

Discovering the Core Semantics of Event from Social Media

Weidong Liu^a, Xiangfeng Luo^{a,*}, Zhiguo Gong^b, Junyu Xuan^a, Ngai Meng Kou^b, Zheng Xu^c

^a*Shanghai University, Shanghai, China*

^b*University of Macau, Macau, China*

^c*The Third Research Institute of Ministry of Public Security, Shanghai, China*

Abstract

As social media is opening up such as Twitter and Sina Weibo¹, large volumes of short texts are flooding on the Web. The ocean of short texts dilutes the limited core semantics of event in cyberspace by redundancy, noises and irrelevant content on the web, which make it difficult to discover the core semantics of event. The major challenges include how to efficiently learn the semantic association distribution by small-scale association relations and how to maximize the coverage of the semantic association distribution by the minimum number of redundancy-free short texts. To solve the above issues, we explore a Markov random field based method for discovering the core semantics of event. This method makes semantics collaborative computation for learning association relation distribution and makes information gradient computation for discovering k redundancy-free texts as the core semantics of event. We evaluate our method by comparing with two state-of-the-art methods on the TAC dataset and the microblog dataset. The results show our method outperforms other methods in extracting core semantics accurately and efficiently. The proposed method can be applied to short text automatic generation, event discovery and summarization for big data analysis.

*Corresponding author.

Email addresses: liuwd@shu.edu.cn (Weidong Liu), luoxf@shu.edu.cn (Xiangfeng Luo), fstzgg@umac.mo (Zhiguo Gong), xuanjunyu@shu.edu.cn (Junyu Xuan), yb27406@umac.mo (Ngai Meng Kou), xuzheng@shu.edu.cn (Zheng Xu)

¹Chinese microblogging website <http://weibo.com/>

Keywords: core semantics, semantic link network, information gradient

1. Introduction

With booming social media, the data explosion of microblog on blogosphere accompanies with hot events. For example, a large volume of microblogs discussed about “*USA Boston Marathon bombing*”, “*the US surveillance program*”
5 “*PRISM*” and so on. Given the microblogs/short texts² about a concrete event, information about the event is unevenly distributed on these “event messages” since some ones might contain much more important and diverse information (e.g., different event time, locations, participants, processes, and opinions) than others (e.g., redundancy and noises in short texts). Besides, these short texts
10 are globally semantic redundant and locally semantic sparse since many short texts contain the same content and local ones only focus on one aspect of the whole event. Understanding the event concisely and thoroughly is impossible when these redundant short texts may crowd out other ones which contain important and diverse information. For example, when we query by keyword
15 “*Ebola*”, Sina Weibo returns redundant Chinese microblogs about “*Ebola of salmon*” and Twitter returns redundant tweets about “*A Italian doctor catches Ebola*” as shown in Fig.1, which crowd out many important microblogs which discuss about outbreak, spreading and control of event “*Ebola*”.

Herein, how to automatically discover the core semantics of event from big
20 social media data is a challenging problem, since it is time-consuming and unpractical to manually find out the core semantics of event from big media data.

Existing methods to solve this problem are summarized as follows:

- 1) Feature-based methods. These methods directly use basic statistic technique on features including word frequency, title words, cure words which are con-
25 sidered for selecting sentences as core semantics[1, 2]. Structural features of

²There is a word limitation of microblog. For example, each tweet in twitter can’t exceed 140 words.



Figure 1: Redundant sentences returned by microblogging services when inputting keyword query "Ebola"

discourse are used to identify core semantics by rhetorical structure analysis, pragmatic analysis, lexical chain, latent semantic analysis[3]. Besides, more features are used in some specified semantics discovery methods whose features include hashtags, timestamps and emotion labels[4].

- 2) graph-based methods. These methods construct graph where short texts as nodes and text-pairwise relations as edges[5, 6, 7, 8]. Top k short texts are selected as core semantics by ranking values of graph-based features or values of Markov random walk on the graph[5, 6, 7, 8]. Besides, such methods can be extended into conditional random fields which identify core semantics by labeling sentences, where the sentence label influences the labels of nearby sentences[9].
- 3) clustering-based methods. These methods cluster short texts into different clusters, and then select some short texts from each cluster to represent the semantics of the cluster[10, 11, 12]. The clustering methods include hierarchical clustering, partitional clustering and semantic-based clustering[13]. Besides, some priori knowledge or constraint conditions in specified domain are considered in clustering[14, 15].
- 4) semantic link-based method. Semantic link-based methods have strong abil-

ities in semantics organization, semantic community discovery and emerging
 45 semantics learning/reasoning[16]. Such methods have been used in semantic
 representation[17], semantic organization[18], semantic interaction [19, 20],
 semantic community discovery[21, 22] and semantic linking space for Cyber-
 Physical Society[23, 20].

5) other methods. These methods include Bayesian topic model-based methods[24],
 50 Neural Networks-based methods, Decision tree-based methods and so on[25,
 26].

However, these methods have the following limitations:

- 1) semantic association loss. The graph-based and cluster-based methods often
 use vector space model to represent short texts and use vector-based similar-
 55 ity methods. Obviously, these similarity-based methods lost many semantic
 association relations;
- 2) high computational cost. The time complexity of most the above methods
 [5, 6, 7, 8, 10, 11, 12], which have to compute text-pairwise similarity, is
 $O(n^2)$. It is unpractical when the text number is large in big data;
- 60 3) redundancy-prone results. The above methods pay less attention on the issue
 of redundancy and result in redundant results since these methods assign
 almost the same values to alike short texts.

To solve the above limitations, we propose a Markov random field based
 method for discovering the core semantics of event:

- 65 1) To avoid semantic association loss, our method makes semantic collaborative
 computation to learn the whole association relation distribution of an event
 by small-scale association relations .
- 2) To reduce computation cost, our method makes probabilistic inference in a
 limited keyword association link network, rather than text-pairwise compu-
 70 tation.
- 3) To be free of redundancy, our method proposes information gradient compu-
 tation by maximizing information gradient of k short texts since information
 gradient decreases when redundancy increases.

Compared with existing methods, the contributions of our method are summarized as follows:

- 1) Our method learns association relation distribution by semantic collaborative computation.
- 2) Our method is efficient by probabilistic inference on semantic association link network.
- 3) Our method obtains redundancy-free core semantics by information gradient computation.

The remainder of the paper is organized as follows. In Section 2, we introduce the preliminaries including some basic definitions and problem formal definition. In Section 3, we propose a framework of Markov random field based method for discovering the core semantics of event. We construct a Markov random field by semantic association collaborative computation, which learns association relation distribution by low-degree relations in Section 4. We propose information gradient computation to maximize coverage of association distribution by the minimum number of redundant-free short texts in Section 5. Experimental results are presented in Section 6. We give the conclusion and future work in Section 7.

2. Preliminary knowledge and problem statement

Before discussing our method, we first introduce some basic concepts which are thoroughly used in this paper and then propose the problem statement of this paper.

2.1. Preliminary knowledge

Semantic representation and inference are two major issues for discovering the core semantics of event. We introduce two basic models in semantic representation and probabilistic inferences.

