Power Series Representation Model of Text Knowledge Based on Human Concept Learning

Xiangfeng Luo, Jun Zhang, Feiyue Ye, Peng Wang, and Chuanliang Cai

***Abstract*—How to build a text knowledge representation model, which carries rich knowledge and has a flexible reasoning ability as well as can be automatically constructed with a low computational complexity, is a fundamental challenge for reasoning-based knowledge services, especially with the rapid growth of web resources. However, current text knowledge representation models either lose much knowledge [e.g., vector space model (VSM)] or have a high complex computation [e.g., latent Dirichlet allocation (LDA)]; even some of them cannot be constructed automatically [e.g., web ontology language, (OWL)]. In this paper, a novel text knowledge representation model, power series representation (PSR) model, which has a low complex computation in text knowledge constructing process, is proposed to leverage the contradiction between carrying rich knowledge and automatic construction. First, concept algebra of human concept learning is developed to represent text knowledge as the form of power series. Then, degree-2 power series hypothesis is introduced to simplify the proposed PSR model, which can be automatically constructed with a lower complex computation and has more knowledge than the VSM and LDA. After that, degree-2 power series hypothesis-based reasoning operations are developed, which provide a more flexible reasoning ability than OWL and LDA. Furthermore, experiments and comparisons with current knowledge representation models show that our model has better characteristics than others when representing text knowledge. Finally, a demo is given to indicate that PSR model has a good prospect over the area of web semantic search.**

***Index Terms*—Cognitive informatics, human concept learning, knowledge representation, semantic search, text understanding.**

I. Introduction

CONSTRUCTING a text knowledge representation model, which carries rich knowledge and has a flexible reasoning ability as well as can be automatically constructed with a lower computational complexity, is a fundamental challenge for reasoning-based web knowledge services (e.g., web semantic search). And especially, it is demanded to be updated with the rapid growth of web resources. With the help of a text knowledge representation model, plenty of applications can be carried out, such as news recommendation [1], web services [2], the construction of association link network [3], and semantic web [4]. Many scholars have attempted to make their researches on this issue, and have gained many achievements [6]–[21]. Generally speaking, there are four main types of knowledge representation models as follows.

1) Statistics models, such as a vector space model (VSM) [5] and latent semantic Indexing (LSI) [6]. The VSM describes text resource only with separate terms, but does not consider the relationships between terms; therefore, this model only can express plain knowledge of text. As to reduce the dimensions of text features (i.e., bag of terms), LSI is proposed, which obtains a high consumption of computing for singular value decomposition. LSI replaces individual terms (i.e., keywords) as the descriptors of documents by independent "artificial concepts" that can be specified by any one of several terms [6]. Therefore, LSI is not in line with the reasoning-based web knowledge services, especially with the rapid growth of web text resources.
2) Cognition-based models, such as an element fuzzy cognitive map (EFCM) [7], [8], and a concept algebra-based model [9], [10]. Taking EFCM as an example, although EFCM improves the ability of text representation models in carrying more knowledge than the VSM and LSI, it does not have richer knowledge than web ontology language (OWL).
3) Probability topic models, such as probabilistic latent semantic indexing [11], latent Dirichlet allocation (LDA) [12], an author topic model (ATM) [13], an author-recipient-topic model [14], correlated topic models (CTMs) [15], etc. Obviously, probability topic models have solid foundation of mathematics, but they possess high complicated computation when the number of dynamic web text resources becomes larger. Taking LDA as an example, it is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics [12]. The key reasoning problem in LDA is that of computing the posterior distribution of the hidden variables, which is intractable for exact reasoning. Therefore, a wide range of approximate reasoning algorithms are considered for LDA, which include Laplace approximation, variational approximation, and Markov chain Monte [12], whose computational amount is much higher than EFCM and LSI.
4) Ontology-based models, such as ontology inference layer (OIL) [16], OWL [17], temporal OWL [18], and simple html ontology extensions [19]. Currently, ontology-based

Manuscript received January 16, 2012; revised May 23, 2012 and September 4, 2012; accepted November 7, 2012. Date of publication July 29, 2013; date of current version December 20, 2013. This work was supported by the National Science Foundation of China under Grant 91024012, Grant 61071110, and by the Shanghai Leading Academic Discipline Project under Grant J50103. This paper was recommended by Associate Editor J. A. Keane.

X. Luo, J. Zhang, F. Ye, and C. Cai are with the High Performance Computing Center, School of Computing Engineering and Science, Shanghai University, Shanghai 200444, China (e-mail: luoxf@shu.edu.cn; zhangjun_haha@shu.edu.cn; yefy@shu.edu.cn; caichuanliang@shu.edu.cn).

P. Wang was with the Jiangnan Institute of Computer Technology, Wuxi 214083, China. He is now with the National High-Performance IC Design Center, Shanghai 201204, China (e-mail: wangpeng_shu@shu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCC.2012.2231674

models are very popular and a lot of applications are carried out, for an instance, personalized web information gathering, which is discussed in [20]. These models really contain rich knowledge. Taking OWL as an example, the components in OWL, referring to classes, properties, instances of classes, and relationships between these instances, are hard to be obtained without human efforts since there are many restrictive syntactic constrains that result in its inconvenient construction. Therefore, it is obvious to get that the construction of the OWL model is semiautomatic. With the help of human efforts to construct OWL, reasoning in it can offer us rich knowledge, such as consistency, subsumption, equivalence, etc. One possible defect of ontology-based models is that the knowledge of web resources cannot be automatically obtained by a machine. Therefore, this type of models does not leverage the contradiction between carrying rich knowledge and constructed automatically in the text knowledge representation process.

