Building the Multi-layer Theory of Association Semantic based on the Power-law Distribution of Linking Keywords

Guangli Zhu, Xiaojun Tang, Shunxiang Zhang
School of Computer Science and Engineering,
Anhui University of Science & Technology
Huainan, China
glzhu@aust.edu.cn,1262989796@qq.com,
sxzhang@aust.edu.cn

Zheng Xu*
The Third Research Institute
the Ministry of Public Security
Shanghai, China
xuzheng@shu.edu.cn
*Corresponding author

***Abstract*—Web information contain plentiful, significant knowledge which is eager to be explored by users. 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. This paper builds the multi-layer theory of association semantic based on the power-law distribution of linking keywords. First, some experiments of four types of keywords with different linking role are done to discover the possible distribution law. Experiment results show that four types of keywords are all reveal power-law distribution. Then, based on the discovered power-law distribution, the multi-layer theory of association semantic is built. The multi-layer theory of association semantic can provide a theoretical support for knowledge recommendation with different particle size on Association Link Network (ALN).**

Keywords- Association Link Network, power-law distribution, multi-layer theory of association semantic, knowledge discovery in Web resources.

I. INTRODUCTION

Web information contain plentiful, significant knowledge including explicit and implicit knowledge [1]. How to organize these Web information for facilitating knowledge discovery has been deeply 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 presents 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 these research about ALN, this paper explores the multi-layer theory of association semantic to provide theoretical support of knowledge discovery, recommendation on different particles/layers for users. The significant contributions of this paper are as follows:

- The power-law distribution of four types of keywords with different linking role. According to the role of association semantic link, two kinds of semantic features are defined. Further, association semantic links are with 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.
- The multi-layer theory of association semantic. Based on the power-law distribution of association keywords, Association Semantic Concentric (ASC) is defined. And corollary about the relations between the ASC and the exponent of power-law distribution is presented. Larger exponent of power-law distribution leads to smaller average value of ASC. In addition, the corollary about changing trend of ASC among any two adjacent layers is proved.

To the best of our knowledge, the multi-layer method of association semantic has not been well studied in existing work.

The rest of this paper is organized as follows. In section II, two semantic features and four kinds of keywords in association links are given. In section III, The distribution of four kinds of keywords is presented. In section IV, the layered theory of association semantic are proposed. In section V, the multi-layer theory is validated by experiments. Finally, section VI concludes the work of this paper.

II. TWO SEMANTIC FEATURES AND FOUR KINDS OF KEYWORDS IN ASSOCIATION LINKS

Different types of keywords have different roles in association semantic links. In this section, we will analyze the

DOI reference number: 10.18293/SEKE2016-167

semantic association tendency (i.e., active traction and passive traction) of keywords to divide these keywords into four types of keywords. Note that these keywords are domain keywords which are extracted by the prior extracting method from Web resources[2].

A. Two 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 association semantic link. 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 resources on a given time window. This can ensure that this keyword can be used as a element of representing domain knowledge.
- The keyword must be used as an antecedent keyword occurred in one or more association semantic links.

This type of semantic feature comes from two types of knowledge. One is well-known knowledge which is human's association cognitive sense hidden in 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 massive data. For example, association link {"polio"->"Bill Gates"} belongs to this type of knowledge. Here, "insulin" and "polio" all have active traction semantic feature.

Usually, a keyword with ATF can be used to describe the attribute of an object, event. Or it is used to describe the part of an object, event. For example, in Figure 1 (a), the keyword "woman" is an object which has been listed 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 the described objects , which have been listed some active traction features such as "search", "iPad" and so on.

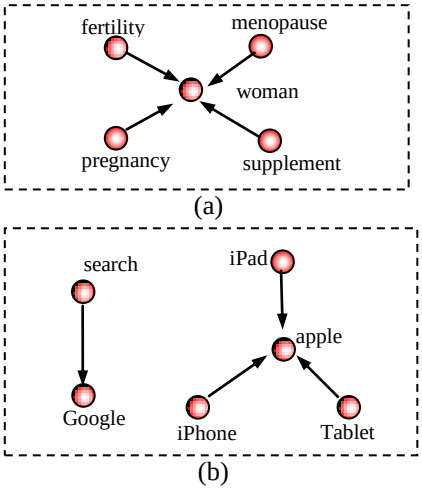


Figure 1. Two semantic features in association semantic link

Definition 2: Passive Traction Feature of Keyword (PTF)

Passive Traction Feature of Keyword (PTF) is another kind of semantic feature which is owned by descendant keyword in a keyword-level association semantic link. 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 resources on a given time window. This kind of keyword is also used as a element of representing domain knowledge.
- The keyword must be used as an descendant keyword occurred in one or more association semantic links.

