Hierarchy-Cutting Model based Association Semantic for Analyzing Domain Topic on the Web

Zheng Xu, *Member, IEEE*, Shunxiang Zhang, Kim-Kwang Raymond Choo, *Senior Member, IEEE*, Lin Mei, Xiao Wei, Xiangfeng Luo, *Member, IEEE*, Chuanping Hu, and Yunhuai Liu, *Member, IEEE*

Abstract—Association link network (ALN) can organize massive web information to provide many intelligent services in the era of Big Data. Effective semantic layered technology not only can provide theoretical support for knowledge discovery in Web resources, but also can improve the searching efficiency of the related information system such as Web information system and industrial information system. How to realize the layer division of association semantic by the hierarcy analysis of ALN is an important research topic. To solve this problem, this paper proposes a hierarchy-cutting model of association semantic. First, some experiments of four types of keywords with different linking roles are conducted to discover the possible distribution law. Experimental results show that these keywords with association role reveal previous power-law distribution. Then, based on the discovered power-law distribution, up-cutting and down-cutting points are presented to divide the association semantic into three layers. At the same time, the theories of hierarcy-cutting model are presented. Finally, the examples of the current core topic and permernent topics belonging to a domain are given. The experiments show that hierarcy-cutting points have high accuracy. The multilayer theory of association semantic can provide a theoretical support for knowledge recommendation with different particle sizes on ALNs.

Index Terms—Association Link Network, Power-Law Distribution, Hierarchy-Cutting Model, Domain Topic.

1 INTRODUCTION

Web information contains plentiful, significant knowledge including explicit and implicit knowledge [1]. How to organize the Web information for facilitating knowledge discovery has been deeply investigated by some researchers. Association link network (ALN) is a kind of semantic link network built by mining the association relations among Web resources for effectively supporting Web intelligent application such as Web semantic association search, Web knowledge discovery, and recommendation [2, 3]. Xu et al. have studied on cloud environment for surveillance data management using video structural description [4], generating temporal semantic context of concepts [5]. Zhu et al. present discovering and learning communities and emerging semantics in semantic link network [6]. With the rapid development of information technology, human kinds are more likely to read and share information by similar intelligent applications. For example, the distributed and collaborative learning [7], semantic representation of scientific documents for supporting e-learning [8], discovering and searching of correlation between shared resources [9], and smart component technologies for human-centric computing [10], etc.

Based on the research about ALN, this paper explores the hierarchy-cutting model of association semantic by network analysis on the keyword-layer ALN. The model not only can provide theoretical support of domain topic (or say it as a kind of knowledge) discovery, recommendation on different particles/layers for users, but also can improve the searching efficiency based on ALN. The significant contributions of this paper are as follows:

- **The discovery of power-law distribution of three types of keywords with association linking role**. According to the role of association semantic link (ASL), two kinds of basic semantic features are defined. Further, ASLs are extracted on a given support and changing supports to compute the distribution of the three kinds of keywords with association role. The keywords with association semantic rolereveal obvious power-law distribution characteristic.
- **The presentation of hierarchy-cutting model of association semantic.** Based on the power-law distribution of association keywords, the basic idea of hierarchy-cutting model is presented by finding the up-cutting point and down-cutting point. And some corollaries about three-layer association semantic are presented. Top-layer association semantic is refined, which can improve the searching efficiency. And the bottom-layer association semantic can provide more details for user's information requirements including query and browsing.

In addition, based on the proposed model, some examples of the current core topic and permanent topics belonging to a domain are given. To our knowledge, the multilayer method of association semantic has not been well studied in the existing work.

The rest of this paper is organized as follows. Section 2

- *Zheng Xu, Lin Mei, and Chuanping Hu are with the Third Research Institute of the Ministry of Public Security, Shanghai, 201142, China.*
- *The corresponding author S.X. Zhang is with the School of Computer Science and Engineering, Anhui University of Science & Technology, Anhui Huainan, 232001, P.R. China. E-mail: sxzhang@aust.edu.cn.*
- *K. K. Choo is with University of Texas at San Antonio, USA.*
- *X. Luo is with Shanghai University, Shanghai, China.*
- *X. Wei is with Shanghai Institute of Technology, Shanghai, China.*
- *Y. Liu is with Beijing Institute of Big Data Research and Peking University Beijing, China.*

xxxx-xxxx/0x$xx.00 © 200x IEEE

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2017.2647986, IEEE Transactions on Industrial Informatics

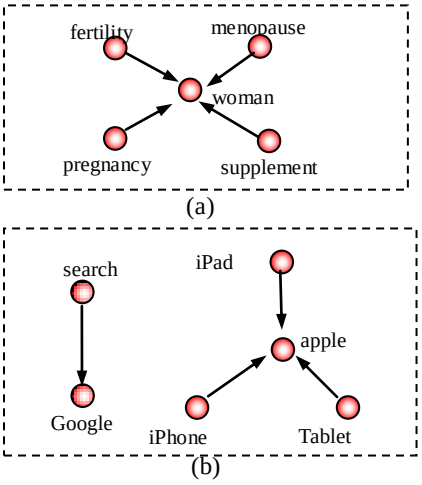

Fig. 1. Two basic association semantic features.

gives the related work. In Section 3, two basic semantic features and four kinds of keywords in association links are given. In Section 4, the distribution of four kinds of keywords is analyzed. In Section 5, hierarchy-cutting model of association semantic is proposed. Section 6 describes the experiments. In Section 7, the examples of domain topic on the Web are given. Finally, conclusions are drawn in Section 8.

2 Related Work

This section reviews two related works. One is the semantic representation and extracting semantic links. The other is about network analysis.

First, the semantic representation of a scientific document is proposed to generate the interconnection of the merged SRDFs belonging to one domain, which can be reflected by keywords' relations and their weights [11]. Luo et al. propose the extracting method of domain keywords of text for text classifying, clustering, and personalized services [12]. Recently, power series representation (PSR) model [13] is proposed, which has a lower computational complexity than latent Dirichlet allocation (LDA). Also, it contains more knowledge than vector space model (VSM).