From the previous analysis, we know that the previous models may involve complicated computation (e.g., CTM, LDA), or lose many text knowledge (e.g., VSM), or cannot be constructed automatically (e.g., OIL, OWL), and even lack the ability of flexible knowledge-based reasoning (e.g., LDA). We know that a text can be regarded as a concept, and the sentences that are contained in the text can be regarded as objects, where the keywords/terms are the attributes of these objects [21]. Based on the concept algebra [22], a power series model of text knowledge representation [power series representation (PSR)] is proposed in this paper, which does not only carry more abundant text knowledge (e.g., relations in web page), but also can be constructed automatically. Thus, the PSR model adapts to the rapid growth of web resources and holds more rich knowledge and a flexible reasoning ability than EFCM and LSI, and has lower computational complexity than LDA and OWL.

This paper is organized as follows. The text representation based on the linearity hypothesis of human concept learning will be introduced in next section. The PSR model will be proposed in Section III. Then, the constructing processes of the PSR model and its reasoning ability will be discussed in Sections IV and V, respectively. In Section VI, an extension of the PSR model based on degree-2 hypothesis is given. Section VII gives experiments to verify that the PSR model has some special characteristics than others whereas Section VIII presents a PSR-based web semantic search system. Finally, conclusion is given in the last section.

II. Text Representation Based on Linearity Hypothesis of Human Concept Learning

A. Concept Algebra and Linearity Hypothesis of Human Concept Learning

Based on empirical evidence, Feldman proposed the *linearity hypothesis* which means human learners are biased toward linear concepts [22], [23]. The linearity here is not directly related to the notion of linear separability of categories, which refers to the separability of positive and negative examples by a hyperplane.

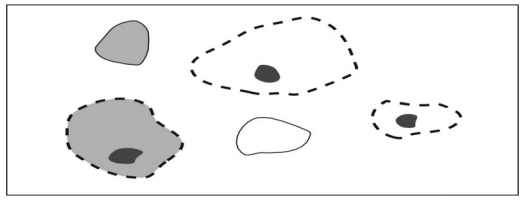


Fig. 1. "World" W containing amoeba-like objects [22].

Linearly separable categories have generally not been found to be easier for humans to learn [24]; although for a contrary view, see [25]. However, it is found that the linearity hypothesis accords with the cognitive behaviors of human beings [26], [27].

As for concept learning, many research findings are given. For example, Boolean concepts [10], prototype-based or exemplar-based models [28], [29], etc. Recently, Yao [30] proposed a conceptual framework for concept learning from the viewpoints of cognitive informatics and granular computing. Herein, we mainly focus on Feldman's work.

Feldman used Fig. 1 as an example. In this figure, the "world" W that consists of the five amoeba-like objects corresponds to a concept of a human concept learning process. A "language" has been chosen to express the structure of this "world" W, which is as simple as a list $\sum = \{\sigma_1, \sigma_2, \ldots, \sigma_d\}$ of abstract property tags, and called the property language [22]. For the given "world," an appropriate language might be {*blob_shaped, shaded, has_nucleus, has_dotted_membrane, large*}, which is abbreviated to $\sum = \{a, b, c, d, e\}$ under the assignment

$a =$ *blob_shaped;*
$b =$ *shaded;*
$c =$ *has_nucleus;*
$d =$ *has_dotted_membrane;*
$e =$ *large,*

and the "world" W encoded into $\sum$ is denoted by $W|_{\sum}$.

The previous example can be described as follows according to Feldman's work [22], [23]:

$\alpha = \{a\}$
$\beta = \{b\}$
$\omega = \{e \rightarrow c,\ c \rightarrow d,\ d \rightarrow c\}$.
$W|_{\sum} \subseteq \{a\} \cdot [\{e \rightarrow c,\ c \rightarrow d,\ d \rightarrow c\} \times \{b\}]$

where α is for the set of constant properties, i.e., the affirmations in "world" W; ω for the set of paired implications; and β for the rest of "world" W, the unconstrained properties.

Every possible "world," which can be produced by these types of concepts alone, can be expressed as the Cartesian product of the lattice for ω with the lattice for β, conjoined with the properties α [22], [23], that is

$$\alpha \cdot [\omega \times \beta], \quad W|_{\sum} \subset \alpha \cdot [\omega \times \beta]. \tag{1}$$

Equation (1) illustrates that human concept learning is a linear process.

B. Text Representation Based on Linearity Hypothesis of Human Concept Learning

Generally speaking, text knowledge is composed of sequential sentences, and a sentence is composed of some terms'

(i.e., keywords) combinations. When we regard the terms as attributes and the sentences as objects in a text, text understanding process can be regarded as a linear learning process of concept. Moreover, in consideration of the object-attribute-relation model [31]–[33], a text can be regarded as a concept, whereas sentences that are contained in the text can be regarded as objects belonging to the text and the terms in sentences are the attributes of these objects.