Usually, those concept keywords with high frequency has this type of semantic feature. For example, 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, what semantic feature a keyword owns, ATF or PTF, is determined by its own semantic meaning. In 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.

B. Four kinds of keywords

Based on the two basic semantic features, active traction and passive traction, all the extracted the keywords to be used as 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 resources and must be appointed/defined as one of the domain keywords by the TF/IDF method. This ensures that it can be used as a 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 association semantic link on a given time window. Certainly, this association semantic link is the 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 resources. 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 association semantic link on a given time window.

Definition 5: Bridging Traction Keyword (BTK)

As a Bridging Traction Keyword (BTK), it has two types of semantic feature. That is, in a association semantic link, 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 resources. Obviously, this type of keywords has more important link role in the semantic representation of Web resources.

Definition 6: Non-Traction Keyword (NTK)

The fourth type of keyword is on the convert to the BTK in the semantic feature. It has not 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 resources to represent the semantic of Web resources.

Based on the definitions of four kinds of keywords, we can simply plot their modes as Figure 2.

Usually, those concept keywords with high frequency has this type of semantic feature. For example, 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.

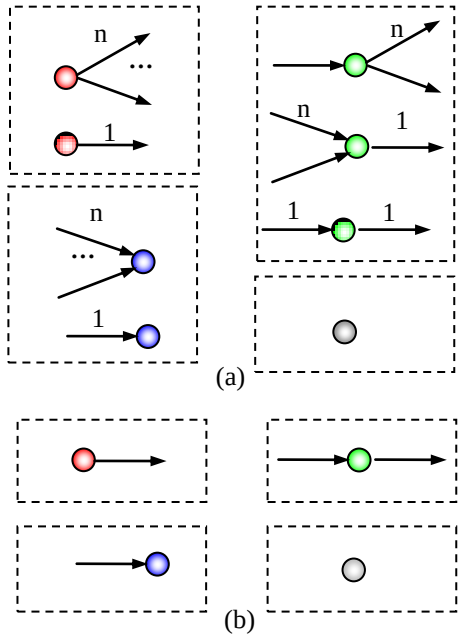


Figure 2. the graph of keywords set

III. THE DISTRIBUTION OF FOUR KINDS OF KEYWORDS

In this section, we use the Web resources of a month belonging to the health domain as the researched data set. We explore the distributions of four kinds of keywords on the different/variable supports of association semantic links (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.

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

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

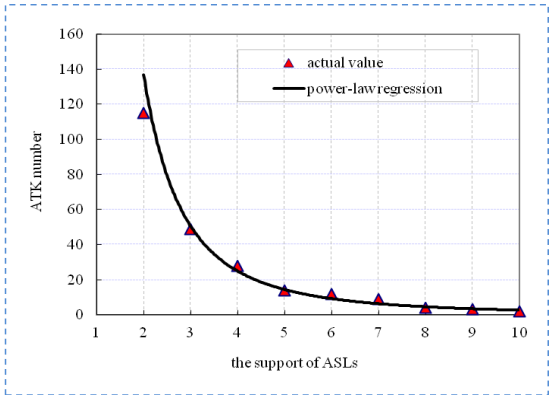
Input: the variable support, the semantic representation of Web resources on a month

Output: the number of different keywords and related ASLs

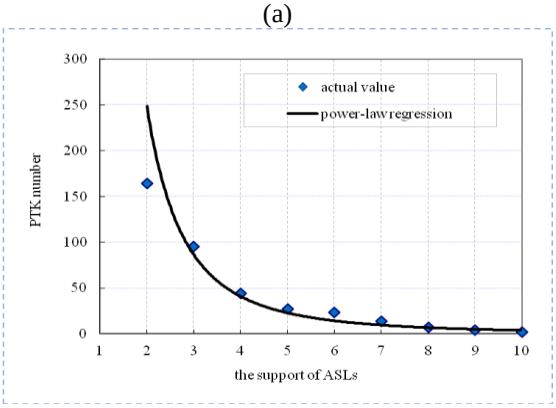
1: set the support of ASLs as 2
2: get the ASLs from the given semantic representation of Web resources by the method of ASL.
3: counting the number of four kinds of keywords occurring in these Web resources, and related ASLs;
4: if the number of the gotten ASLs is 0, then goto step 6;
5: else adding the support of ASL, and goto 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, so the support can be simplified as an integer variable.