100 Event representation is just as human beings learn concepts from an event
 [17, 27, 28, 29, 30, 31, 32, 33], where each concept consists of association rela-
 tions. Inspired by [17], we adopt association relations to represent an event as
 follows:

Definition 1 (*Event Power Serial Representation, E-PSR*). E-PSR is rep-
 105 resented as,

$$E-PSR = \{\Phi_k | 0 \leq k \leq 2\} \quad (1)$$

$$\Phi_k = \{\phi_{k,i}\} \quad (2)$$

where Φ_k denotes a k-degree association relation set; $\phi_{k,i} = w_0^{(k,i)}, w_1^{(k,i)}, \dots, w_{k-1}^{(k,i)} \rightarrow w_k^{(k,i)}$ is a k-degree association relation with support value which is calculated by,

$$sup(\phi_{k,i}) = \frac{\sum_{s_l \in e} I(w_0^{(k,i)}, w_1^{(k,i)}, \dots, w_k^{(k,i)} | s_l)}{\sum_{s_l \in e} I(s_l | s_l)} \quad (3)$$

110 where $I(A|B)$ is an indicator function whose value is 1 if $A \subseteq B$ and 0 otherwise;
 $e = \{s_l | 1 \leq l \leq n\}$ denotes an event which consists of n sentences.

For example, an event contains two short texts, $e = \{s_1, s_2\}$:

s_1 : *That boy stands on the left, whose t-shirt is red.*

s_2 : *Two girls stand on the right, whose skirts are also red.*

115 E-PSR of the event e includes association relations as follows:

$\Phi_0 = \{red\}; \Phi_1 = \{boy \rightarrow left, girl \rightarrow right, boy \rightarrow t-shirt, girl \rightarrow skirt\}.$

E-PSR simply obtains low-degree association relation distribution $\{\Phi_k | 0 \leq k \leq 2\}$ rather than the whole association relation distribution $\{\Phi_k | 0 \leq k\}$, since how to obtain $\{\Phi_k | k > 2\}$ is still a unsolved problem. As such, we may
 120 ask the following question: Is the above problem solved by inference based on probabilistic model?

Markov random field model(MRF) is an undirected probabilistic graphical model. We propose event Markov random field for semantic representation and inference.

125 **Definition 2 (*Event Markov Random Field, E-MRF*).** E-MRF is represented by,

$$E-MRF = \langle G, P(X) \rangle \quad (4)$$

where $G = \langle X, E \rangle$ is an undirected graph, where X denotes a set of random variables and E denotes a set of dependence relations between X ; P denotes a joint probability distribution over X , which is calculated by:

$$P(X) = \mu_0 \prod_{c_i \in C} \Psi(X_{c_i}) \quad (5)$$

130 where c_i denotes a maximal clique, which is a full connective sub-graph of G ; $C = \{c_1, c_2, \dots\}$ denotes a maximal clique set; X_{c_i} is a maximal clique variable of c_i ; $\Psi(X_{c_i})$ is a non-negative potential function over X_{c_i} ; $\mu_0 = \frac{1}{\sum_x \prod_{c_i \in C} \Psi(x_{c_i})}$ is a normalization factor.

Besides, the E-MRF has strong scalability to satisfy different applications
135 since the potential functions $\Psi(x_{c_i})$ of E-MRF can be defined flexibly to generate different $P(X)$.

If $P(X)$ is known, the support value of a k -degree association relation $\phi_{k,i}$ ($k > 2$) is inferred by,

$$P(\phi_{k,i}) = \sum_x P(x) I(\phi_{k,i} | x) \quad (6)$$

where $I(\phi_{k,i} | x)$ is an indicator function whose value is 1 if x is consistent
140 with $\phi_{k,i}$ and 0 otherwise; x is consistent with $\phi_{k,i}$ if $x_{(w_{w_t}^{(k,i)})} = 1$, $(w_{w_t}^{(k,i)} \in \phi_{k,i}, x_{(w_{w_t}^{(k,i)})} \in x)$.

2.2. Problem statement

In this paper, our task is to learn association relation distribution of an event
and to cover the distribution by k sentences as the core semantics of event. The
145 k sentences which can cover the association relation distribution of an event is the core semantics of event.

Supposing an event $e = \{s_l | 1 \leq l \leq n\}$, consists of n sentences, such as microblogs, tweets or comments, the k sentences discovered from e should satisfy the following properties:

- 150 1) Priority for sentence with frequently discussed content, since such sentences include more core association relations.
- 2) Priority for sentence with new content, since such sentences provide user with more information about the event.

Definition 3 (*information gradient of k -sentences, $IG(S_k)$).* $IG(S_k)$ reflects how frequent and how novel the content of k sentences is in an event, which
155 is calculated by,

$$IG(S_k) = IG(S_{k-1}) + IG(s_{(k)}|S_{k-1}) \quad (7)$$

where $S_k = \{s_{(l)}|1 \leq l \leq k\}$ denotes a set of k sentences; $s_{(l)}$ denotes the l^{th} sentence; $IG(s_{(k)}|S_{k-1})$ denotes conditional information gradient of the sentence $s_{(k)}$ conditioned on S_{k-1} .

160 $IG(s_{(k)}|S_{k-1})$ is approximated 0 when $s_{(k)}$ provides no novel or indifferent content conditioned on S_{k-1} .

$$IG(s_{(k)}|S_{k-1}) \approx 0 \text{ if } IG(S_k) - IG(S_{k-1}) \approx 0 \quad (8)$$

To satisfy above properties, the core semantic of event is defined as,

Definition 4 (*Core Semantics of Event, $CS(e)$).* the core semantic of event is obtained by maximizing information gradient of k sentences,

$$CS(e) = \arg \max_{|S_k|=k} IG(S_k) \quad (9)$$

165 We list the notations of the above definitions in table 1, which are thoroughly used in this paper.

3. Proposed method

To solve the problem defined in equation 9, a framework of Markov random field based method for discovering core semantics is shown in Fig.2. In the
170 following, we introduce the whole framework by 4 steps. 1st and 2nd steps mainly obtain short texts for each event. The most adopted algorithms to obtain

Table 1: Notations used in the paper

Symbols	Description
$\phi_{k,i}$	$\phi_{k,i} = w_0^{(k,i)}, w_1^{(k,i)}, \dots, w_{k-1}^{(k,i)} \rightarrow w_k^{(k,i)}$ is a k-degree association relation
Φ_k	$\Phi_k = \{\phi_{k,i} i < \Phi_k \}$ denotes a k-degree association relation set
$E-PSR$	$E-PSR = \{\Phi_k 0 \leq k \leq 2\}$ is a set of association relations set
X	$X = \{X_{w_i} 0 \leq i \leq X \}$ denotes a random variable set
$G = \langle X, E \rangle$	$G = \langle X, E \rangle$ is an undirected graph, where E denotes dependence of X
c_i	c_i is a maximal clique which is a full connective sub-graph of G
C	$C = c_i$ is a set of maximal cliques
X_{c_i}	a maximal clique random variable of c_i
$\Psi(X_{c_i})$	a non-negative potential function over X_{c_i}
μ_i	a parameters of Markov random field
$P(X)$	joint probability of X
\hat{X}	\hat{X} is a sub-set of variables X
$P(\hat{X})$	marginal probability of \hat{X} compared with joint probability $P(X)$
$E-MRF$	$E-MRF = \langle G, P(X) \rangle$ denotes an event Markov random model
$IG(S_k)$	information gradient of $S_k = \{s_{(l)} 1 \leq l < k\}$
$IG(s_{(k)} S_{k-1})$	conditional information gradient of a sentence $s_{(k)}$ conditioned on S_{k-1}