In links/rules extracting, the algorithm of significant association rules between the items in the database is presented, which incorporates buffer management and novel estimation and pruning techniques [14]. The proposed relation extraction method mines generalized associations of semantic relations conveyed by the textual content of Web documents [15]. Xu et al. mine and annotate a semantic relation with temporal, concise, and structured information, which can release the explicit, implicit, and diversity semantic relations between entities [16]. These links have been widely applied in many fields such as co-experiencing in multiple spaces in lifetime [17] and organizing multimedia Big Data [18]. Andrei et al. have explored the graph structure in the Web, and found the power-law of degree distribution [19]. Ordinarily, the connection topology is assumed to be either completely regular or completely random. To interpolate between regular and random networks [20], with the help of recursion relations derived from the self-similar structure, Zhang et al. obtain the solution of average path length [21]. A theoretical approach for metric properties of uncorrelated random networks with hidden variables was presented. The approach was applied to calculate the exact expression for average path length in random networks [22]. It is commonly known that there exist short paths between vertices in a network showing the small-world effect [23]. Yet vertices, for example, the individuals living in society, usually are not able to find the shortest paths, due to the very serious limit of information. To theoretically study this issue, here the navigation process of launching messages toward the designated targets is investigated on a variant of the one-dimensional small-world network (SWN) [24]. Recently, some researchers have explored the spatial patterns of close relationships across the lifespan [25].

To our knowledge, the existing technologies on association semantic network cannot efficiently support the analysis of domain topic in large-scale Web pages because of two basic issues: 1) The existing methods fall short of multiple-layer model of large-scale Web resources for supporting the discovery of domain topic. 2) The existing network analysis mainly focuses on the characteristic analysis; there is no method or algorithm of division of association semantic to satisfy different users' information requirement. In this paper, we focus on how to realize the hierarchy-cutting of association semantic for analyzing the domain topic.

3 The Basic Semantic Feature of Keywords in ASL

In this section, two basic semantic features are presented to analyze the semantic association tendency of keywords. Further, all keywords in ALN are divided into four types of keywords with different association roles. Note that these keywords are domain keywords, which are extracted by the prior extracting method from Web resources [2].

3.1 Two basic semantic features

Definition 1: Active Traction Feature of Keyword (ATF)

Active Traction Feature of Keyword (ATF) is a kind of semantic feature, which is owned by antecedent keyword in a keyword-level ASL. In our research, a keyword with active traction feature must satisfy the following two conditions:

- A keyword k_m with ATF must be extracted from a domain Web resource on a given time window. This can ensure that this keyword can be used as an element of representing the domain knowledge.
- The keyword must be used as an antecedent keyword occurring in one or more ASLs.

This type of semantic feature comes from two types of knowledge. One is the well-known knowledge, which is human's association cognitive sense hidden in the human mind. For example, association link {"insulin"->"diabetes"} in health domain belongs to this type of knowledge. The other is unknown knowledge, which must be mined from Big Data. For example, association link {"polio"->"Bill Gates"} belongs to this type of knowledge. Here, "insulin" and "polio" have active traction semantic feature.

Usually, a keyword with ATF can be used to describe the attribute of an object or event. Or it is used to describe the part of an object or event. For example, in Figure 1(a), the keyword "woman" is an object, which has been listed with four attributes (described using keywords with active traction feature) such as "supplement," "menopause," "pregnancy," and "fertility." In Figure 1(b), "Google" and "apple" are two described objects, which have been listed with some active traction features such as "search," "iPad," and so on.

Definition 2: Passive Traction Feature of Keyword (PTF)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2017.2647986, IEEE Transactions on Industrial Informatics

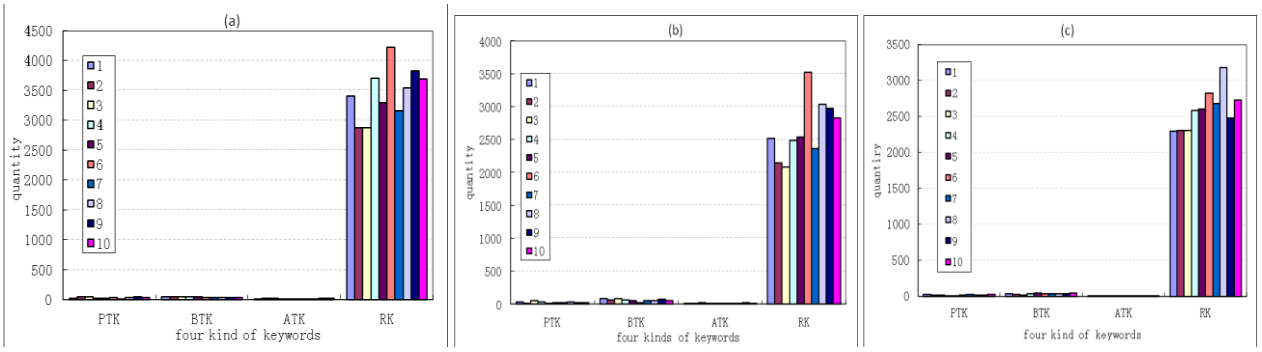

Fig. 3. The distribution of four kinds of keywords at a given support

Passive Traction Feature of Keyword (PTF) is another kind of semantic feature, which is owned by descendant keyword in a keyword-level ASL. In our research, a keyword with passive traction feature must satisfy the following two conditions:

- A keyword k_m with PTF must be extracted from a domain Web resource on a given time window. This kind of keyword is also used as an element of representing domain knowledge.
- The keyword must be used as a descendant keyword occurring in one or more ASLs.

Usually, these concept keywords with high frequency have this type of semantic feature. For example, the association link {"pregnancy"->"woman"} in human health domain (see Figure 1(a)), "woman" has passive traction feature. Similar keyword example with PTF include "emission," "climate" in environment domain, "Google" and "apple" in Internet domain.

In general, which semantic feature a keyword owns, ATF or PTF, is determined by its own semantic. Anyway, it is the basic unit of semantic representation of Web resources. Semantic feature is the source of the semantic link from a Web resource to other Web resources. A Web resource is linked to other Web resources, which are related to its ATF and PTF keywords. This problem will be discussed in the next section.

3.2 Four kinds of keywords

Based on the two basic semantic features, active traction and passive traction, all the extracted keywords that need to be used as a representation of domain knowledge from the Web resource [2] are divided into the following four types.

Definition 3: Active Traction Keyword (ATK)

For a keyword k_m, if it belongs to active traction keyword, it must satisfy the following two conditions:

- This keyword must be extracted from a domain Web resource and must be appointed/defined as one of the domain keywords by the TF/IDF method. This ensures that it can be used as an element of representing the semantic of Web resources.
- This keyword only has the semantic feature of active traction. That is, it must occur as the antecedent keyword of a keyword-level ASL on a given time window. Certainly, this ASL is a part of the keyword-level association semantic network.

