Semantic Fingerprint of Hot Event Acquisition based on Graph Theory

Maoyuan Zhang[1,a] , Qiongyao Meng[2,b],Xiaohang Pan[3,c]

[1,2,3]School of Computer, Central China Normal University, Wuhan, 430079, China

[a]email: zhangmy@mail.ccnu.edu.cn, [b]email:1126186992@qq.com

[c]email: panxiaohang_love@126.com

Keywords: Semantic Fingerprint; Hot Events; Relevancy; Graph-based

Abstract. Now, there are more and more hot event of network. The event itself is complex, and have too much useless information. In this paper, we propose a formal method which uses semantic fingerprint to represent the hot event of network, and it can be very intuitive expression important information of hot events. Based on the graph theory, by the event title, the text sentence as the node, through the interaction between them, get hot events on behalf of the sentence. Then use the TF-IDF method to get the feature word from the sentence expression the semantic fingerprints of hot events. The semantic fingerprint is applied to the retrieval model to prove the validity of the event semantic fingerprint method.

Introduction

With the development of the internet, every event in the world can connect through the network, as the increase of the network information, more and more hot events attract people's attention, and the event include too much useless information. So it is an important question that how to expression a hot event to make people understand the important information of a hot event directly through the representation of the event.

The concept of semantic fingerprint [1] is proposed based on the Chinese semantic knowledge acquisition of the Internet encyclopedia. Semantic fingerprints including the semantic tags' semantic description and the semantic tags refer to the concept. It is depict by semantic tags' related word group and every word's degree of contribution to the semantic fingerprints. The degree of contribution is the relevancy degree between the word and the semantic tags in the semantic fingerprints. Similar, if we use the event represent the semantic tags and the event's feature word represent the semantic fingerprints, we can found a relation between the event and the event's feature word. So we can understand the event's importance information through the event's semantic fingerprints.

The event's semantic fingerprints

American psychologist Harry Lorayne's [4] book said that: "If you want to keep in mind that any new information, you have to make this information and relating it to something you already know or remember. Contact, closely related to memory, means to bundle two or more things together, or make the connection between them." In the same way, when people saw a new event, often establishing a connection between the existing information and the new event. Then to determine whether they have interest to the new event. For example, people have interest for the word like university students, roommate and poisoned when the see the hot events of Fudan University poisoning case. So, if we build a link between the word and the event through the semantic fingerprints, this word can expression the event. And we will know the important information of the event through the word.

So, the paper aims to propose a method to expression the event's semantic fingerprints. The method make up by event and event's related words. The event represent the semantic tags and the event's related words represent the semantic fingerprint. The relevancy degree between the related word and the event represent the semantic fingerprint's contributive degree. The form of the semantic fingerprints like this:

Event: Semantic fingerprints (contributive degree)

The semantic fingerprints represent the hot events' related word and the contributive degree (between 0 and 1) represent the related word's contributive degree.

Using event title in graph-based approach for semantic fingerprints

The paper put forward using event title in graph-based approach for semantic fingerprints of events. First, use the sentence of the event and the event's title as the node of graph, the similarity between the sentence and the title and the similarity between the sentences as the side of graph. Then sort the nodes after get the node's weight through the interaction of the graph's node. And choose the top M sentences as the key sentences. The last, use the TF-IDF method get the semantic fingerprint of event from the key sentences. The graph's construction as follow:

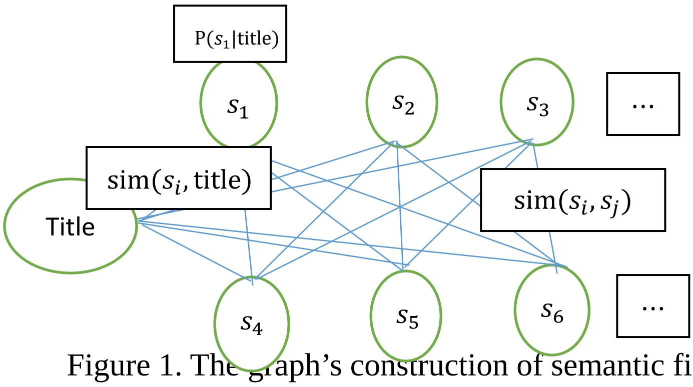

Figure 1. The graph's construction of semantic fingerprint

In the figure above, s_i(i=1,2,3…n,n is the number of the sentence) represent the sentences of event. The symbol "title" represent the title of the event. The symbol "sim(s_i,title)" represent the similarity between title and sentence. The symbol "sim(s_i, s_j)" represent the similarity between sentence and sentence. The symbol "P(s_i|title)" represent the weight of the sentence. We will choose the top M sentences as the key sentences after sort the sentences' weight.