In this paper, we regard the terms as attributes of describing text knowledge, a sentence or paragraph as an object, therefore a text which consists of many sentences and paragraphs can be regarded as a concept. Then, we represent the text knowledge according to the linearity hypothesis of human concept learning. Thus, based on the similarity between the concept and the text, an example about text fragments is given as follows.

$\mathbf{S_1}$: That *boy* stands on the *left*, whose *t-shirt* is *red*.

$\mathbf{S_2}$: Two *girls* stand on the *right*, whose *skirts* are also *red*.

Referring to the previous discussions and the linearity hypothesis of human concept learning, we find a property language $\sum_{\text{text}}$, which includes all the properties from the text fragment including $\mathbf{S_1}$ and $\mathbf{S_2}$.

$\sum_{\text{text}} = \{$*boy, left, red, girl, right, t-shirt, skirt*$\}$, and

$\sum_{\text{text}} = \{a, b, c, d, e, f, g\}$ under the following assignments,

$a =$ *boy*, $b =$ *left*, $c =$ *red*, $d =$ *girl*, $e =$ *right*, $f =$ *t-shirt*, $g =$ *skirt*.

After analyzing the text fragment, we can find that the clothes of all people including one *boy* and two *girls* are *red*. And it is apparent that the person who is *boy* must be *left* and wears a *t-shirt*, while the person who is *right* must be a girl and wears a *skirt*. Therefore, the previous text fragment can be represented as

$\alpha = \{c\}$ $\beta = \{\}$

$\omega = \{a \rightarrow b, a \rightarrow f, d \rightarrow e, d \rightarrow g\}$

$$W|_{\sum \text{text}} = \{c\} \cdot [\{a \rightarrow b, a \rightarrow f, d \rightarrow e, d \rightarrow g\} \times \{\}].$$

Referring to the *linearity hypothesis* of human concept learning [22], [23], different knowledge from a text can be regarded as different attribute sets in human concept learning.

Definition 2.1 (Text assertions): The keywords/terms that stand for the common knowledge in a text are referred as text assertions.

Text assertions have the simplest and the most easily understood information in a text. Moreover, in this paper, text assertions are also defined as the keywords/terms that often appear in a text. In another word, the terms, whose term frequences are very high, will be seen as text assertions. It is because that such terms standing for the common sense knowledge of a certain text always appear frequently in it. For example, given a text focusing on internet, "*internet*" and "*computer*" are appropriate to be seen as the common sense knowledge of this text, besides each of them obtains a high term frequence and, thus, will be considered as text assertions in a text knowledge representation process.

Definition 2.2 (Text association rules): Text association rules are referred to these terms' relations that are causations and reflect the semantic relationships in a text.

Generally, text association rules are mined from those terms that are adjacent to each other and own a high co-occurrence appearance in a text.

According to human concept learning [22], different text association rules are represented as the following forms:

$$a \xrightarrow{r_1} b; \quad ab \xrightarrow{r_2} c; \quad \cdots; \quad \overbrace{ab \cdots}^{k} \xrightarrow{r_i} m$$

where $a, b, c, \ldots, m$ are terms in a text; k is the number of the antecedent of an implication; $r_1, r_2, \ldots, r_i$ denote corresponding confidences of the association rules.

According to the *linearity hypothesis* of human concept learning [22], there are some descriptions as follows.

α: $\{\}$ denotes text assertions that correspond with the affirmation (i.e., assertion keyword) in human concept learning, which contain the common sense belonging to a text.

ω: $\{\}$ denotes text association rules that correspond with the extensive implication in human concept learning, which contain the relationships among terms that are extracted from text.

β: $\{\}$ denotes isolated terms that have the same meaning in human concept learning, which are not related with other terms in text and need background knowledge to understand them. Therefore, β affects the understanding difficulty of a text.

According to [22], we utilize the previous three sets to represent text knowledge as $\alpha \cdot [\omega \times \beta]$, where $\cdot$ denotes adding α to every element in the set ($[\omega \times \beta]$) and $\times$ denotes the Cartesian product of the two sets (ω and β). If β is empty, $\omega \times \beta$ equals to ω, which means that no isolated terms affect the relationships set ω.

For the text fragment $\mathbf{S_1}$ and $\mathbf{S_2}$, if we regard the text fragment as a text T and suppose that $\{$*boy, left, red, girl, right, t-shirt, skirt*$\}$ is the set of terms of this text, the knowledge of this text can be represented as follows:

α: $\{$*red*$\}$;

ω: $\{$*boy*$\rightarrow$*left, boy*$\rightarrow$*t-shirt, girl*$\rightarrow$*right, girl*$\rightarrow$*skirt*$\}$;

β: $\{\}$.

Therefore, the text can be represented as

$$W|_{\sum} = \alpha \cdot [\omega \times \beta] = \{red\} \cdot [\{boy \rightarrow left,\ boy \rightarrow t\text{-}shirt,\ girl \rightarrow right,\ girl \rightarrow skirt\} \times \{\ \}].$$

From the previous qualitative analysis, we know that the knowledge representation model of concept algebra has good capability to describe text knowledge.