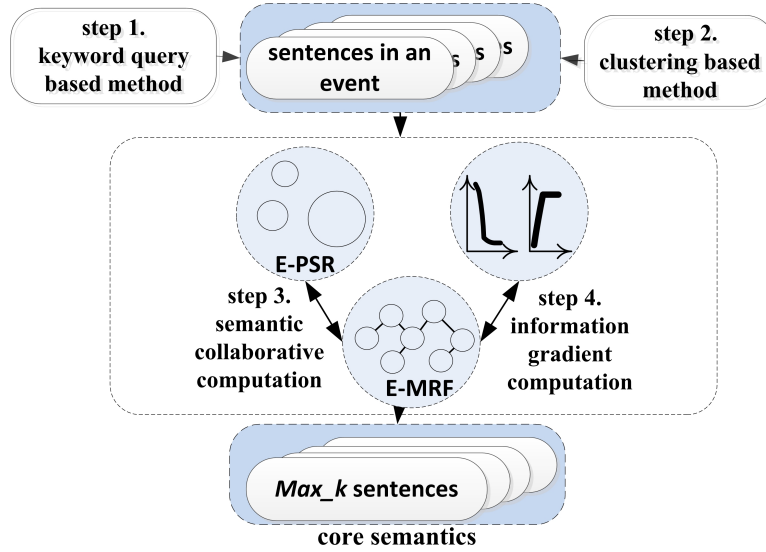


Figure 2: A framework of Markov random field based method for discovering the core semantics of event

these short texts are: 1) query based algorithms, users query some keywords to get related sentences of an event; 2) event detection or clustering based algorithms, such algorithms discover sentences which belong to different events [34, 35, 36]. To ensure the data is valid, we apply the state-of-the-art event discovery methods to reduce the possible negative effects on core semantics discovery [37, 38, 39, 40]. Besides, association relation based representation in our model can further reduces these adverse effects caused by noise and irrelevant short texts. 1st and 2nd steps are not the focus of this paper.

We mainly focus on 3th-4th steps for discovering core semantics of an event.

Obtaining semantic association relation distribution is a basic issue for core semantics discovery. Our Markov random field based method collaborates with power serial representation model to learn the association relation distribution of an event. 3th step gives semantic collaborative computation between E-MRF and E-PSR, which mutually benefits each other: E-PSR reduces computational cost of construction of E-MRF by low-degree association relations $\{\Phi_k | 0 \leq k \leq 2\}$; E-MRF extends E-PSR with inference ability on high-degree association relation $\{\Phi_k | k > 2\}$.

Although, the above method has built a probabilistic graphic model which has strong abilities in semantic representation and inference, it assigns alike sentences similar probability values no matter how redundant they are. Obviously, such method is redundancy-prone.

To obtain redundancy-free results, 4th step makes information gradient computation, which assigns redundant sentences with lower conditional information gradient by equation 8 since these redundant sentences contain little novel content. If k sentences can maximize information gradient defined in equation 9, the k sentences are the core semantics of event which maximally covers the association relation distribution.

4. Semantic collaborative computation

200 To learn an E-MRF model and make semantic inference thereafter, we col-
laborates E-MRF with E-PSR because: 1) semantic association consistence. E-
MRF keeps association consistence with E-PSR by association relations which
cover almost 90% semantic association [17]; 2) learning efficiency improvement.
E-PSR reduces the graphic scale of E-MRF with smaller and sparser associa-
205 tion relations $\{\Phi_k | 0 \leq k \leq 2\}$ which cause efficient probabilistic inference on
E-MRF.

4.1. Collaborative computation of E-MRF with E-PSR

E-MRF is an undirected probabilistic graphic model and E-PSR is an asso-
ciation relation set. Structural consistence and association consistence between
210 E-MRF and E-PSR are critical issues for learning association relation distribu-
tion. To guarantee these consistences, we collaborate E-MRF with E-PSR by
structural collaborative computation and potential value collaborative compu-
tation.

4.1.1. Structural collaborative computation

215 Structural collaborative computation mainly solves issues: 1) how dose E-
PSR form graphic structure of E-MRF; 2) what's the parameter structure
of E-MRF formed by E-PSR. Referred to a factor graph theory [41], given
E-PSR= $\{\Phi_k | 0 \leq k \leq 2\}$, we solve the above issues by 1st-4th steps. For 1),
we form an undirected graph structure by 1st-2nd steps; For 2), we form the
220 parameter structure by 3th-4th steps.

step 1. add a variable X_{w_i} in E-MRF, only if keywords w_i currents in E-PSR;
step 2. link X_{w_i} and X_{w_j} in E-MRF, only if w_i and w_j co-occur in an association
relations in E-PSR;

step 3. map each association relation into a maximal clique, only if the keywords
225 in association relations are contained by the maximal clique;

step 4. set the parameter structure for each maximal clique variable by,

$$\Psi_{X_{c_i}} = \prod_{\phi_i \in \{\Phi_k | 0 \leq k \leq 2\}} \mu_i^{I(X_{c_i} | \phi_i)} \quad (10)$$

where X_{c_i} denotes a maximal clique variable; μ_i is a parameter of an association relation ϕ_i in E-PSR; $I(X_{c_i} | \phi_i)$ is an indicator function whose outcome is 1 when association relation ϕ_i is consistent with the maximal clique variable X_{c_i} ,
 230 0 otherwise.

According to 1st-4th steps, we propose a structural collaborative algorithm
 1.

Algorithm 1 structural collaborative algorithm

Input: an event, $e = \{s_l | 1 \leq l < n\}$

Output: a graphic structure $G = \langle X, E \rangle$ and a parameter structure $\{\Psi_{X_{c_i}}\}$ of E-MRF

1. mine E-PSR by algorithm [17]
 2. form $G = \langle X, E \rangle$ by 1st-2nd steps
 3. form $\{\Psi_{X_{c_i}}\}$ by 3th-4th steps
 4. **return** $G = \langle X, E \rangle$ **and** $\{\Psi_{X_{c_i}}\}$
-

4.1.2. Potential value collaborative computation

To guarantee semantic association consistence between E-PSR and E-MRF,
 235 we propose potential value collaborative computation which learns potential values of E-MRF from E-PSR. It is easily found that $sup(\phi_{k,i})$ defined in equation 3 is unbiased estimation of marginal probability $P(\phi_{k,i})$ defined in equation 6 by,

$$|\sum_x P(x) I(\phi_{k,i} | x) - sup(\phi_{k,i})| < \varepsilon \quad (11)$$

where $sup(\phi_{k,i})$ denotes the support value of $\phi_{k,i}$.

240 The parameter μ_i defined in equation 10 should satisfies equation 11. Using local item sets to construct a MRF model is first proposed by Pavlov [42],

we adopted an iterative scaling (IS) algorithm [43, 44, 45] to learn parameters $\{\mu_i | 0 \leq i \leq |E-PSR|\}$ by,

$$\mu_0^{t+1} = \mu_0^t \times \frac{1 - \text{sup}(\phi_i)}{1 - P^t(\phi_i)} \quad (12)$$

$$\mu_i^{t+1} = \mu_i^t \times \frac{\text{sup}(\phi_i) \times (1 - P^t(\phi_i))}{P^t(\phi_i) \times (1 - \text{sup}(\phi_i))} \quad (0 < i) \quad (13)$$

Herein, we propose a potential value collaborative algorithm 2.