Definition 4: Passive Traction Keyword (PTK)

Similar to the definition of ATK, if a keyword k_m belongs to passive traction keyword, it also must satisfy the following two conditions. One is that it must be one of the domain keywords extracted from a domain Web resource. The other is that it only has the semantic feature of passive traction. That is, it must occur as the descendent keyword of a keyword-level ASL on a given time window.

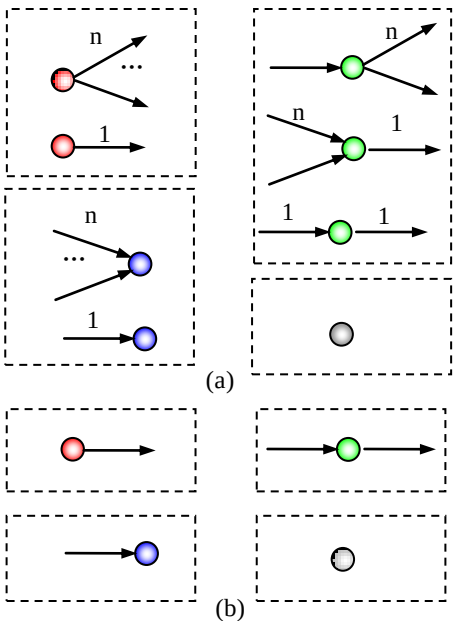

Fig. 2. Four kinds of keywords based on two basic association semantic features

Definition 5: Bridging Traction Keyword (BTK)

As a Bridging Traction Keyword (BTK), it has two types of semantic features. That is, in an ASL, it occurs as the antecedent keyword, which has the semantic feature of active traction. In another association link, it appears as descendent keyword, which has the semantic feature of passive traction. Certainly, it is a necessary condition that it is one of the domain keywords extracted from a domain Web resource. Obviously, this type of keyword has a more important link role in the semantic representation of Web resources.

Definition 6: Non-Traction Keyword (NTK)

The fourth type of keyword is the conversion to the BTK in the semantic feature. It does not have the two semantic features of active traction and passive traction. It is only used as one of the domain keywords extracted from a domain Web resource to represent the semantic of Web resources. Based on the definitions of the four kinds of keywords, we can simply plot their modes as in Figure 2.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2017.2647986, IEEE Transactions on Industrial Informatics

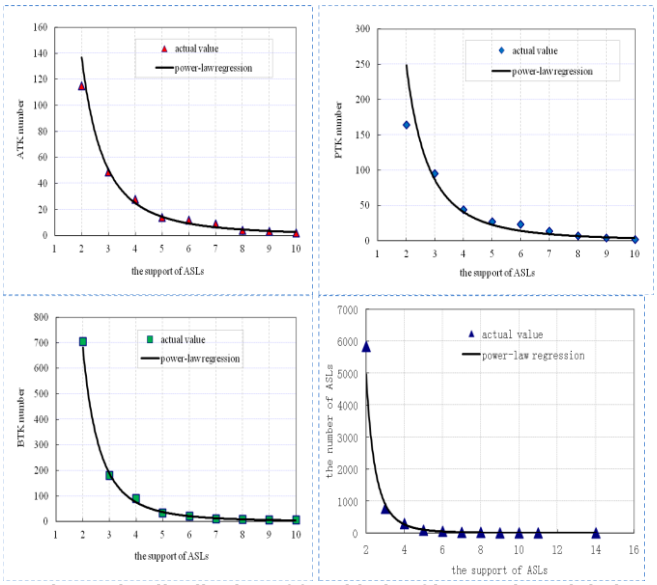

Fig. 4. The distribution of four kinds of keywords and their related ASLs

Usually, these concept keywords with high frequency have this type of semantic feature. For example, in the association link {"pregnancy"->"woman"} in human health domain (see Fig. 1(a)), "woman" has passive traction feature. Similar keyword example with PTF include "emission," "climate" in the environment domain, and "Google" and "apple" in the Internet domain.

4 THE DISTRIBUTION OF FOUR KINDS OF KEYWORDS

This section aims to explore the distributions of four kinds of keywords on two conditions: a given/fixed support and different/variable supports of ASLs. At the same time, the distributions of all keywords and ASLs are analyzed to do comparisons with the distributions of four kinds of keywords.

4.1 The Distribution of Keywords at a given/fixed support and confidence

In this kind of distribution analysis, data sources are Web resources of three domains (i.e., health, internet, and environment) on Reuters Website. Then, these Web resources are produced in a month according to a series of steps: extracting domain keywords, semantic representation, and mining ASLs. Note that the ASLs are mined at a given support and confidence by the first mining and second re-mining [2]. The support and confidence are set as 2% and 50%, respectively.

Based on the above steps, all the keywords are analyzed to classify them into one of the four kinds of keywords defined in Section 3.2. The statistical results of the four kinds of keywords have been plotted in Fig. 3. From Fig. 3, in the keywords of a domain, the proportion of NTK keywords is very high. It is more than 95% according to the statistical results of the three domains. At the same time, the proportion of these keywords with association role (i.e., constructing ASLs, antecedent keyword, and descendant keyword) is very low. The statistical results are, respectively, 2.2%, 3.5%, and 3.8% for the three domains, health, internet, and environment. According to these statistical data, the result about "association link/chain role" can be concluded as follows.

In the domain web resources text, the keywords with the semantic association link/chain role are very small. Most other keywords do not appear in the ASLs, and they do not have the role of semantic association. Actually, they only play the role

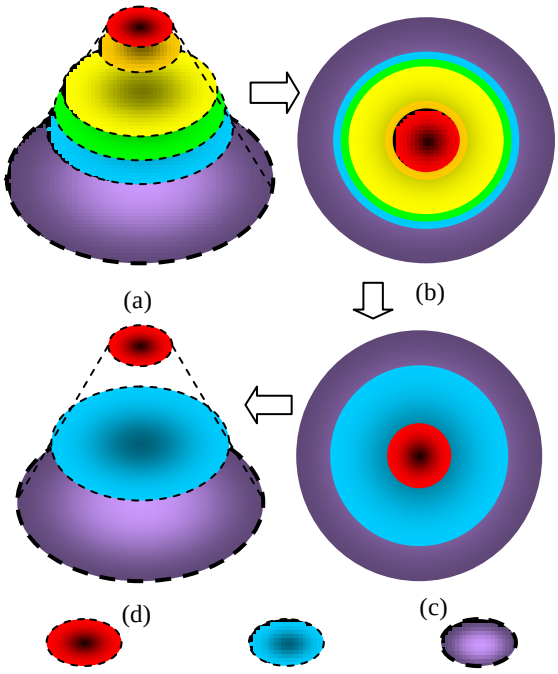

Fig. 5. The basic idea of a hierarchy-cutting model

of each text semantic description and representation.