III. Power Series Representation Model of Text Knowledge

A. Extended Concept Algebra of Human Concept Learning

Based on the linearity hypothesis of human concept learning in [22] and [23], Feldman regards a "world" or a "concept" with affirmation and implication; that is to say, the "world" or "concept" is described by linear terms. However, there are many higher order terms in reality; therefore, he extended the concept algebra to more general hierarchy.

According to the definition of *implication* in [22], the "implication" is linear and means paired causality. However, there are

higher order causal regularities among more than two attributes in real world. Therefore, the paired causality should be extended to a general form. The general form of extensive implication can be defined as

$$\sigma_1\sigma_2\ldots\sigma_k \rightarrow \sigma_0$$

where the subscript index k here, called the degree of the implication, is the number of properties that are antecedent of the implication, which serves as a measure of the complexity of the regularity.

Obviously, the paired implication $\sigma_1 \rightarrow \sigma_2$ is the special form of the general case. After defining the extensive implication, a "world" or a "concept" set can be described more fruitfully.

For an object set x, the set of implication of degree k, that is contained in the series expansion of x, is denoted as $\mathbf{\Phi}_x^k$. And the full power series for observation x [denoted as $\phi(x)$] is as follows:

$$\begin{aligned}&0:\mathbf{\Phi}_x^0(\text{minimum degree})\\&1:\mathbf{\Phi}_x^1\\&\ldots\\&D-1:\mathbf{\Phi}_x^{D-1}(\text{maximum degree}).\end{aligned}$$

For any object set x that is defined over a language $\sum$ of size D, there is a minimal irredundant set of implications

$$\begin{aligned}W|_{\sum} &= \mathbf{\Phi}_x^0\mathbf{\Phi}_x^1\cdots\mathbf{\Phi}_x^{D-1}\\&= \bigwedge_{k=0}^{D-1}\mathbf{\Phi}_x^k\\&= \wedge\phi(x).\end{aligned} \tag{2}$$

Comparing the simple algebra that is discussed in Section II with the full power series shown in (2), we know that the former is only able to express linear concepts while the latter can describe any discrete-featured patterns. On the web, there are a huge number of web pages, which are discrete and involved with all kinds of relationships; therefore, it is an important task to distinguish different terms and represent the relationships among them. Considering that there are some similarities between the concepts and the web pages, we try to study the representation of text knowledge (e.g., web pages) based on human concept learning. The PSR model of text knowledge will be discussed in the next section.

B. Power Series Representation Model of Text Knowledge

In order to describe the PSR model of text knowledge conveniently, some definitions will be given in advance.

Definition 3.1 (Degree of text association rule): The number of text association rule's antecedent keywords/terms is defined as the degree of text association rule.

For example, the degree of $a \rightarrow b$ is 1; the degree of $a, b \rightarrow c$ is 2.

From the viewpoint of human concept learning [20], the degree of implication reflects the understanding difficulty of concepts and relationships among them. Similarly, the degree of text association rule measures the complexity of relationships among keywords/terms. The higher the degree of text association rule, the more complicated the text association rule, and vice versa. For instance, there are two text association rules in a text on the domain of web services

$$\begin{aligned}&\textit{WSDL} \rightarrow \textit{machine} - \textit{readable language}\\&\textit{WSDL}, \textit{UDDI}, \textit{SOAP} \rightarrow \textit{Service framework}.\end{aligned}$$

The degree of the former is 1, and that of the latter is 3. It is obvious that the latter is more difficult to be understood than the former, because there are more complicated association among *WSDL, UDDI, SOAP*, and *Service framework*. Specially, the largest degree of the text that contained D-terms is $D-1$. Text assertion is a special text association rule whose degree is 0.

Definition 3.2 (Degree-k text association rule set): One text T, which contains keywords/terms $\sum = \{K_1, K_2, \ldots, K_D\}$, is denoted as $T|_{\sum}$. And the set of degree-k text association rules is denoted as $\mathbf{\Phi}_x^k$, where $x \subseteq \sum$.

In Section III-A, the concept is expanded as the form of power series. Similarly, we can expand the text association rules as following definitions.

Definition 3.3 (Power series expansion of text association rules): Assuming a text that contains D keywords/terms, power series expansion of text association rules is a set of all text association rules whose degrees are from 0 to D–1. Referring to the expansion of power series of human concepts, power series expansion of text association rules is represented as

$$\begin{aligned}&0:\mathbf{\Phi}_x^0(\text{text assertions})\\&1:\mathbf{\Phi}_x^1\\&\ldots\\&k:\mathbf{\Phi}_x^k\\&\ldots\\&D-1:\mathbf{\Phi}_x^{D-1}(\text{text association rules of the largest degree}).\end{aligned}$$

Definition 3.4 (Power series representation model of text knowledge): Referring to the concept algebra of human concept learning [22], [23], there is a set of text association rules that correspond to any text T; therefore, there is the model

$$T|_{\sum} = \mathbf{\Phi}_x^0 \wedge \mathbf{\Phi}_x^1 \wedge \cdots \wedge \mathbf{\Phi}_x^{D-1} = \bigwedge_{k=0}^{D-1}\mathbf{\Phi}_x^k \tag{3}$$

where $\sum$ denotes all the keywords/terms in text T; D denotes the number of terms; and $\mathbf{\Phi}_x^k$ denotes a set of *degree-k* text association rules in $x \subseteq \sum$.