Algorithm 2 potential value collaborative algorithm

Input: $G = \langle N, E \rangle$ and $\{\Psi_{X_{c_i}}\}$ and E-PSR

Output: parameter values $\{\mu_i | 0 \leq i \leq |E-PSR|\}$ of E-MRF

1. while (Not all support values satisfy equation 11)
 2. for (i over constrains)
 3. update μ_0^{t+1} by equation 12
 4. update μ_i^{t+1} by equation 13
 5. **return** $\{\mu_i | 0 \leq i \leq |E-PSR|\}$
-

245

4.2. Basic semantic computation in E-MRF

Given $E-MRF = \langle G, P(X) \rangle$ learned by algorithm 1-2, we had obtain the joint probability distribution of X since we calculate probability of sub-variables \hat{X} by,

$$P(\hat{X}) = \sum_x P(x) I(\hat{X}|x) \quad (14)$$

250 where \hat{X} is a subset of random variables x ; $I(\hat{X}|x)$ is an indicator function with outcome 1 when \hat{X} is consistent with x , 0 otherwise.

Based on equation 14, we make some basic semantic computations. For example, we can calculate how frequently a sentence is discussed by,

$$P(s_i) = \sum_x P(x) I(s_i|x) \quad (15)$$

255 where $I(s_i|x)$ is an indicator function with outcome 1 when s_i is consistent with x , 0 otherwise.

A common issue among equations 11-15 is to calculate the marginal probability of E-MRF. Junction tree algorithm is a general probabilistic inference framework for calculating joint probability, marginal probability and condition probability [46] by decomposing a global joint probability computation into a linked set of local computations. Referring to [46], we adopt the junction tree algorithm to calculate marginal probability.

If a sentence with higher frequency is selected as core semantics by equation 15, large number of alike sentences will be selected as well since their similar contents and thus result in undesirable redundancy.

5. Information gradient of k-sentences

Redundancy conflicts with cognitive psychology [47], since undue redundancy has limited effect on memory activity. For human beings, repeated memorizing a word or sentence content has limited effect in human memory process as shown in Fig.3. Fig.3(a) shows that memory activity value changes with repetition. The memory activity significantly increases with increasing repetition before about 23 times and then gets stable thereafter. Fig.3(b) shows that gradient of the memory activity changes with increasing repetition. The memory gradient gradually decreases into 0 when the repetition increases.

5.1. The computation of information gradient

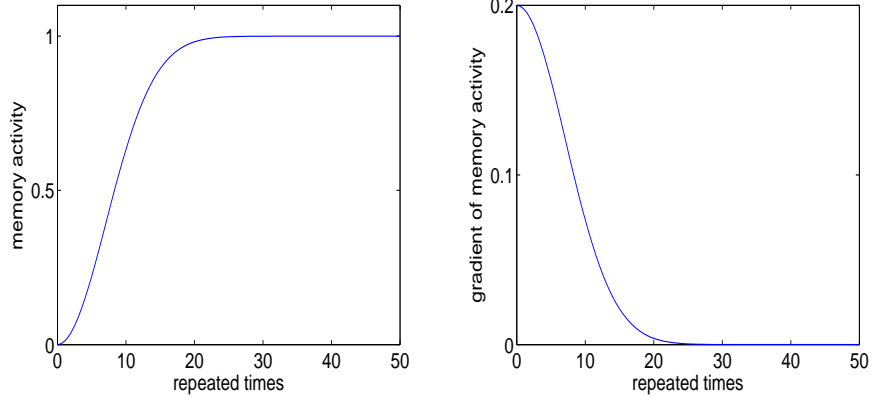
Inspired by memory activity lines in Fig.3, we propose information gradient which decreases as redundancy increase as follows.

Definition 5 (Information Gradient of Association Relation, $IG_t(\phi_i)$).

$IG_t(\phi_i)$ reflects the amount of association information in ϕ_i when it is described at t^{th} time

$$IG_t(\phi_i) = \begin{cases} 1 - EXP(-\lambda_i \times t^\alpha), & t = 1; \\ EXP(\lambda_i \times (t^\alpha - (t-1)^\alpha)), & t > 1. \end{cases} \quad (16)$$

$$t = \sum_{s_{(l)} \in S_{k-1}} I(\phi_i | s_{(l)}) \quad (17)$$



(a) memory activity changes with repeated times (b) memory gradient changes with repeated times

Figure 3: memory activity trend line changes with repetition

where $IG_t(\phi_i) \rightarrow [0, 1]$; $I(\phi_i|s_{(l)})$ is an indicator function, whose outcome is 1 when ϕ_i is consistent with $s_{(l)}$; t is the repetition of ϕ_i described by $S_{k-1} = \{s_{(l)} | 1 \leq l < k-1\}$; λ_i is a decay rate of ϕ_i .

From equation 16-17, we know that $IG_t(\phi_i)$ value is conditioned on repetition
 285 t of ϕ_i , which has the following properties by following lemmas:

Lemma 6. $IG_t(\phi_i)$ gradually decreases with t increases.

$$IG_t(\phi_i) > IG_{t+1}(\phi_i) \quad (18)$$

PROOF. Suppose $\sigma = IG_t(\phi_i) - IG_{t+1}(\phi_i)$ and $\rho(t) = EXP(-\lambda \times t^\alpha)$

Then,

$$\sigma = 2 \times EXP(-\lambda \times t^\alpha) - EXP(-\lambda \times (t-1)^\alpha) - EXP(-\lambda \times (t+1)^\alpha)$$

$$290 \quad \sigma = 2 \times \rho(t) - \rho(t-1) - \rho(t+1)$$

since $\rho(x) = EXP(-\lambda \times x^\alpha)$ is a concave function, $\sigma < 0$ and therefore lemma 6 is proofed.

Lemma 7. $IG_t(\phi_i)$ decreases more sharply with t increases.

Suppose $F(x) = \sum_x IG_x(\phi_i) = 1 - EXP(-\lambda \times x^\alpha)$ and $t_2 > t_1$

$$(F(t_2 + h) - F(t_2)) / (h \times (F(\infty) - F(t_2))) > (F(t_1 + h) - F(t_1)) / (h \times (F(\infty) - F(t_1))) \quad (19)$$

295 PROOF. Suppose $\sigma(t, h) = (F(t + h) - F(t)) / (F(\infty) - F(t))$ then,

$$\lim_{h \rightarrow 0} \sigma(x, h) / h = \lambda \times x^\alpha$$

$$\sigma = 2 \times \rho(t) - \rho(t - 1) - \rho(t + 1)$$

since $\rho(x) = EXP(-\lambda \times t^\alpha)$ is a monotone increasing function,

$\sigma(t_2, h) / h > \sigma(t_1, h) / h$ when $t_2 > t_1$ and lemma 7 is proofed

300 From lemma 6-7, we know that $IG_t(\phi_i)$ is higher when the association relation ϕ_i is first described. $IG_t(\phi_i)$ gradually decreases with increasing redundant contents of sentences. $IG_t(\phi_i)$ denotes the amount of association information of ϕ_i at t^{th} time. What's influence of $IG_t(\phi_i)$ on association relation distribution of an event? $IG_t(\phi_i)$ launches its influence on the potential function due by,

$$\Psi_{S_{k-1}}(X_{c_i}) = \prod_{\phi_i \in \Phi_{k \leq 2}} (\mu_i \times IG_t(\phi_i))^{I(X_{c_2} | \phi_i)} \quad (20)$$

305

$$t = \sum_{s_{(l)} \in S_{k-1}} I(\phi_i | s_{(l)}) \quad (21)$$

Compared equation 20 with equation 10, it is found that $IG_t(\phi_i)$ exerts influence on parameters μ_i in equation 20.