In addition, the proportion of the keywords with the semantic association link/chain role in the different domains also reflects the clustering characteristics of the Web resources. For example, the semantic link network of Web resources of health domain has higher clustering coefficient [2]. Accordingly, its proportion of keywords with "link/chain" is smaller than the other two domains. Environment domain has smaller clustering coefficient. Accordingly, its proportion of keywords with "link/chain" is higher than the other two domains.

4.2 The distribution of four kinds of keywords at the changing support

Next, the analysis is carried out on different supports. So, we first give the calculation procedure of the distribution by adjusting the support as follows.

Algorithm 1: Analyzing the distribution of four kinds of keywords and their related ASLs

Input: the variable support, the semantic representation of Web resources on a month

Output: the number of different keywords and related ASLs

1: set the support of ASLs as 2
2: get the ASLs from the given semantic representation of Web resources by the method of ASL
3: count the number of four kinds of keywords occurring in these Web resources, and their related ASLs
4: if the number of obtained ASLs is 0, then go_to step 6
5: else add the support of ASL, and go_to step 2
6: end

Note that, in this analyzing algorithm, the support is simplified as an integer variable. Usually, the support is the proportion of the document frequency and the total number of Web resources. Because the total number of Web resources is fixed in our adjusting process, the support can be simplified as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2017.2647986, IEEE Transactions on Industrial Informatics

an integer variable.

According to this adjusting process, the distribution of four kinds of keywords, the related ASLs, and their regression results are plotted in Figure 4. In this regression analysis, different regression methods are adopted. Based on the probable trend, we employ power-law regression to analyze the distributions of ATK, PTK, BTK, and ASLs. Multinomial regression is selected to analyze the distribution of NTK. From Fig. 4, we can find that the distributions of ATK, PTK, BTK, and ASLs follow power-law function. And the distribution of NTK follows four times the multinomial function. Further, the Web resources of Internet and environment domain are selected to analyze the distribution of the four kinds of keywords and their related ASLs. The achieved parameters of the regression analysis are listed in Table 1.

TABLE 1
The power-law exponent and its complex correlation coefficient of regression analysis on four different supports

		ATK	*PTK*	*BTK*	*ASL*	*NTK*
H	*b*	2.25	2.87	2.39	3.76	/
	R	0.98	0.99	0.96	0.99	0.99
I	*b*	2.71	3.47	2.77	4.10	/
	R	0.97	0.987	0.94	0.99	0.98
E	*b*	2.19	3.11	2.95	4.21	/
	R	0.98	0.97	0.96	0.99	0.99

In Table 1, "*b*" denotes the exponent of power-law regression and "*R*" denotes the complex correlation coefficient of regression analysis. From Table I, we can find the following coincident result:

Result of regression analysis: four kinds of keywords and their related ASLs except NTK strictly follow the power-law distribution. Higher complex correlation coefficients show the correctness of the regression analysis. More importantly, the "core semantic" gradually occurs with the bigger support of ASLs.

5 HIERARCHY-CUTTING MODEL

In this section, we first give the basic idea of the layered theory. Then, the layered theory based on semantic concentricity degree is presented. Third, the computing steps of hierarchy-cutting model are presented.

5.1 The basic idea of hierarchy-cutting model

From the result of regression analysis in Section 3, all the keywords with semantic traction feature, ATK, PTK, and BTK follow the power-law distribution. At the same time, the keywords and their related ASLs are becoming less with the occurrence of larger support of ASLs. This means that some "basic semantic" is **stripped** and the "core semantic" gradually occurs.

If we regard the keywords-level association semantic network at a given support as a filled circle, then we have some concentric circles on different supports (from Figure 5(a) to Figure 5(b)). Further, these concentric circles can be simplified/reduced as Figure 5(c). Subsequently, the simplified concentric circles can be mapped into a multilayer semantic as shown in Figure 5(d).

Definition 7: Three-layer Association Semantic (TL-AS)

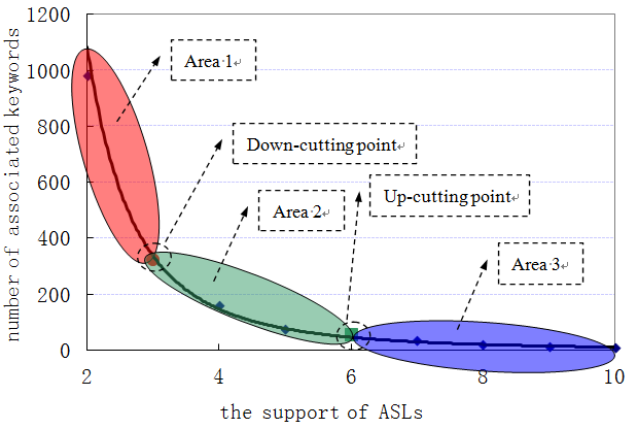

Fig. 6. The up and down hierarchy-cut point based on power-law distribution of association semantic

For the Web resources on a given time window, we can get some keywords-level association semantic network (k-ALN) $G_1, G_2 \cdots, G_m$ (corresponding to some concentric circles) on different supports $\{fre_1, fre_2, \cdots, fre_m\}$. If we can simplify them as Figure 5(c), then three-layer association semantic can be defined as follows:

- G_1 can be named as the "basic semantic" of the Web resources on the given time window. This kind of semantic corresponds to the largest (i.e., outer) of concentric circles in Fig. 5(c).
- G_i can be named as the "main semantic" of the Web resources on the given time window. This kind of semantic corresponds to the second large concentric circle in Fig. 5(c).
- G_j can be named as the "core semantic" of the Web resources on the given time window. This kind of semantic corresponds to the smallest concentric circle in Fig. 5(c).

Actually, this basic idea is reasonable from the mathematical theory. Because the keywords with "link/chain" reveal obvious power-law distribution, inflexion points are easily found. In Figure 6, the power-law distribution curve is divided into three areas according to two inflexion points, i.e. up-cutting point and down-cutting point. Therefore, the main tasks of hierarchy-cutting model are exploring the related theory and completing steps.