Acquisition of characteristic sentences

The first step to obtain the semantic fingerprints of events is get the feature sentences of event. The method is use the influence of the title sentence to the text sentence and the text sentence to the text sentence to get more important sentence. The symbol "title" represent the title of event and the symbol "s" represent the sentence. In the condition to give the title, the weight of sentence can calculate use the formula as follow:

$$\text{P(s|title)} = \text{d}\sum_{v\in C}\frac{sim(s,v)}{\sum_{z\in C} sim(z,v)}P(v|title) + (1-d)\frac{sim(s,title)}{\sum_{z\in C} sim(z,title)} \qquad (1)$$

In the formula, the symbol "C" represent the set of event. The symbol "v" represent the sentence of the event set. The symbol "sim()" represent the similarity of the sentences. The symbol "$P(v|title)$" represent the weight of the sentence. The symbol "d" and "1-d" respectively represent the influence of other sentence and the title to a sentence. It value between zero and one. The formula can transformation as follow:

$$\text{P(k + 1)} = M^T P(k) \qquad (2)$$

$$\mathrm{M} = \mathrm{dA} + (1 - \mathrm{d})\mathrm{B} \tag{3}$$

In the formula, the matrix A represent the similarity between the sentences and the matrix B represent the similarity between the title and the sentence. The matrix M represent the total influence from the others. The symbol "$P(k)$" represent the sentence's weight after k times calculate.

The specific steps of the method are as follows:

1. Calculate the value of the matrix A and B. The element $A_{ij} = sim(s_i, s_j)$ in matrix A represent the similarity between the sentences $B_{ij} = sim(s_i, title)$ in matrix B represent the similarity between the sentence and the title.
2. According the experimental results to adjustment the parameter d to calculate the matrix M.
3. The symbol "$P(k)$" represent the weight of the sentence and it initial value set as $\frac{1}{N}$ (N is the number of the sentences). When the value of the formula $\left|\left|\mathrm{P}(\mathrm{k}+1) - \mathrm{P}(\mathrm{k})\right|\right| < \epsilon$(The symbol "$\epsilon$" is the threshold), stop the iterative calculation.
4. Rank the sentences use the weight of the sentences and choose the top M sentences as the key sentences.

Acquisition of semantic fingerprint

This section uses the method of TF-IDF to extract the feature words as the semantic fingerprint of the event from the feature sentences which are obtained in the 3.1 section. TF-IDF is a statistical method used to assess the importance of a word to a document or a corpus of one of the documents. The importance of words is proportional to the number of times it appears in the document, but it is inversely proportional to the frequency of the word in the corpus.

TF representation the frequency of the word, is a word appearing in the document frequency, on behalf of the importance of the word. The importance of representation form as follows.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{4}$$

Among them, $n_{i,j}$ said the ith word in j file. $\sum_k n_{k,j}$ said the number of word in j file.

IDF expressed the inverse document frequency, that is, the number of words in the document appeared. The form as follows.

$$idf_i = log\frac{|D|}{1+|\{j: t_i \in d_j\}|} \tag{5}$$

Among them, $|D|$ represent the event set. $(1 + |\{j: t_i \in d_j\}|)$ said the number of files contain t_i of the word. The front plus 1 is to prevent the denominator is 0.

Finally get a word TF-IDF calculation method is as follows.

$$tfidf_{i,j} = tf_{i,j} * idf_i \tag{6}$$

And, $tfidf_{i,j}$ represent the importance of words.

Section 3.1 using the title event based method to obtain the text feature sentence. In this section, the characteristics of sentences as a document, calculate the characteristic sentence in every word of the TF-IDF values, to sort the words according to word TF IDF values, take the top n words as event semantic fingerprint.

Experiment

In order to verify the validity of the semantic fingerprint acquisition method, we obtain the news events from the Internet through the information collection technology. The topic detection and tracking (TDT) method of the network public opinion monitoring technology is used to get the hot events on the network. We selected some hot event that cased the social's great concern. By using the semantic fingerprint calculation method, we obtain the semantic representation of the event semantic fingerprint as follows:

Semantic tags	Semantic fingerprints（related word and contributive degree）
Shanghai stampede	(stampede:0.96666664), (bund:0.9638889),(Shanghai:0.94722223), (crowded:0.72672457),(injured:0.7182698),(Chen yi square:0.66524255), (death:0.6337654),(accident:0.6018324), (security:0.5532967), (people flow:0.54584),(wounded:0.5197265),(new year:0.51543194), (administration:0.45622352),(Victims:0.42189682),(family member:0.392585), (serious:0.37849545),(Huangpu district:0.35785913),(public:0.35268897)…
MAS MH370 out of contact	(MAS:0.6944444),(MAS flight:0.6792929)(disappear:0.6515151), (out of contact:0.560971),(aircraft:0.55733216),(Malaysia:0.5156701), (Australia:0.439182),(passenger:0.4140948), (search: 0.4073897), (aviation:0.3817094),(Kuala Lumpur: 0.38112032),(radar:0.3765858), (Beijing: 0.37339744),(family member :0.36656612),(satellite:0.3582503), (China:0.3548387),(Indian ocean:0.3450861),(sea area:0.31058174)…
Kun Ming Railway Station violent& terrorist	(Kun Ming Railway Station:0.9680851),(terrorist:0.75551677), (violence:0.7464024),(split:0.54937) (Yunnan:0.5162964), (square:0.48060408),(masses;0.44080225),(thug:0.419114), (attack:0.4118178),(China:0.39888734),(death:0.3436386), (Victims:0.34064114) ,(society:0.3303911),(injured:0.29035342), (killed :0.28605524),(Xinjiang:0.247080), (ticket office: 0.21316703), (unity:0.20978758)…
Occupy Central with Love and Peace, OCLP	(OCLP:0.74814373),(Hong Kong:0.7291667),(NPC:0.66192913), (society:0.6202088),(citizen:0.6139992),(government:0.56099194), (general election;0.54060245),(illegal:0.5211922),(break the law:0.517576), (develop:0.51224303),(assembly:0.5100603),(communication;0.50629), (order:0.49258828),(democracy:0.41593075),(stability:0.41570467), (break:0.40004492), (reason:0.39789867),(basic law:0.38797513)…