Supposing $S_{k-1} = \{s_l | 1 \leq l \leq k-1\}$ has described the association relation distribution before adding $s_{(k)}$, the association relation distribution is calculated

310 by,

$$P_{S_{k-1}}(X) = \mu_0 \prod_{c_i \in C} \Psi_{S_{k-1}}(X_{c_i}) \quad (22)$$

where $\Psi_{S_{k-1}}(X_{c_i})$ is refereed by equation 20.

5.2. The maximization of information gradient

Referred as equation 22 and equation 9, $IG(s_{(k)} | S_{k-1})$ is calculated by,

$$IG(s_{(k)} | S_{k-1}) = \sum_x P_{S_{k-1}}(x) I(s_{(k)} | x) \quad (23)$$

where $P_{S_{k-1}}(x)$ denotes a joint probability distribution of x before adding $s_{(k)}$;

315 $I(s_{(k)} | x)$ is an indicator function, whose outcome is 1 when x is consistent with $s_{(k)}$, 0 otherwise.

Lemma 8. $IG(s_{(i)}|S'') \geq IG(s_{(i)}|S')$ if $S'' \subseteq S'$

PROOF. $IG(s_{(k)}|S') = P_{S'}(X)I(s_{(k)}|x)$

$IG(s_{(k)}|S'') = P_{S''}(X)I(s_{(k)}|x)$

320 According lemma 6, $IG_{t^{s''}}(\phi_i) \geq IG_{t^{s'}}(\phi_i)$, since $t^{s''} < t^{s'}$
so $IG(s_{(i)}|S'') \geq IG(s_{(i)}|S')$, lemma 8 is proofed.

Referred as equation 9 in problem definition, the core semantics of event is obtained by maximizing information gradient of k sentences is calculated by,

$$CS(e) = \arg \max_{|S_k|=k} IG(S_k) \quad (24)$$

where $G(S_k) = IG(S_{k-1}) + IG(s_{(k)}|S_{k-1})$; $IG(s_{(k)}|S_{k-1}) = \sum_x P_{S_{k-1}}(x)I(s_{(k)}|x)$

325 If $IG(S_k)$ satisfies lemma 8, then $IG(S)$ is a submodular function [48]. For a submodular function, it has been proofed that the CELF method can obtain a near-optimal solution for maximizing information gradient of k sentences [48]. The equation 24 is maximized by the algorithm 3.

Algorithm 3 a solution for information gradient maximization

Input: an event $e = \{s_l | 1 \leq l < n\}$

Output: $CS(e)$

1. while ($sizeOf(CS(e)) \leq k$)
 2. $s_{(k)} = \arg \max_{s_i} IG(s_i | MAX_S_k)$
 3. $CS(e) = CS(e) \cup s_{(k)}$
 4. **return** $CS(e)$
-

6. Experiments

330 In this section, we conduct some experiments to validate the correctness of our method.

6.1. Datasets and evaluation measurement

To evaluate our method, we use TAC2008-TAC2010³ as dataset 1 and real-world microblog data crawled from Sina Weibo as dataset 2. Table 2 gives some statistics about the dataset 1 and dataset 2.

- 1) Dataset 1 includes 138 topics where each topic has a topic statement (title and narrative) and 20 relevant documents which have been divided into 2 sets: document set A and document set B. In this paper, we use set A in our experiments, where each topic has average 262 sentences and 4 manually generated summaries with 100-word length respectively.
- 2) Dataset 2 includes 20 events with total 725300 microblogs. For each event, we crawl microblogs in 30 days since its beginning timestamp. Besides we also collect the titles of news about these events from Baidu news⁴ in the same period. More details are shown in Table 3

Table 2: Description of datasets

Dataset source	Dataset 1	Dataset 2
	TAC2008-TAC2010	microblogs
Total # topics/event	138	20
Avg.# sentences in a topic	262	36255
Avg.# keywords	840	21367
Total # sentences	128414	725300

Baseline methods

We compare our methods with following state-of-the-art methods:

- 1) Cluster-based Conditional Markov random Walk Model (ClusterCMRW) [7]: it clusters sentences first and then ranks sentences in each clustering.
- 2) Cluster-based HITS Model (ClusterHITS) [7]: it clusters sentences first and then regards each clustering as hub and each sentence as authority. It uses hub value to rank clusters and use the authority value to rank sentences.

³<http://www.nist.gov/tac/data/index.html>

⁴<http://news.baidu.com>

Table 3: Description of microblog dataset in Sina Weibo

Event	Beginning Timestamp	#microblogs	#Title of Baidu News
The crisis in the Korean Peninsula	2013-03-08	105150	7288
USA Boston Marathon bombing	2013-04-15	42987	4531
the US surveillance program PRISM	2013-08-22	31939	2912
The crisis in Syria	2013-09-05	84172	9875
China's declaration of an air de- fense zone	2013-11-23	40414	3280
China's first moon rover, Yutu, or Jade Rabbit	2013-12-02	20621	1078
The crisis in Ukraine	2014-02-22	13317	3149
Malaysia Airline's flight 370 disap- peared	2014-03-08	62126	10368
Declaration of independence of Au- tonomous Republic of Crimea	2014-03-16	72079	5314
Sunflower Student Movement in Taiwan	2014-03-18	2358	153
South Korea's ferry accident	2014-04-16	34388	2172
981 drilling platform	2014-05-27	2552	257
Establishment of Shanghai Pilot Free Trade Zone	2013-09-30	1562	819
2014 FIFA World Cup Brazil	2014-06-13	128472	22206
Islamic State in Iraq and Syria	2014-06-29	4612	2037
The crash of Malaysia Airline's Boeing-777	2014-07-17	16248	2591
Scottish referendum	2014-08-05	1681	234
WHO issued that Ebola became an international public health emer- gency	2014-08-08	7533	4332
Taiwan gutter oil scandal	2014-09-04	12692	1457
Alibaba will begin I.P.O. Process in U.S.	2014-09-09	40399	4721
North Korean government an- nounced the withdrawal of nonag- gression treaty with South Korea	19 2013-03-08	105150	7288

Evaluate measurement

We use a widely used evaluation toolkit ROUGE [49] for evaluation. It measures summaries by counting the overlaps between the system generated summaries and human-written summaries as reference summaries. We mainly use ROUGE-1, ROUGE-2 and ROUGE-SU4 in our experiments.

ROUGE-N is calculated as follows:

$$ROUGE-n = \frac{\sum_{s \in ref} \sum_{n-gram \in s} I(n-gram|ref, gen)}{\sum_{s \in ref} \sum_{n-gram \in s} I(n-gram|ref)} \quad (25)$$

Where n denotes word length of n -gram; $I(n-gram|ref, gen)$ is an indicator function whose values is 1 when n -gram in the generated summary and reference summaries; $I(n-gram|ref)$ outcomes 1 when n -gram in reference summaries.

ROUGE-SU4 is calculated as follows:

$$ROUGE-SU4 = \frac{\sum_{s \in ref} (\sum_{skip2-gram \in s} I(skip2-gram|ref, gen) + \sum_{1-gram \in s} I(1-gram|ref, gen))}{\sum_{s \in ref} (\sum_{skip2-gram \in s} I(skip2-gram|ref) + \sum_{1-gram \in s} I(1-gram|ref))} \quad (26)$$

Where S4 denotes skip-bigram of any word pair in sentences whose word distance is at most 4; U denotes unigram.