5.2 The layered theory based on semantic concentricity degree

Based on the basic idea of the layered theory, we present the detailed layered theory and its related concept.

Definition 8: Association Semantic Concentricity (ASC)

Association semantic concentricity is the repeatability score of association semantic contained in two k-ALNs G_i, G_j. For the Web resources on a given time window, if two supports $\{fre_i < fre_j\}$ are given, two related k-ALNs G_i, G_j can be achieved by Luo's method [3]. ASC $ASC(G_i, G_j)$ can be defined as

$$ASC(G_i, G_j) = R_{Gj} / R_{Gi}, R_{Gi} = \sqrt{Count_i / \pi} \quad (1)$$

where R_{Gi} and R_{Gj} denote the semantic radius of G_i and G_j. $Count_i$ denotes the number of ASLs in G_i.

Corollary 1. The value field of ASC

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2017.2647986, IEEE Transactions on Industrial Informatics

For the Web resources on a given time window, G_i and G_j are two k-ALNs generated on different supports $fre_i < fre_j$, then we have $0 < ASC(G_i, G_j) \le 1$.

According to Definition 7, for two supports fre_i and fre_j, if there is $fre_i < fre_j$, then, based on the generating method of k-ALNs,

$$G_j \subseteq G_i \tag{2}$$

Correspondingly, we have

$$R_{Gj} \le R_{Gi} \tag{3}$$

That is,

$$ASC(G_i, G_j) \le 1 \tag{4}$$

In addition, for any ALN, we have

$$R_{Gj} > 0 \tag{5}$$

So,

$$ASC(G_i, G_j) > 0 \tag{6}$$

Combining formulas (4) and (6), we have

$$0 < ASC(G_i, G_j) \le 1 \tag{7}$$

Corollary 2. The relations between the ASC and the exponent of power-law distribution

Larger exponent of power-law distribution leads to smaller average value of ASC. It is true on the contrary.

For the Web resources on a given time window, we can set m support fre, and ALNs $G_1, G_2 \cdots, G_m$ can be obtained. Larger exponent of power-law distribution means the changing velocity of association semantic among these ALNs. This inevitably leads to smaller average value of ASC.

Corollary 3. The changing trend of ASC

For m ALNs $G_1, G_2 \cdots, G_m$, if the number of their ASLs follow power-law distribution, then we have the sequence of ASC, $ASC(G_1, G_2)$, $ASC(G_2, G_3) \cdots$, $ASC(G_{m-1}, G_m)$, which is an increasing sequence.

Proof: Obviously, we only prove $\forall i \in (1, m)$, $ASC(G_{i-1}, G_i) < ASC(G_i, G_{i+1})$.

According to the definition of ASC, we have

$$ASC(G_{i-1}, G_i) = R_{Gi} / R_{Gi-1}$$

$$= \sqrt{a * fre_i^{-b} / \pi} \Big/ \sqrt{a * fre_{i-1}^{-b} / \pi} = [(i-1)/i]^{b/2} \tag{8}$$

Similarly,

$$ASC(G_i, G_{i+1}) = [i/(i+1)]^{b/2} \tag{9}$$

Therefore,

$$ASC(G_{i-1}, G_i) / ASC(G_i, G_{i+1}) = [(i^2-1)/i^2]^{b/2} < 1 \tag{10}$$

That is,

$$ASC(G_{i-1}, G_i) < ASC(G_i, G_{i+1}) \tag{11}$$

Thus, Corollary 3 is proved.

5.3 The computing steps of hierarchy-cutting model

Corollary 3 proves that the standard sequence of association semantic concentricity (ASC) is a strictly increasing sequence in the case of power-law distribution. However, the actual sequence of ASC is not a strictly increasing sequence. There is a certain difference between the standard sequence and the actual sequence. When the number of ASLs containing keyword-ALN of adjacent layer changes sharply, the semantic layer may have a large transition/jumping (such as the semantic transition/jumping from the "main semantic" to the "core semantic"). Thus, the following hypothesis of hierarchy-cutting of association semantic is proposed.

Hypothesis about association semantic cutting point: If it is called the theory of "standard ASC" on the condition of fully satisfying the power-law distribution. Obviously, when the difference between the actual ASC and the standard ASC is larger, larger transition/jumping occurs between two adjacent keyword-ALNs. Accordingly, the largest maximum and the second maximum transition/jumping location (i.e., a given support) can be viewed as two hierarchy-cutting points.

Based on the above hypothesis, the computing steps of hierarchy-cutting model can be described as follows:

(1) Computing the number of ASLs

On the m supports $fre = \{ fre_1, fre_2, \cdots fre_m \}$, the keyword-ALNs $\{ G_1, G_2, \cdots, G_m \}$ can be generated according to the extracted ASLs. Based on the regression analysis, the function between the number of ASLs $count$ and support fre should follow the equation

$$count = \alpha * fre^{-\beta} \tag{12}$$

where α and β are two constants. According to equation (12), the possible/theoretical value of the number of ASLs can be achieved, $\{ count_1', count_2', \cdots, count_m' \}$.

(2) Computing the association semantic concentricity

According to the real value of the number of ASLs, the semantic radius of $\{ G_1, G_2, \cdots, G_m \}$ can be computed. Further, association semantic concentricity of any two adjacent layers can be computed as $ASC(G_1, G_2)$, $ASC(G_2, G_3)$, $\cdots$, $ASC(G_{m-1}, G_m)$. In addition, the possible/theoretical value of association semantic concentricity of any two adjacent layers are also computed by equation (12), $ASC'(G_1, G_2)$, $ASC'(G_2, G_3)$, $\cdots$, $ASC'(G_{m-1}, G_m)$.

(3) Completing the hierarchy-cutting

This step aims to find the down-cutting point and up-cutting point. It can be described as finding two obvious changing/increasing values of ASC of any two adjacent layers, i.e.

$$\Delta ASC_i = ASC'(G_{i-1}, G_i) - ASC(G_{i-1}, G_i) \tag{13}$$

Suppose ΔASC_i and ΔASC_j are the maximum and secondary values by equation (13), then the up-cutting point of support frequency is i, and down-cutting point of support frequency is j.