Table 1. We obtained semantic fingerprint of the events

Result analysis

According to the method of this paper, we obtain the semantic fingerprint of the event. The semantic fingerprint of events is used to describe an event. From the experimental results, we can see that this method can get a good description of the semantic fingerprint. We put the semantic fingerprint of the event into the retrieval model and retrieve the relevant documents to prove that the semantic fingerprint of the event in this paper can be described as an event.

We choose 5 news events from the news corpus, each event has 120 related documents, a total of 600 news text as a database. We select the first six related words in the event semantic fingerprint as the query word, and use the Lucene search tool to retrieve the relevant documents in the database.

The experimental results are compared with TF-IDF and BM25, and the results are as follows:

	Recall	Precision	F-value
TF-IDF	0.728	0.781	0.754
BM25	0.760	0.831	0.792
Event Semantic Fingerprint	0.846	0.876	0.861

Table 3 Similarity comparison in information retrieval

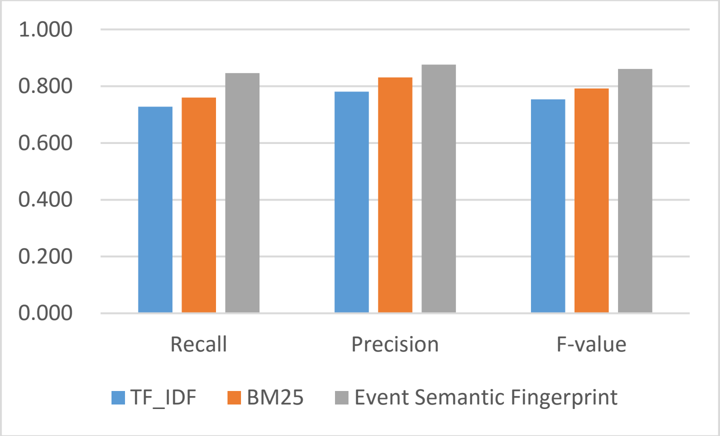

Figure 2. The comparison results using histogram

From the contrast experiment, we can see that the feature of semantic fingerprint of event is more able to express an event. It proves the superiority of the event semantic fingerprint.

Conclusions and Outlook

In this paper, a method of using semantic fingerprint to represent the hot event is proposed. And using based-graph approach to calculation the relevance of the title sentence and the text sentence as well as the text sentence. Get the first m sentence as the characteristic sentence by sort of the weight. Then use the TF-IDF method to get the characteristic word as the semantic fingerprint of the event from the characteristic sentence. The semantic fingerprint is applied to the retrieval model to prove the validity of the event semantic fingerprint method.

In the future, according to the characteristics of the event semantic fingerprint will be applied to the event related to the discovery of hot events.

Acknowledgement

In this paper, the research was sponsored by the Humanity and Social Science Youth Foundation of Ministry of Education of China (Project No. 15YJC870029) and National Social Science Foundation (Project No. 12&2D223).

Reference:

[1] Liu Yang, Tingting He, Xinhui Tu. Based on the network encyclopedia of the Chinese semantic knowledge acquisition [C]. The fifth national youth computational linguistics seminar [c](YWCL2010) ;2010.

[2] Guohu Wang, Haifeng Deng, Yalei Wang. A Study on Public Opinion Relevancy of Network Hot Issues [J]. Journal of Information.2012.07

[3] Lin Zhao, Lide Wu, Xuanjing Huang. Using query expansion in graph-based approach for

query-Focused multi-document summarization [J]. Information Processing and Management.2008.7.

[4] Harry Lorayne. Harvard magic memory I+II+III+IV. Xi’an: Shaanxi normal university press, p19, 2009.8.

[5] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias et al. “Placing Search in Context: The Concept Revisited”. ACM Transactions on Information System, 20(1):116-131, January 2002.

[6] Long C, Huang ML, Zhu XY et al. A new approach for update multi-document summarization [J]. Journal of computer science and technology 25(4):739-749 July 2010.