6.2. Experimental setup

For evaluation of our method, we use dataset 1 and dataset 2 to conduct the experiments. The sentences in dataset 1 and dataset 2 are tokenized and stemmed by Stanford parser tools⁵. Our method for discovering core semantics is conducted as follows:

- 1) To obtain distribution of association semantics and enable semantics inference by low-degree association relations, we construct E-MRF model by a graph structure algorithm 1 and a potential value collaborative algorithm 2 in section 4.
- 2) To solve problem of maximizing information gradient, we select k sentence as the core semantics of event by algorithm 3 in section 5.

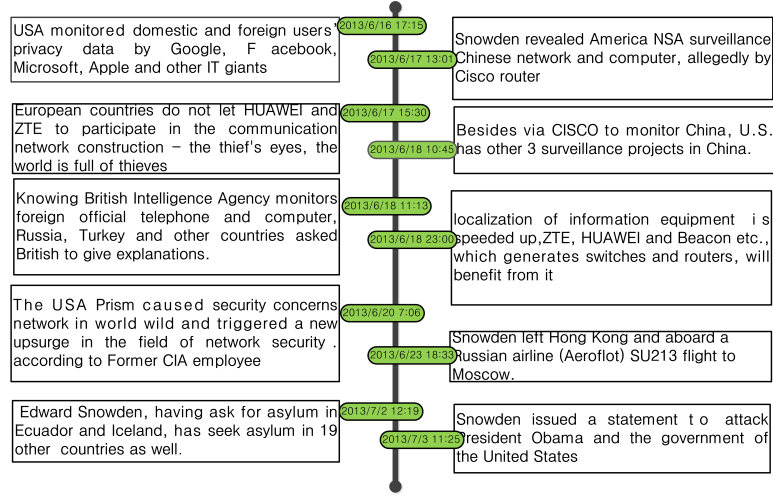


Figure 4: 10 sentences as the core semantics of event "The US surveillance program PRISM"

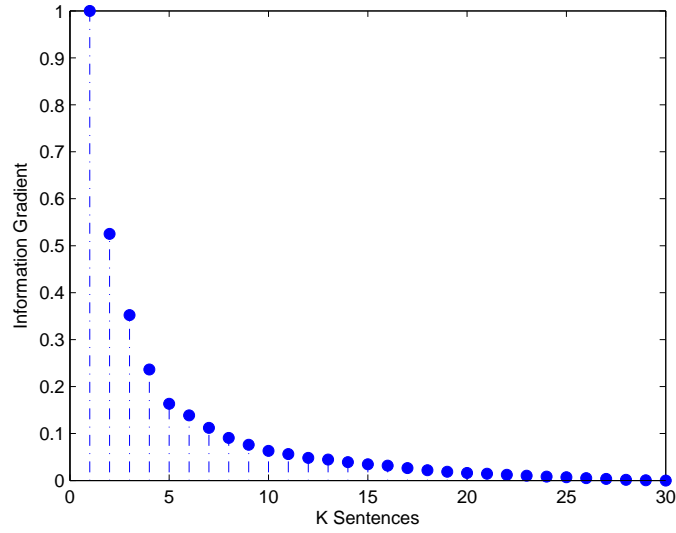


Figure 5: The normalized information gradient of k^{th} sentence ($k \leq 30$)

Fig.4 shows the results chronologically, the 10 sentences are selected from 31939 microblogs in event “The US surveillance program PRISM” from Sina Weibo. These 10 sentences mainly discuss different aspects of the event: s_1 answers which companies involving the surveillance program: “Google, Facebook, Microsoft, Apple”; s_2 is about “USA National Security Bureau monitors China network by the Cisco router”; s_3 answers “Why European countries refuse Huawei and ZTE to participate in communication network construction”; s_4 is about “Besides Cisco, 3 other surveillance projects are in china”; s_5 is about “Since British Intelligence Agency monitors foreign official telephone and computer, Russia and Turkey asked British to give explanation”; s_6 is about “Prism promoted information equipment localization”; s_7 is about “network security received increasing attention after Prism”; s_8 - s_{10} mainly about “the whereabouts of Snowden”.

Dividing information gradient by the maximum value, we normalize the information gradient. Fig.5 shows the normalized information gradient of k^{th} sentence ($1 \leq k \leq 30$). It shows that the value of information gradient decreases from 1 to a stable value which is approaching to 0. As more sentences are selected as core semantics of an event, the incoming sentence contains lower information gradient since most semantic association relations of the event have been described. The most of semantic association is covered by the first 15 sentences and the information gradient of remains sentences is extremely weak.

6.3. Experimental results

To evaluate our method on dataset 1, we compare our method with two baseline methods under three measurements as described in section 6.1. We extract k sentences ($1 \leq k \leq 15$) from each topic by our method and other two baseline methods. Table 4 compares the three methods by ROUGE-1, ROUGE-2 and ROUGE-SU4 under k sentences. Table 4 shows that our method always outperforms other two baseline methods on ROUGE-1, ROUGE-2 and ROUGE-SU4

⁵<http://nlp.stanford.edu/software/lex-parse.shtml>

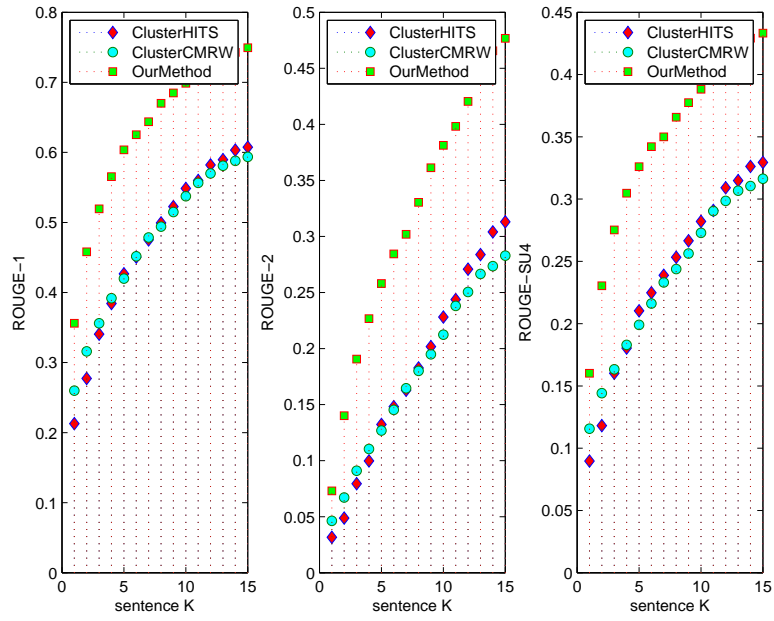


Figure 6: ROUGE-1,ROUGE-2 and ROUGE-SU4 between ClusterHITS, ClusterCMRW and Our method in dataset 1

Table 4: Comparison of different core semantics discovery methods on dataset 1

Evaluate measurement		ClusterCMRW	ClusterHITS	Our method	Sig.
Rouge-1	2	0.3159	0.2773	0.4582	0.00
	4	0.3918	0.3842	0.5653	0.00
	6	0.4516	0.4505	0.6252	0.00
	8	0.4941	0.4994	0.6701	0.01
	10	0.5373	0.5486	0.6986	0.00
	12	0.5700	0.5820	0.7218	0.00
	14	0.5879	0.6033	0.7427	0.00
Rouge-2	2	0.0671	0.0489	0.1401	0.01
	4	0.0910	0.0794	0.2267	0.00
	6	0.1104	0.0998	0.2844	0.00
	8	0.1269	0.1323	0.3304	0.00
	10	0.1453	0.1481	0.3814	0.00
	12	0.1645	0.1629	0.4205	0.00
	14	0.1801	0.1828	0.4656	0.00
Rouge-SU4	2	0.1181	0.1181	0.2304	0.01
	4	0.1602	0.1602	0.2752	0.00
	6	0.1804	0.1804	0.3048	0.00
	8	0.2103	0.2103	0.3261	0.00
	10	0.2248	0.2248	0.3421	0.00
	12	0.2388	0.2388	0.3500	0.00
	14	0.2533	0.2533	0.3658	0.00

405 with significant value ($sig. \leq 0.01$). Besides, Fig. 6 gives a comparison of average values of ROUGE-1, ROUGE-2, ROUGE-SU4 between ClusterCMRW, ClusterHITS and Our Method. The results show that our method has obviously higher values on ROUGE-1, ROUGE-2 and ROUGE-SU4.

