for each document
- **Step 3:** Three graduates are responsible for selecting 5 events from the pool as the standard answers set of important events (denoted by ES) for each document
- **Step 4:** Three graduates discuss and determine the order of 5 events for each document.

**The comparisons of ranking effectiveness:** We analyze the results using Recall and Kendall' T. Recall is given by:

$$R(\text{recall}) = \frac{|CS \cap ES|}{|ES|} \quad (2)$$

where, CS is the event set obtained by using the proposed algorithm, which is ranked by descending order according to the event weight and ES is the event set given by the pooling technology.

Kendall' T is defined as follows:

$$T = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \quad (3)$$

where, $n_c$ is the number of concordant pairs and $n_d$ is the number of discordant pairs in the event set.

The Kendall' T has the following properties: (1) If the agreement between the two rankings is perfect (i.e., the two rankings are the same), the coefficient has value 1. (2) If the disagreement between the two rankings is perfect (i.e., one ranking is the reverse of the other), the coefficient has value -1. (3) For all other arrangements the value lies between -1 and 1 and increasing values imply increasing agreement between the rankings. If the rankings are completely independent, the coefficient has value 0 on average.

We employ the algorithm in Fig. 2 to rank events based on the 40 adjacency matrixes of event relation graphs. Set d = 0.85. Table 3 shows the experimental results.

From Table 3, the average Recall of 40 documents is 74.5% and the average Kendall' T is 0.374. The result of ranking events by proposed method is greatly consistent with human evaluating.

**The comparisons of convergence:** We employ the algorithm in Fig. 2 to rank events based on the 40 adjacency matrixes of event relation graphs in different precisions (1E3~1E9), the iterative times of the algorithms are shown in Table 4.

Table 3: The comparison of ranking effectiveness

| Comparison | PageRank |
|---|---|
| Average Recall | 74.5% |
| Average Kendall'T | 0.374 |

Table 4: Average times of iteration of different precisions for 40 documents

| | 1E-3 | 1E-4 | 1E-5 | 1E-6 | 1E-7 | 1E-8 | 1E-9 |
|---|---|---|---|---|---|---|---|
| PageRank | 21 | 23 | 27 | 29 | 34 | 37 | 41 |

From Table 4, the iterative times increase with the improvement of precisions, but the iterative times of the algorithm proposed can be acceptable.

We further construct 40 event relation graphs for 40 documents manually according to whether events have the same elements, but do not consider the position that events exist. Using PageRank algorithm to rank events, the average Recall is 73.5% and the average Kendall'T is 0.375. The result obtained by our method is greatly close to the performance of the latter. But our method is automatic and it is more practical than the latter.

## CONCLUSION AND FUTURE WORK

We take events as the basic semantic unit for texts to study the method of identifying events and ranking event for a single document. The key technique is based on the analysis of event relations to construct event relation graph as the representation model for a single document, further applying PageRank algorithm to compute event weight.

The further research will focus on consummating three kinds of lists of special verbs and how to rank events for a collection of documents. Based on the important events, we will study the method of selecting sentences to generate automatic summarization for a single document or multi-documents.

## ACKNOWLEDGMENTS

This research is supported by the National Natural Science Foundation of China (No.60975033), the Shanghai Leading Academic Discipline Project (J50103) and the Selecting and Training Excellent Young Teacher Foundation of Shanghai (shu-07027).

## REFERENCES

Filatova, E. and V. Hatzivassiloglou, 2003. Domain-independent detection, extraction and labeling of atomic events. Proceedings of Recent Advances in Natural Language Processing Conference, (RANLP'03), Borovetz, Bulgaria, pp: 145-152.

Li, Z., B. Wang, M. Li and W. Ma, 2005. A probabilistic model for retrospective news event detection. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 15-19, ACM, New York, USA., pp: 106-113.

Li, W.J., M.L. Wu, Q. Lu, W. Xu and C.F. Yuan, 2006. Extractive summarization using inter-and intra-event relevance. Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics, Jul. 17-18, Association for Computational Linguistics, Morristown, New Jersey, USA., pp: 369-376.

Pustejovsky, J., 2000. Events and the Semantics of Opposition, in Events as Grammatical Objects. CSLI Publications, Stanford, pp: 445–482.

Qindong, S., W. Qian and Q. Hongli, 2009. The algorithm of short message hot topic detection based on feature association. Inform. Technol. J., 8: 236-240.

Vanderwende L., M. Banko and A. Menezes, 2004. Event-centric summary generation. Proceedings of the Document Understanding Conference, (DUC'04), Boston, MA., pp: 127-132.

Yang, Y., T. Pierce and J. Carbonell, 1998. A study of retrospective and on-line event detection. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 24-28, ACM, New York, USA., pp: 28-36.

Zhong, Z.M., Z.T. Liu, W. Zhou and J.F. Fu, 2009. The model of event relation representation. J. Chinese Inform. Proc., 23: 75-79.

Zhou, W., 2007. Several Concept-Based Knowledge Representations and Related Approaches. Shanghai University, Shanghai.