6 EXPERIMENTS

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2017.2647986, IEEE Transactions on Industrial Informatics

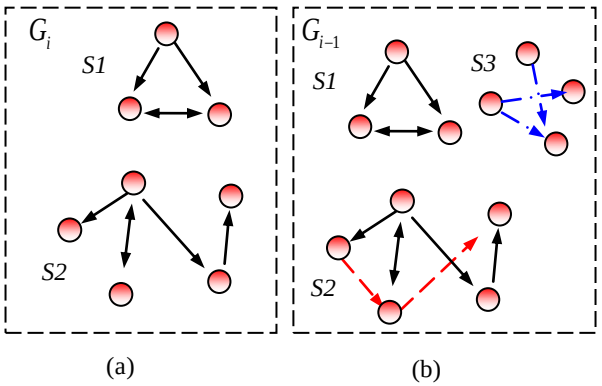

Fig. 7. Possible constitution of two keyword-level ALNs of adjacent layers

In this section, we present the experimental data and result analysis to verify the multilayer theory of association semantic.

6.1 Experiment for layered theory

Three domain news data are selected from the Website http://www.reuters.com, including health, environment, and internet, to build keyword-level ALNs on different supports. The time window of experimental data is set as a month.

For each domain news data, we independently execute the extracting method of ASL on an adjustable support. The initial value of the adjustable support is set as 2. Then, it is increased gradually. The increasing step is 1. The ASC of any two adjacent layers of keyword-level ALNs are computed by Definition 7. The computed results are listed in Table 2.

Table 2
The ASC of three domains on different supports

sup	*H-ASC*	*I-ASC*	*E-ASC*
2,3	0.343132	0.414126	0.412968
3,4	0.637229	0.628666	0.474075
4,5	0.686762	0.639767	0.61808
5,6	0.758913	0.682575	0.641689
6,7	0.703211	0.662589	0.845154
7,8	0.788811	0.781736	0.894427
8,9	0.779194	0.797724	0.707107
9,10	0.8044	0.755929	无
aver	**0.687706322**	**0.670389113**	**0.656214**

In Table 2, the column "*sup*" denotes two supports of two adjacent layers. "*H-ASC,*" "*I-ASC,*" and "*E-ASC,*" respectively, denote the association semantic concentricity of health, Internet, and environment news domains. And the average value of association semantic concentricity of each domain is listed in the last row of Table 2. From Table 2, the following conclusions can be drawn.

- Larger exponent of the power-law distribution leads to smaller average value of ASC. It verifies the correctness of Corollary 2. We compare the results of power-law exponent listed in Table 1 and the average value of association semantic concentricity listed in Table 2. The keyword-level ALN of environment news domain has the largest power-law exponent and the smallest association semantic concentricity.
- It has larger ASC at larger support between two adjacent layers. It verifies the correctness of Corollary 3. From Table 2, although a few decreases in association semantic concentricity have occurred, this kind of increasing trend of two adjacent layers in total is obvious.

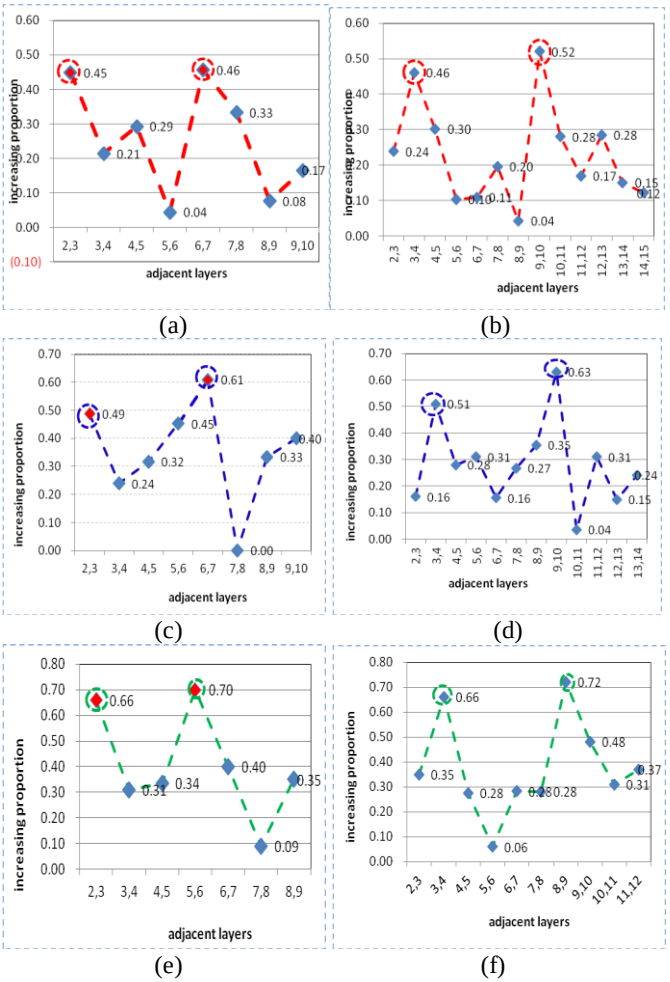

Fig. 8. Experiments for up- and down-cutting points based on three domains

6.2 Evaluation method of hierarchy-cutting model

The evaluation of hierarchy-cutting model is mainly the precision of cutting points including up-cutting point and down-cutting point.

Let G_{i-1} and G_i denote two keyword-level ALNs of adjacent layers. Figure 7 gives the possible constitution of two keyword-ALNs of adjacent layers. In Figure 7, G_i is made up of two subgraphs, S_1 and S_2. And G_{i-1} is made up of three subgraphs, S_1, S_2, and S_3. From Figure 7, G_{i-1} has been added to two kinds of ASLs. One is making the semantic of old subgraph more plentiful, which is plotted by dashed line in subgraph S_2. The other is new occurring ASLs, which constitute a new subgraph. Obviously, the proportion of this kind of new subgraph is larger, the effect is better as the cutting point of hierarchy model.