The microblog data in dataset 2 has no standard reference data. To validate
 410 the efficiency of our method, we collected headlines of Baidu news in the same periods as reference data of dataset 2 since the headlines are condensed for the news. We expect that the core semantics of an event should be contained by these news titles. To verify our method is efficient on dataset 2, we conduct experiments as follows: For each event in dataset 2, we extract k sentences
 415 ($0 < k \leq 30$) from each event by our method and two other baseline methods respectively. For each event, we use the title of Baidu news of this event as reference data. We calculate gram-1, gram-2 and gram-SU4 by comparing machine generated sentences with title of Baidu news for each event. We compare our method with two baseline methods under ROUGE-1, ROUGE-2 and
 420 ROUGE-SU4 before 14 sentences in table 5. The results show that our method significantly preforms better than other baseline methods with significant value ($sig. \leq 0.01$).

Fig.7 shows the comparison of our method with other baseline methods on ROUGE-1,ROUGE-2 and ROUGE-SU4. We compare these methods under k
 425 sentences ($0 < k \leq 30$). It shows that our method obtain higher ROUG-1, ROUG-4 value than other methods do before 23 sentences; after 23 such advantages get week. Such phenomenon is caused by that the core semantics of these events have been coved by the 23 sentences and that adding new sentences dose not increase information gradient. However, our method has higher ROUGE-
 430 2 than other two methods. From the above analysis, we can verify that our method performs better for discovering the core semantics of event.

Table 5: Comparison of different core semantics discovery methods on dataset2

Evaluate measurement		ClusterCMRW	ClusterHITS	Our method	Sig.
Rouge-1	2	0.06994	0.08817	0.15833	0.00
	4	0.08987	0.14667	0.17528	0.00
	6	0.13014	0.16765	0.22743	0.00
	8	0.18694	0.22742	0.22340	0.00
	10	0.18991	0.23124	0.26261	0.00
	12	0.24205	0.23802	0.29207	0.00
	14	0.25265	0.25159	0.30161	0.01
Rouge-2	2	0.00128	0.00180	0.00385	0.01
	4	0.00333	0.003078	0.00487	0.00
	6	0.00487	0.004104	0.00590	0.00
	8	0.00590	0.004360	0.00795	0.00
	10	0.00641	0.00693	0.00821	0.00
	12	0.00667	0.00821	0.00923	0.00
	14	0.00769	0.00821	0.01231	0.00
Rouge-SU4	2	0.03189	0.03945	0.06954	0.01
	4	0.04134	0.06548	0.07800	0.00
	6	0.05927	0.07476	0.10097	0.00
	8	0.08413	0.09908	0.10142	0.00
	10	0.08539	0.10404	0.11665	0.00
	12	0.10809	0.10773	0.12998	0.00
	14	0.11304	0.11638	0.13502	0.00

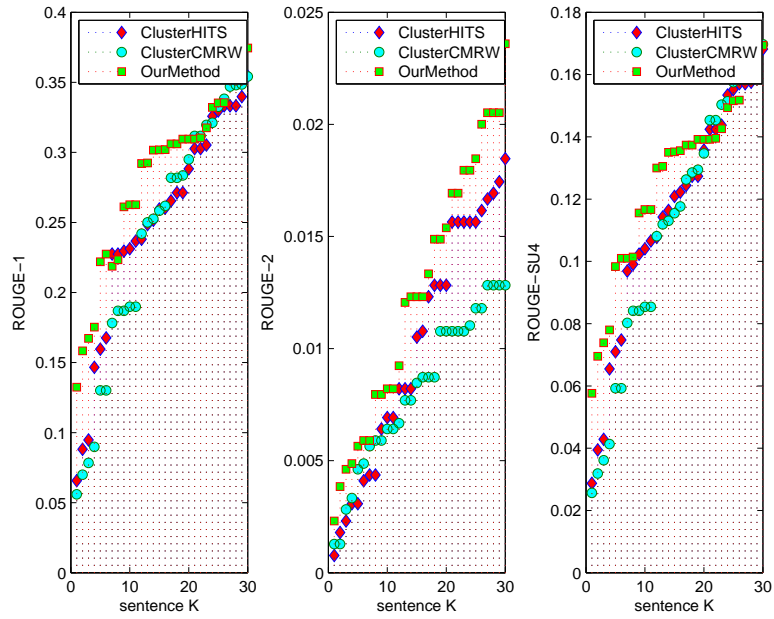


Figure 7: ROUGE-1,ROUGE-2 and ROUGE-SU4 between ClusterHITS, ClusterCMRW and Our method in dataset 2

7. Conclusion and future work

With widely use of open social media such as Twitter and Sina Weibo, the occurrence of a real-world hot event often causes large volumes of user-contributed “event messages” which contain different aspects of the event. Discovering the core semantics of event becomes a challenging problem since the core semantics of an event has been flooded by volumes of short texts which contain redundancy, noise and irrelevant content. The major challenges include:

- 1) how to learn association relations distribution by small-scale association relations;
- 2) how to maximize coverage of association relation distribution by the minimum number of short texts.

To solve the above challenging issues, the Markov random field based method extracts k sentences as the core semantics of event by,

- 1) semantic collaborative computation between event Markov random field and event power serial representation, which obtains association distribution by small scale association relations efficiently.
- 2) information gradient computation for maximizing information gradient of k sentences, which generates redundancy-free results by maximizing information gradient with the minimum number of short texts.

To evaluate our method, we compare our method with other state-of-the-art methods on TAC standard dataset and a large scale microblog dataset. The results show that our method outperforms other two baseline methods in discovering the core semantics of event.

Some users may consider how to organize these extracted short texts in semantic coherent way and others may focus on what’s the influence of other factors for discovering core semantics, such as temporal factor, user information and so on. We plan to explore these problems in future work.