After analyzing the PSR model of text knowledge, we find it complicated to represent text knowledge. Therefore, it is an important task to find some special features of text to simplify it. The simplicity of the PSR model of text knowledge will be discussed in Section IV-D.

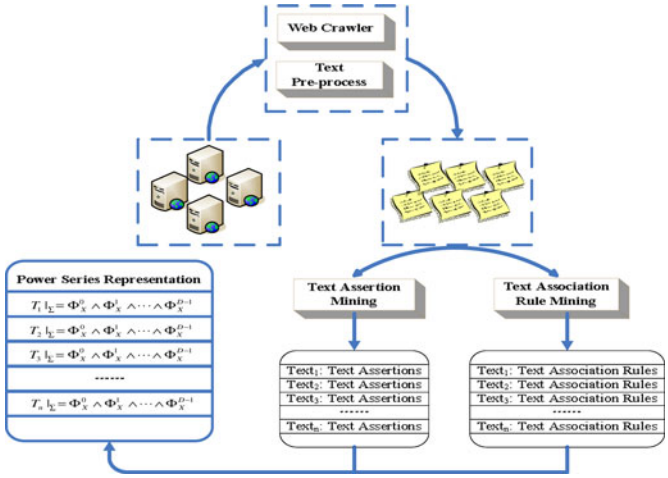


Fig. 2. Whole process of constructing the PSR model of text knowledge.

IV. Constructing the Power Series Representation Model of Text Knowledge

A. Main Tasks in the PSR Constructing Process

In Section III, the PSR model of text knowledge has been proposed; therefore, how to automatically construct it becomes a very important issue.

In our model, we find that the construction process of power series model contains two tasks: one is the extraction of text assertions and the other is the extraction of text association rules. Thus, in this section, we will mainly focus on these two mining algorithms. The whole process of constructing the PSR model is illustrated in Fig. 2. As shown in this figure, we first download text resources from websites such as Reuters[1] and Yahoo.[2] Then, we use a part-of-speech tagging tool to parse the terms in those texts, and only nouns or noun phrases are kept. After that, based on the existing word extraction algorithms such as term frequency–inverse document frequency [5], we only extract meaningful terms (i.e., keywords) to represent the texts. Furthermore, we will put up a further mining task on texts, including two parts, i.e., 1) text assertion mining; and 2) text association rule mining. Finally, these two parts will be combined to make up the PSR model for text knowledge.

Moreover, as considering the mining complexity of text association rule, we also propose a hypothesis that is called degree-2 hypothesis to simplify the PSR model. Based on the degree-2 hypothesis, the complexity of constructing power series model will be greatly reduced.

B. Constructing Algorithm of Text Assertions

As defined in Definition 2.1, text assertions are certain terms that stand for the common knowledge in a text, and moreover, they are also defined as the terms that often appear in the text. In other words, this paper will take those terms as text assertions, whose term frequencies in a text are higher than a given threshold. According to this definition, the constructing algorithm of text assertions is given as follows.

1) For every text, calculate each term's frequency, by which the terms are ranked in descending order.
2) According to Salton *et al*. [5], the terms with the highest term frequencies are always the top 4% in the ranked list. Consequently, on the basis of step 1, we will select the top 4% terms in the ranked list as text assertions for each text.

To better understand the construction process of text assertions, a simple example is given as follows.

Assuming that a text contains 50 terms, and according to those terms' frequencies, they are listed as {Sun: 43; apple: 35; Schaeffer: 19; policymaker: 18; warfighter: 17; guideline: 14; NSA: 14; harden: 12; assurance: 10; infrastructure: 7 . . .}. Elements (e.g., Sun: 43) in this list are composed of two parts: the former refers to the term (e.g., Sun) while the latter means the frequency of the term (e.g., 43). Based on this list, we will only choose the top 4% terms as text assertions, which are the sun and apple. In other words, because of the highest appearances of sun and apple among 50 terms, they are regarded as text assertions in the text.

C. Constructing Algorithm of Text Association Rules

There are mature algorithms to obtain the association rules such as [36] and [37]. Cognitive scientists find that the sentences are regarded as the basic unit when human reads texts; three concepts are referred when human beings begin to read texts [34], [35], which are 1) back browsing, i.e., browsing the previous text knowledge again when needed; 2) preselection of interested area, i.e., sentence fragment length of human selecting to understand easily text knowledge; 3) adjustment of interested area, i.e., increasing or decreasing sentence fragment length according to reading experiences.

Therefore, from the viewpoint of cognitive scientists, it is necessary for us to set *sliding window* in the extracting process of text association rules to simulate the human reading process. When human beings read a sentence, they cannot understand the whole sentence at once, but they can process a part of a sentence, that is to say, there is a model to process the *interested area* (i.e., the size of sliding window). Therefore, the use of *sliding step* can simulate the human reading process. After reading two adjacent sentences, human will combine both sentences (back processing). The ideology of sliding window is an imitation to the human's reading process; it has a high reference value and theoretical basis in the text knowledge representation process.