Therefore, the evaluation method of hierarchy model can convert into the computation of the proportion of newly occurring ASLs. For the series of keyword-level ALNs $\{ G_1, G_2, \cdots, G_m \}$, the proportion series of added new occurring ASLs can be achieved, which is denoted as $\{ pn_2, \cdots, pn_m \}$. Then, the next task is finding the maximum and secondary values in $\{ pn_2, \cdots, pn_m \}$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2017.2647986, IEEE Transactions on Industrial Informatics

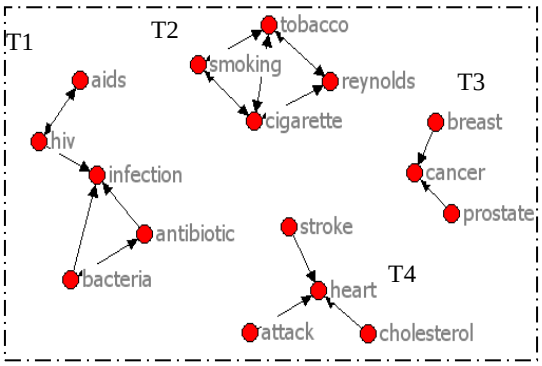

Fig. 9. Examples of current core topic of the health domain

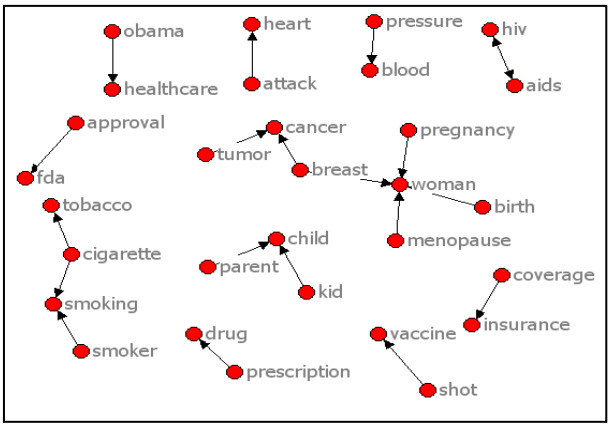

Fig. 10. Examples of long-term topic of the health domain

6.3 Experimental results for hierarchy-cutting model

Based on the proposed evaluation method, three groups of experiments have been carried out and the results are plotted in Figure 8. The first group is Figure 8(a) and (b), which shows the experimental results of health domain on two sizes of ALN. The other two groups are, respectively, internet and environment domains. From Figure 8, the following conclusions can be drawn.

- **Two obvious peak values**. From the three groups of experiments, the proportion series of added new occurring ASLs between adjacent layers are all occurring between two obvious peak values. These two peak values ensure the precision of the hierarchy-cutting model.
- **The stability of peak values**. Two peak values have stability on the given support. For three domains, the supports of down-cutting point are about 2.7%, 2.3%, and 2.0%, respectively. And the supports of up-cutting point are about 8.5%, 8.0%, and 7.8%, respectively.

7 DISCOVERY OF DOMAIN TOPIC ON THE WEB

In this section, two kinds of domain topics on the Web are given. One is the current core topic of a domain; the other is long-term topic.

7.1 Examples of current core topic of a domain

Current core topic of a domain can be defined as a kind of topic focused by users during a short time (e.g., a day, a week, a month). The data resources of this paper are processed within a month on the environment of Big Data. So, it can satisfy the time requirement of mining the current core topic of a domain. In addition, the layer of "core semantic" separated by hierarchy-cutting model can provide support for the discovery of current core topic. Actually, the discovery of current core topic can be simplified as the searching of semantic subgraph from the layer of "core semantic." Figure 9 has given examples of current core topic of the health domain.

Fig. 9 shows four core topics of health domain on 2015.11. The first topic discusses about the "aids" including infection, antibiotic, etc. The second topic tells us about the problem of "smoking" and health. The third topic describes a new technology about "cancer treatment" problem such as breast cancer, prostate cancer, and so on. The fourth topic discusses the relation between "heart attack" and "cholesterol."

7.2 Examples of long-term topic of a domain

Long-term topic of a domain can be defined as a kind of topic focused by users during a long time (e.g., a year or more). The layer of "main semantic" can provide support for the discovery of long-term topic. Similar to the discovery of current core topic, finding long-term topic can be simplified as the searching of semantic subgraph from the layer of "main semantic." Figure 10 has given examples of long-term topic of the health domain.

Fig. 10 presents the semantics of 11 long-term topics in the health domain. They include: healthcare plan followed by Obama, heart disease, human blood disease, AIDS, all kinds of food safety licensing issues approved by the U.S. Food and Drug Administration (FDA), women's health including pregnancy, menopause, breast cancer and so on, tobacco and smoking, the influence of education mode to kids disease; medical insurance coverage, vaccine shot, and drug prescriptions.

8 CONCLUSIONS

The hierarchy-cutting model is presented to realize three-layers division of association semantic for ALNs, which is suitable for organizing Web resources on the era of Big Data. It not only can provide theoretical support of the analysis of domain topic (e.g., current core topic, long-term topic belonging to a domain), but also can satisfy users' information requirement on different particles/layers. The power-law distribution of keywords with association linking role has been discovered by regression analysis. According to the role of ASL, two kinds of basic semantic features are defined. Further, ASLs are extracted on different supports to compute the distribution of four kinds of keywords. All the keywords with association semantic role reveal obvious power-law distribution characteristic by regression analysis. The hierarchy-cutting model of association semantic has been built. Based on the power-law distribution of association keywords, the basic idea, theory, and steps of hierarchy-cutting model are presented. The proposed model has completed the division of three-layer association semantic, that is, "basic semantic," "main semantic," and "core semantic." "Core semantic" can provide the support for discovering current core topic. And the other two semantic layers can provide more details of the related topic.

In future, we need further efforts to explore the organization

and recommendation of the analyzed topic based on the hierarchy-cutting model.

ACKNOWLEDGMENTS

This work was supported by the Natural Science Foundation of Anhui Province Universities (No. KJ2015A111), in part by the National Science Foundation of China under Grant 61300202, 61332018, 61471232, and the Science Foundation of Shanghai under Grant 16ZR1435500.

REFERENCES

[1] S.X. Zhang, K. Lu, W. Liu, X. Yin, and G. Zhu, Generating associated knowledge flow in large-scale web pages based on user interaction. Computer Systems Science and Engineering , 30(5):377-389, 2015.

[2] S.X. Zhang, X.F. Luo, J.Y. Xuan, et al., Discovering small-world in association link networks for association learning. World Wide Web, 17(2):229-254, 2014.

[3] X.F. Luo, Zh. Xu, J. Yu, et al., Building association link network for semantic link on web resources. IEEE Trans. Autom. Sci. Eng. 8(3):482-494, 2011.

[4] Zh. Xu et al., Semantic based representing and organizing surveillance big data using video structural description technology. The Journal of Systems and Software, 102:217-225, 2015.