References

- 460 [1] A. Nenkova, K. McKeown, A survey of text summarization techniques, in: Mining Text Data, Springer, 2012, pp. 43–76.
- [2] L. Vanderwende, H. Suzuki, C. Brockett, A. Nenkova, Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion, *Information Processing & Management* 43 (6) (2007) 1606–1618.
- 465 [3] D. Marcu, The rhetorical parsing, summarization, and generation of natural language texts. toronto, university of toronto, Ph.D. thesis, Tesis doctoral (1998).
- [4] R. Yan, L. Kong, C. Huang, X. Wan, X. Li, Y. Zhang, Timeline generation through evolutionary trans-temporal summarization, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, pp. 433–443.
- 470 [5] G. Erkan, D. R. Radev, Lexrank: graph-based lexical centrality as salience in text summarization, *Journal of Artificial Intelligence Research* (2004) 457–479.
- 475 [6] R. Mihalcea, S. Hassan, Using the essence of texts to improve document classification, in: Proceedings of RANLP, Citeseer, 2005.
- [7] X. Wan, J. Yang, Multi-document summarization using cluster-based link analysis, in: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2008, pp. 299–306.
- 480 [8] V. Qazvinian, D. R. Radev, Scientific paper summarization using citation summary networks, in: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, Association for Computational Linguistics, 2008, pp. 689–696.

- 485 [9] D. Shen, J.-T. Sun, H. Li, Q. Yang, Z. Chen, Document summarization using conditional random fields., in: IJCAI, Vol. 7, 2007, pp. 2862–2867.
- [10] X. Cai, W. Li, R. Zhang, Combining co-clustering with noise detection for theme-based summarization, *ACM Transactions on Speech and Language Processing (TSLP)* 10 (4) (2013) 16.
- 490 [11] G. Kumaran, J. Allan, Text classification and named entities for new event detection, in: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, 2004, pp. 297–304.
- [12] H. Becker, M. Naaman, L. Gravano, Selecting quality twitter content for events., *ICWSM* 11.
- 495 [13] R. Irfan, C. K. King, D. Grages, S. Ewen, S. U. Khan, S. A. Madani, J. Kolodziej, L. Wang, D. Chen, A. Rayes, et al., A survey on text mining in social networks, *The Knowledge Engineering Review* 30 (02) (2015) 157–170.
- 500 [14] Y. Sun, J. Han, J. Gao, Y. Yu, itopicmodel: Information network-integrated topic modeling, in: *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, IEEE, 2009, pp. 493–502.
- [15] S. Basu, I. Davidson, K. Wagstaff, *Constrained clustering: Advances in algorithms, theory, and applications*, CRC Press, 2008.
- 505 [16] H. Zhuge, Communities and emerging semantics in semantic link network: Discovery and learning, *Knowledge and Data Engineering, IEEE Transactions on* 21 (6) (2009) 785–799.
- [17] X. Luo, J. Zhang, F. Ye, P. Wang, C. Cai, Power series representation model of text knowledge based on human concept learning, *Systems, Man, and Cybernetics: Systems, IEEE Transactions on* 44 (1) (2014) 86–102.
- 510

- [18] X. Luo, Z. Xu, J. Yu, X. Chen, Building association link network for semantic link on web resources, *Automation Science and Engineering, IEEE Transactions on* 8 (3) (2011) 482–494.
- [19] H. Zhuge, Interactive semantics, *Artificial Intelligence* 174 (2) (2010) 190–204.
- [20] H. Zhuge, Cyber-physical societythe science and engineering for future society, *Future Generation Computer Systems* 32 (0) (2014) 180 – 186.
- [21] B. Xu, H. Zhuge, Automatic faceted navigation, *Future Generation Computer Systems* 32 (0) (2014) 187 – 197.
- [22] J. Chen, H. Zhuge, Summarization of scientific documents by detecting common facts in citations, *Future Generation Computer Systems* 32 (0) (2014) 246 – 252.
- [23] H. Zhuge, Semantic linking through spaces for cyber-physical-socio intelligence: A methodology, *Artificial Intelligence* 175 (5) (2011) 988–1019.
- [24] A. Celikyilmaz, D. Hakkani-Tür, Discovery of topically coherent sentences for extractive summarization, in: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, Association for Computational Linguistics, 2011, pp. 491–499.
- [25] D. Das, A. F. Martins, A survey on automatic text summarization, *Literature Survey for the Language and Statistics II course at CMU* 4 (2007) 192–195.
- [26] S. Tabassum, E. Oliveira, A review of recent progress in multi document summarization, in: *Doctoral Symposium in Informatics Engineering*, 2015.
- [27] B. Hayes-Roth, F. Hayes-Roth, Concept learning and the recognition and classification of exemplars, *Journal of Verbal Learning and Verbal Behavior* 16 (3) (1977) 321–338.

- [28] J. P. Minda, J. D. Smith, Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation., *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28 (2) (2002) 275.
- [29] Y. Yao, Interpreting concept learning in cognitive informatics and granular computing, *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 39 (4) (2009) 855–866.
- [30] J. Feldman, The simplicity principle in human concept learning, *Current Directions in Psychological Science* 12 (6) (2003) 227–232.
- [31] J. Feldman, An algebra of human concept learning, *Journal of mathematical psychology* 50 (4) (2006) 339–368.
- [32] J. Feldman, A catalog of boolean concepts, *Journal of Mathematical Psychology* 47 (1) (2003) 75–89.
- [33] J. Feldman, Minimization of boolean complexity in human concept learning, *Nature* 407 (6804) (2000) 630–633.
- [34] F. C. T. Chua, S. Asur, Automatic summarization of events from social media., in: *ICWSM, Citeseer*, 2013.
- [35] L. Shou, Z. Wang, K. Chen, G. Chen, Sumblr: continuous summarization of evolving tweet streams, in: *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, ACM*, 2013, pp. 533–542.
- [36] J. Leskovec, L. Backstrom, J. Kleinberg, Meme-tracking and the dynamics of the news cycle, in: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM*, 2009, pp. 497–506.
- [37] S. B. Kaleel, A. Abhari, Cluster-discovery of twitter messages for event detection and trending, *Journal of Computational Science* 6 (2015) 47–57.

- [38] D. Zhou, L. Chen, Y. He, An unsupervised framework of exploring events on twitter: Filtering, extraction and categorization, in: Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.
- [39] X. Zhou, L. Chen, Event detection over twitter social media streams, The VLDB Journal/The International Journal on Very Large Data Bases 23 (3) (2014) 381–400.
- [40] G. Petkos, S. Papadopoulos, Y. Kompatsiaris, Two-level message clustering for topic detection in twitter., in: SNOW-DC@ WWW, 2014, pp. 49–56.
- [41] M. I. Jordan, Graphical models, Statistical Science (2004) 140–155.
- [42] D. Pavlov, H. Mannila, P. Smyth, Beyond independence: Probabilistic models for query approximation on binary transaction data, Knowledge and Data Engineering, IEEE Transactions on 15 (6) (2003) 1409–1421.
- [43] F. Jelinek, Statistical methods for speech recognition, MIT press, 1997.
- [44] J. N. Darroch, D. Ratcliff, Generalized iterative scaling for log-linear models, The annals of mathematical statistics (1972) 1470–1480.
- [45] D. Pavlov, H. Mannila, P. Smyth, Probabilistic models for query approximation with large sparse binary data sets, in: Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., 2000, pp. 465–472.
- [46] C. Huang, A. Darwiche, Inference in belief networks: A procedural guide, International Journal of Approximate Reasoning 15 (3) (1996) 225–263.
- [47] D. Rundus, Analysis of rehearsal processes in free recall., Journal of experimental psychology 89 (1) (1971) 63.
- [48] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, N. Glance, Cost-effective outbreak detection in networks, in: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2007, pp. 420–429.

- [49] C.-Y. Lin, E. Hovy, Automatic evaluation of summaries using n-gram co-occurrence statistics, in: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, Association for Computational Linguistics, 2003, pp. 71–78.