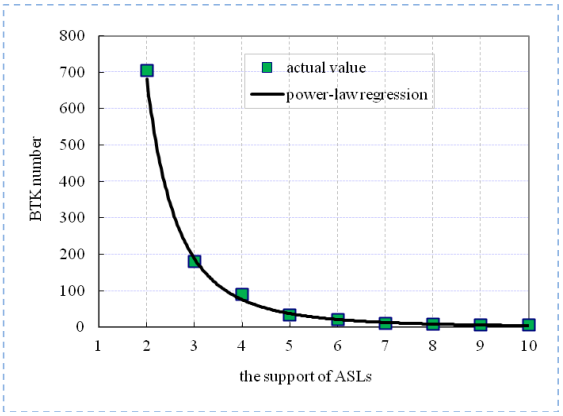
According to this adjusting process, the distribution of four kinds of keywords, the related ASLs and their regression results are plotted as Figure 3.



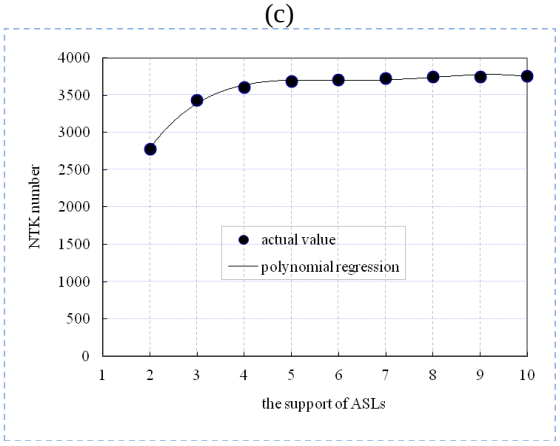
(a)



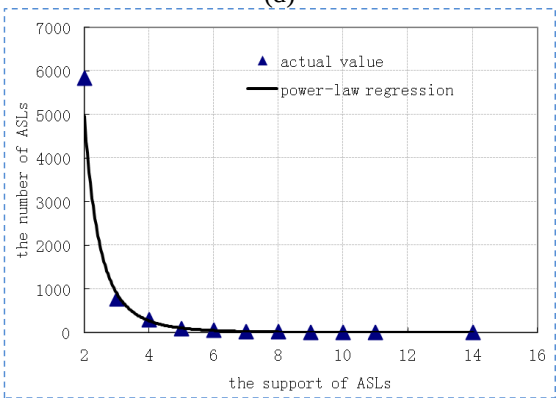
(b)



(c)



(d)



(e)

Figure 3. The distribution of four kinds of keywords and related ASLs

In our 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. The multinomial regression is selected to analyze the distribution of NTK. From Figure 3, we find that the distributions of ATK, PTK, BTK and ASLs follow power-law function. And the distribution of NTK follows the four times multinomial function.

Further, the Web resources of Internet and environment domain are selected to analyze the distribution of four kinds of keywords and related ASLs. The achieved parameters of regression analysis are listed in Table I.

TABLE I. THE POWER-LAW EXPONENT AND ITS COMPLEX CORRELATION COEFFICIENT OF REGRESSION ANALYSIS ON FOUR DIFFERENT SUPPORT

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

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

IV. THE LAYERED THEORY OF ASSOCIATION SEMANTIC

In this section, we first give the basic idea of the layered theory. Then, the layered theory based on semantic concentric degree is presented.

A. The basic idea of the layered theory

From the result of regression analysis in section 3, all the keywords with semantic traction feature, ATK, PTk and BTK, follow power-law distribution. At the same time, the keywords and related ASLs are becoming less with the occurring of larger support of ASLs. This means that some "basic semantic" are 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 4(a) to Figure 4(b)). Further, these concentric circles can be simplified/reduced as Figure 4(c). After that, the simplified concentric circles can be mapped into multi-layer semantic shown as Figure 4(d).

Definition 7: Three-layer Association Semantic(TL-AS) 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 simplified them as Figure 4(c). Then, three-layer association semantic can defined as the following.

- G_1 can be named as "basic semantic" of the Web resources on the given time window. This corresponds to the outer of concentric circles in Figure 3(c).
- G_i can be named as "main semantic" of the Web resources on the given time window. This corresponds to the middle of concentric circles in Figure 3(c).

- G_j can be named as "core semantic" of the Web resources on the given time window. This corresponds to the inner of concentric circles in Figure 3(c).