[5] Zh. Xu et al. Generating temporal semantic context of concepts using web search engines. Journal of Network and Computer Applications, 43:42-55, 2014.

[6] H. Zhuge, Communities and emerging semantics in semantic link network: Discovery and learning. IEEE Transactions Knowledge and Data Engineering, 21(6):785-799, 2009.

[7] Q. Li, R.W.H. Lau, T.K. Shih, et al., Technology supports for distributed and collaborative learning over the internet. ACM Trans. on Internet Technology, 8(2):10:1-10:24, 2008.

[8] X.F. Luo, N. Fang, et al., Semantic representation of scientific documents for the e-science knowledge grid. Concurrency and Computation: Practice and Experience, 20(7):839-862, 2008.

[9] N.Y. Yen, R.H. Huang, J.H. Ma, Q. Jin, and T.K. Shih, Intelligent route generation: Discovery and search of correlation between shared resources. Int. J. Commun. Syst. 26:732-746, 2013.

[10] J.P. James, C. Antonio, H. Chang, and K. Andrew, Introduction to the thematic issue on ambient and smart component technologies for human centric computing. JAISE 6(1): 3-4, 2014.

[11] H. Zhuge and X.-F. Luo, Automatic generation of document semantics for the e-science knowledge grid. The Journal of Systems and Software, vol. 79, pp. 969-983, 2006.

[12] X.-F. Luo and N. Fang, Experimental study on the extraction and distribution of textual domain keywords. Concurrency and Computation: Practice and Experience, 20:1917-1932, 2008.

[13] X. Luo, J. Zhang , F. Ye , P. Wang, and C. Cai, Power series representation model of text knowledge based on human concept learning. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 44(1):86-102, 2014.

[14] R. Agrawal, T. Imielinski, and A. Swami, Mining association rules between sets of items in large databases. In Proc. of the 1993 International Conf. Management of Data (SIGMOD 93), pp. 207-216, 1993.

[15] T. Jiang, A. Tan, and K. Wang, Mining generalized associations of semantic relations from textual web content. IEEE Trans. Knowledge and Data Eng., 19(2):164-179, 2007.

[16] Zh. Xu et al. Mining temporal explicit and implicit semantic relations between entities using web search engines. Future Generation Computer Systems, 37:468-477, 2014.

[17] H. Zhuge. Semantic linking through spaces for cyber-physical-socio intelligence: A methodology. Artificial Intelligence, 175:988-1019, 2011.

[18] C. Hu, Zh. Xu, et al. Semantic link network based model for organizing multimedia big data. IEEE Transactions on Emerging Topics in Computing, 2(3):376-387, 2014.

[19] B. Andrei, K. Ravi, and M. Farzin, Graph structure in the Web. Computer Networks, 33:309-320, 2000.

[20] D.J. Watts and S.H. Strogatz, Collective dynamics of "Small-World" networks. Nature, 393:440-442, 1998.

[21] Zh. Zhang, L. Chen, Sh. Zhou, et al. Analytical solution of average path length for Apollonian networks, Physical Review E 77, 017102-1~017102-4, 2008.

[22] A. Fronczak, P. Fronczak, and J.A. Hołyst, Average path length in random networks. Physical Review E, 70, 056110-1~056110-7, 2004.

[23] M.E.J. Newman, The structure and function of complex networks. SIAM Rev. 45:167-256, 2003.

[24] J.M. Kleinberg, Navigation in a small world. Nature 406:845, 2000.

[25] H.H. Jo, J. Saramaki, R.I.M. Dunbar, and K. Kaski, Spatial patterns of close relationships across the lifespan. Science Reports 4:6988, 1-7, 2014.

Zheng Xu received Diploma and Ph.D. from the School of Computing Engineering and Science, Shanghai University, Shanghai, in 2007 and 2012, respectively. He is currently working in the Third Research Institute of Ministry of Public Security. He has authored or co-authored more than 70 publications including IEEE Trans. on Fuzzy Systems, IEEE Trans. on Automation Science and Engineering, IEEE Trans. on Cloud Computing, IEEE Trans. on Emerging Topics in Computing, IEEE Trans. on Systems, Man, and Cybernetics: Systems, IEEE Trans. on Big Data, etc.

Shunxiang Zhang received his Ph.D. from the School of Computing Engineering and Science, Shanghai University, Shanghai, in 2012. He is a professor at Anhui University of Science and Technology, China. His current research interests include Web Mining, Semantic Search, and Complex Network.

Kim-Kwang Raymond Choo received his Ph.D. in Information Security in 2006 from Queensland University of Technology, Australia. He is currently a cloud technology endowed associate professor at University of Texas at San Antonio, an associate professor at the University of South Australia.

Xiao Wei is an associate professor of Shanghai Institute of Technology. His main research interests include Web Mining, Semantic Search, and E-learning.

Xiangfeng Luo is a professor in the School of Computers, Shanghai University, China. Currently, he is a visiting professor at Purdue University, USA. His main research interests include Web Wisdom, Cognitive Informatics, and Text Understanding. He has authored or co-authored more than 100 publications

Lin Mei received his Ph.D. from Xi'an Jiaotong University, Xi'an, China, in 2000. He is a Research Fellow. From 2000 to 2006, he was a Postdoctoral Researcher with Fudan University, Shanghai, China; the University

of Freiburg, Freiburg in Breisgau, Germany; and the German Research Center for Artificial Intelligence. He is currently the Director of the Technology R&D Center for the Internet of Things with the Third Research Institute of the Ministry of Public Security, China.

Chuanping Hu received his Ph.D. from Tongji University, Shanghai, China, in 2007. He is a Research Fellow and the Director of the Third Research Institute of the Ministry of Public Security, China. He is also a specially appointed Professor and a Ph.D. supervisor with Shanghai Jiao Tong University, Shanghai, China. He has published more than 20 papers, has edited five books, and is the holder of more than 30 authorized patents. He is the chairman of ACM Shanghai Chapter.

Yunhuai Liu is a professor in the Third Research Institute of Ministry of Public Security, China. He received Ph.D. from Hong Kong University of Science and Technology (HKUST) in 2008. His main research interests include wireless sensor networks, pervasive computing, and wireless network. He has authored or co-authored more than 50 publications and his publications have appeared in IEEE Trans. on Parallel and Distributed Systems, IEEE Journal of Selected Areas in Communications, IEEE Trans. on Mobile Computing, IEEE Trans. on Vehicular Technology, etc.