In our experiments, we generate text association transactions by *sliding window*, which means that the terms from previous location to the current window location form a transaction, and then a single text is regarded as a set of transactions; therefore, we can mine the efficient association rules among terms. In addition, as the sliding window is an imitation of the human's reading process, most of the text knowledge should be contained in the extracted text association rules.

According to Miller's theory in short-term memory, i.e., "the magical number seven, plus or minus two" [38], the size of sliding window is set as 9. In another word, if the distance

[1][Online]. Available: http://www.reuters.com
[2][Online]. Available: http://www.yahoo.com

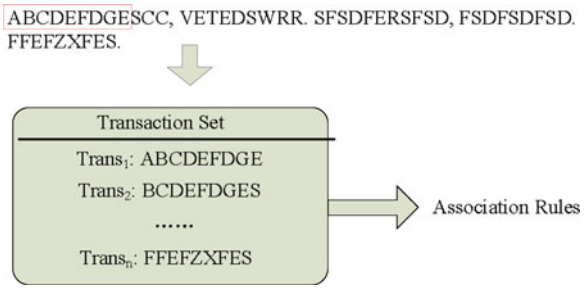


Fig. 3. Extracting method of association rules. Each character in this figure denotes a term.

between two terms is larger than 9, we will consider that they are two irrelevant terms.

Herein, we take an example to explain our experiment that is shown in Fig. 3, which reflects the extracting method of association rules. In this figure, the size of sliding window is 9, which means that one sliding window contains nine terms, and the sliding step is 1, which refers to the interval between two adjacent sliding windows. For a text T in a text set, based on the human reading process, the steps to discover association rules from text are shown as follows.

1) According to the extracting algorithm of keywords [36], keywords (i.e., terms) are extracted, which is denoted as $\{K_1, K_2, K_3, \ldots, K_D\}$.
2) Extraction of degree-k text association rules: According to the extracting algorithm of text association rules [37], the steps of the extraction of degree-k text association rules from text T is given as follows.
   a) Set initial thresholds for confidence and support, which are denoted as θ_1 and θ_2, respectively. Herein, the support θ_2 is referred to the absolute support, which is defined as the minimum times of two keywords coappearing.
   b) For each $K_i \rightarrow K_j \ \ (i, j \in \{1, 2, \ldots, D\}, i \neq j)$, we calculate the frequency of certain windows where K_i and K_j appear, and denote it as T_{ij}; the frequency of certain windows where K_i appears, is denoted as T_i; $K_i \rightarrow K_j$ can be regarded as a degree-1 text association rule, if $T_{ij}/T_i \geq \theta_1, T_{ij} \geq \theta_2$.
   c) For $K_iK_j \rightarrow K_u \ \ (i, j, u \in \{1, 2, \ldots, D\}, i \neq j \neq u)$, we calculate the frequency of certain windows where K_i, K_j, and K_u appear, and denote it as T_{iju}; the frequency of certain windows where K_i and K_j is denoted as T_{ij}; $K_iK_j \rightarrow K_u$ can be regarded as a degree-2 text association rule, if $T_{iju}/T_{ij} \geq \theta_1, T_{iju} \geq \theta_2$.

Degree-k $(3 \leq k \leq D-1)$ text association rules can be extracted with the previous same steps.

D. Degree-2 Hypothesis—Simplicity of the Power Series Representation Model

According to the mining algorithm of association rules from a text that was discussed in Section IV-C, several distributions on text association rules with different relation degrees are shown

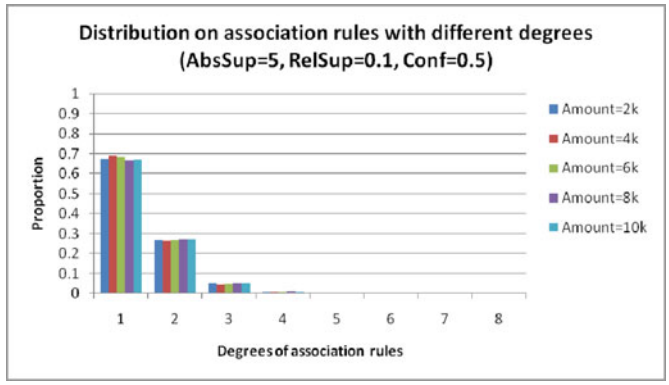


Fig. 4. Distribution on association rules with different degrees under different scales of texts. AbsSup (Absolute Support) is used to count the co-occurrences of two terms. RelSup (Relative support) represents the ratio of the co-occurrences of two terms to the number of transactions. Conf (Confidence) reflects the probability of the occurrence of term A under the condition that term B has appeared. The amount refers to the different scales of texts, which are 2000, 4000, 6000, 8000, and 10 000.