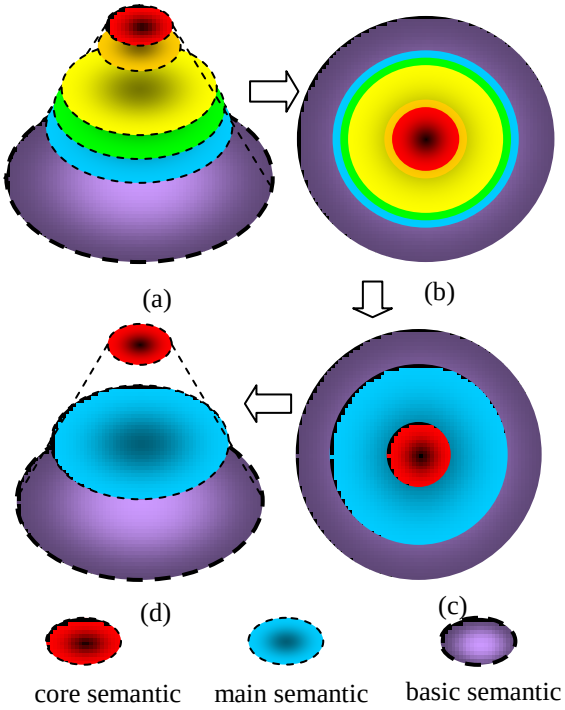


Figure 4. The basic idea of the layered theory

B. The layered theory based on semantic concentric degree

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

Definition 8: Association Semantic Concentricity (ASC)

Association semantic concentric is the repeatability score of association semantic containing 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

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 the Definition 7, for two supports fre_i and fre_j, if there is $fre_i < fre_j$.Then, based on generating method of k-ALNs, there be

$$G_j \subseteq G_i \quad (2)$$

Correspondingly, we have,

$$R_{Gj} \le R_{Gi} \quad (3)$$

That is,

$$ASC(G_i, G_j) \le 1 \quad (4)$$

In addition, for any an ALN, we have

$$R_{Gj} > 0 \quad (5)$$

So,

$$ASC(G_i, G_j) > 0 \quad (6)$$

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

$$0 < ASC(G_i, G_j) \le 1 \quad (7)$$

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 supports fre, ALNs $G_1, G_2 \cdots, G_m$, can be gotten. Larger exponent of power-law distribution means that 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,

$$ASC(G_{i-1}, G_i) = R_{Gi} / R_{Gi-1}$$
$$= \sqrt{a * fre_i^{-b} / \pi} \Big/ \sqrt{a * fre_{i-1}^{-b} / \pi} = [(i-1)/i]^{b/2} \quad (8)$$

Similarly,

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

Therefore,

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

That is,

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

Thus Corollary 3. is proved.

V. EXPERIMENT & RESULT ANALYSIS

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

A. Experimental data

Three domain news data are selected from Website http://www.reuters.com, including health, environment and internet, to building Web keywords-level ALN 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 association semantic link on a adjustable support. The initial value of the adjustable support is set as 2. Then, it is increased gradually. The increasing step is 1. The association semantic concentric of any two adjacent layer keywords-level ALNs are been computed by the definition 7. The computed results are listed in Table II.

B. The layered theory based on semantic concentric degree

In Table II, the column "*sup*" denotes two supports of two adjacent layers. "*H-ASC*" , "*I-ASC*" and "*E-ASC*" respectively denote the association semantic concentric of health, Internet and environment news domains. And the average value of association semantic concentric of each domain are listed in the last row of Table II.

From Table II, we can conclude the following conclusions:

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

TABLE II. THE ASSOCIATION SEMANTIC CONCENTRIC 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**

VI. CONCLUSIONS

The multi-layer theory of association semantic not only can provide theoretical support for knowledge services such as knowledge discovery in Web resources, knowledge recommendation with different particle size but also can improve the searching efficiency of searching system. Our contributions are as the follows:

(1) We have discovered the power-law distribution of four types of keywords with different linking role. According to the role of association semantic link, two kinds of semantic features are defined. Further, association semantic links are with 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.

(2) We have built the multi-layer theory of association semantic. Based on the power-law distribution of association keywords, two corollaries are presented. One is 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. The other is about changing trend of ASC among any two adjacent layers.

ACKNOWLEDGMENT

This work was supported by the Natural Science Foundation of Anhui Province Universities (No. KJ2015A111), in part by the National Science and Technology Major Project under Grant 2013ZX01033002-003, in part by the National Science Foundation of China under Grant 61300202, and in part by the Science Foundation of Shanghai under Grant 13ZR1452900.

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, K. Andrew, Introduction to the thematic issue on Ambient and Smart Component Technologies for Human Centric Computing. JAISE 6(1): 3-4, 2014.