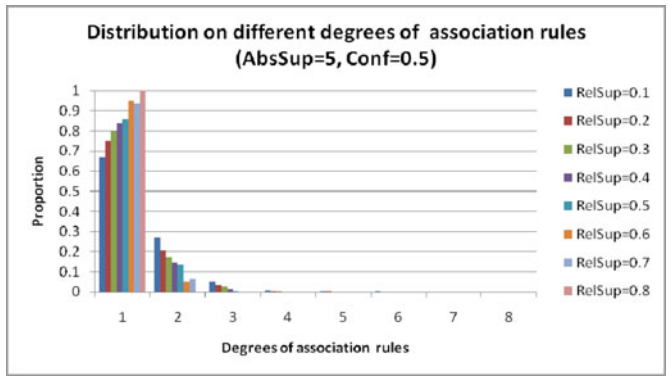


Fig. 5. Distribution on association rules with different degrees under different relative support thresholds.

in Figs. 4 and 5. For avoiding the noises that are carried by short-length texts, we fix a threshold called absolute support, denoted as *AbsSup*, which is used to count the co-occurrences of two terms in all transactions, which are discussed in Section IV-C. Meanwhile, relative support, denoted as *RelSup*, is used to represent the ratio of co-occurrences of two terms to the number of transactions. Confidence, denoted as *Conf*, reflects the probability of the occurrence of term A under the condition that term B has appeared.

Fig. 4 shows the distributing situation of text association rules with different relation degrees under different scales of texts. In Fig. 4, the numbers of experimental texts are 2000, 4000, 6000, 8000, and 10 000. All the texts belong to the *environment news* that crawled from *reuters.com*. The x-axis refers to the degrees of text association rules, while the y-axis means their accordingly proportions in all of the text association rules.

The values of *AbsSup*, *RelSup,* and *Conf* mean that identifying a frequent item whether a text association rule depends on such a condition that its absolute support, relative support, and confidence are no less than 5, 0.1, and 0.5 (see Fig. 4), respectively. It can be seen that among all of the text association rules the proportion of degree-1 and degree-2 text association

rules is more than 90% while the others only occupy less than 10%. Therefore, it can be concluded that under different scales, degree-1 and degree-2 text association rules contain more than 90% text knowledge.

Similarly, to further validate our hypothesis, Fig. 5 is given to show that under the same confidence, how the text association rules distribute with the variation of relative support. And the result is apparent that no matter what value relative support is set, the total proportion of degree-1 and degree-2 text association rules is higher than 90% all along.

From Figs. 4 and 5, we shall conclude that among all of the text association rules, degree-1 and degree-2 text association rules make up most of the knowledge in a text. We find that degree-1 and degree-2 text association rules are the main component of all the text association rules. Therefore, the degree-2 hypothesis is given as follows.

Hypothesis 4.1 (Degree-2 hypothesis of the PSR model): If one text is composed of text assertions, degree-1 and degree-2 text association rules, in another word, if one text can be represented by $\mathbf{\Phi}_x^0 \wedge \mathbf{\Phi}_x^1 \wedge \mathbf{\Phi}_x^2$, the most knowledge of the text has already been contained in them.

Base on the previous hypothesis, the PSR-based *degree*-2 *hypothesis* model of text knowledge representation can be simplified as

$$T|_{\sum} = \mathbf{\Phi}_x^0 \wedge \mathbf{\Phi}_x^1 \wedge \mathbf{\Phi}_x^2. \tag{4}$$

Compared with *linear hypothesis* [22], *degree*-2 *hypothesis* is the expansion of *linear hypothesis* in human concept learning. *Linear hypothesis* reflects the nature language constrains that is affected by the nature of human intelligence and cognitive behaviors. The complex relationships among the concept's attributes are hard to be obtained by human beings because of some physiological and psychological restrictions such as brain storage capacity, logical thinking ability, etc. In fact, for the majority of concept learning process, it is sufficient to obtain a concept's knowledge with degree-1 relationships between attributes because human beings have the associated ability and the prior knowledge to explain the concept's context [39] with the help of degree-1 association rules. However, for a machine, as there is no storage capacity and logical thinking limitation, it is able to obtain more complex semantic relationships among keywords/terms, which makes up the lack of associated ability and prior knowledge in the text machine understanding process.

V. Reasoning Based on Degree-2 Hypothesis of Text Knowledge

Degree-2 *hypothesis of* the PSR model does not only simplify the PSR model of text knowledge, but also makes the PSR model to have a flexible reasoning ability. In this section, the PSR-based reasoning operations of text knowledge will be discussed based on this hypothesis, including three parts: basic reasoning, general extended reasoning, and advanced extended reasoning.

A. Basic Reasoning

We propose three basic reasoning operations, which are *intersection* operation, *union* operation, and *subtraction* operation between two texts. The common information can be reasoned by the *intersection* operation; the whole information among texts can be collected by the *union* operation; and the difference between two texts can be found by the *subtraction* operation between texts. Based on the three basic operations, we can not only deduce the latent information of texts, but also consider more complicated reasoning rules (e.g., the extended rules in Section V-B). In this aspect, our model performs better than other models such as the VSM, EFCM, and OWL. The VSM has a weak reasoning ability because this text representation model is only consisted of terms rather than the relations between them. EFCM just considers the degree-1 association rules; therefore, the reasoning rules of this model are too simple. Ontology-based models (e.g., OIL, OWL) can express rich knowledge, but the reasoning rules of these models are too strict, which cannot meet with the diversity and irregularity of web resources. Probability topic models (e.g., ATM, CTM) are based on probability and statistic, with many mathematical assumptions pregiven (e.g., Laplace approximation, and Markov chain Monte). However, these assumptions are not suitable, since the web resources are massive, dynamic, and uncertain. In our model, text knowledge can be represented automatically, and can accord with the mass, dynamic, and uncertainty of web resources. At the same time, the reasoning process is easier to be implemented in our model, which also contains richer knowledge.

Definition 5.1: The *intersection* operation of two texts $T_a = \mathbf{\Phi}_{x_a}^0 \wedge \mathbf{\Phi}_{x_a}^1 \wedge \mathbf{\Phi}_{x_a}^2$ and $T_b = \mathbf{\Phi}_{x_b}^0 \wedge \mathbf{\Phi}_{x_b}^1 \wedge \mathbf{\Phi}_{x_b}^2$, which stands for the information that appears in both two texts at the same time, is denoted as $I\ (T_a, T_b)$

$$\begin{aligned} I(T_a, T_b) &= (\mathbf{\Phi}_{x_a}^0 \wedge \mathbf{\Phi}_{x_a}^1 \wedge \mathbf{\Phi}_{x_a}^2) \cap (\mathbf{\Phi}_{x_b}^0 \wedge \mathbf{\Phi}_{x_b}^1 \wedge \mathbf{\Phi}_{x_b}^2) \\ &= (\mathbf{\Phi}_{x_a}^0 \cap \mathbf{\Phi}_{x_b}^0) \wedge (\mathbf{\Phi}_{x_a}^1 \cap \mathbf{\Phi}_{x_b}^1) \wedge (\mathbf{\Phi}_{x_a}^2 \cap \mathbf{\Phi}_{x_b}^2) \\ &= \mathbf{\Phi}_{x_\cap}^0 \wedge \mathbf{\Phi}_{x_\cap}^1 \wedge \mathbf{\Phi}_{x_\cap}^2 \end{aligned}$$

where x_a is the set of terms of text T_a, x_b is the set of terms of text T_b, and $x_\cap = x_a \cap x_b$ is the intersection of x_a and x_b.

Intersection operation of two texts contains the common information, including text assertions and text association rules. In this paper, *intersection* operation will be discussed in three situations as the rest situations are not meaningful to the understanding of texts.

1) If $\mathbf{\Phi}_{x_\cap}^0$, $\mathbf{\Phi}_{x_\cap}^1$, and $\mathbf{\Phi}_{x_\cap}^2$ are $\emptyset$, it means that T_a is completely different from T_b.
2) If only $\mathbf{\Phi}_{x_\cap}^0$ and $\mathbf{\Phi}_{x_\cap}^1$ are not $\emptyset$, T_a and T_b probably belong to the same theme but focus on several different aspects, for example, as shown in Table I, it is obvious that $\mathbf{\Phi}_{x_\cap}^0$ and $\mathbf{\Phi}_{x_\cap}^1$ in the intersection of $text_a$ and $text_b$ are not $\emptyset$ while $\mathbf{\Phi}_{x_\cap}^2$ is $\emptyset$, in another word, it indicates that $text_a$ and $text_b$ are talking about similar contents. However, if there were more elements in $\mathbf{\Phi}_{x_\cap}^0$ and $\mathbf{\Phi}_{x_\cap}^1$, it would be that $text_a$ and $text_b$ were almost discussing the same topic;
3) If only $\mathbf{\Phi}_{x_\cap}^0$ is not $\emptyset$, T_a and T_b tend to have some kinds of relations between them but probably focus on different

TABLE I
TEXTS REPRESENTED BY A PSR-BASED DEGREE-2 HYPOTHESIS MODEL

Title	***Text$_a$:*** *Sun and Apple almost merged three times*	***Text$_b$:*** *Skepticism surrounds Apple, Sun report*	***Text$_c$:*** *Apple Trees: Where and How to Plant*
Source	*http://www.theregister.co.uk/2006/01/12/sun_apple_snapple/*	*http://news.cnet.com/Skepticism-surrounds-Apple,-Sun-report/2100-1033_3-202914.html*	*http://www.doityourself.com/stry/plantappletree*
Φ^0_x	*Apple; Sun; McNealy; joy*	*Sun; Apple; Herwick; analyst*	*tree; apple; soil; foot; root*
Φ^1_x	*time →Apple; cell →phone; Apple →Sun; co-founder →Sun; CEO →Apple; CEO →joy; phon →cell; deal →SPARC; iPod →McNealy; machine →answering; Sun →Apple; joy →Apple; answering →machine; SPARC →deal*	*Apple →Sun; culture →Apple; culture →Sun; analyst →Apple; analyst →Sun; company →Apple; company →Sun; Sun →Apple*	*tree →apple; year →tree; year →apple; foot →tree; size →tree; size →apple; apple →tree*
Φ^2_x	*Apple,CEO →joy; Apple,McNealy →Sun; CEO,joy →Apple; McNealy,Sun →Apple; Sun,joy →Apple*	*Apple,culture →Sun; Apple,analyst →Sun; Apple,company →Sun; culture,Sun →Apple; analyst,Sun →Apple; company,Sun →Apple*	$\emptyset$

themes, for example, in Table I, through the intersection of $text_a$ and $text_c$, only $\mathbf{\Phi}^0_{x_\cap}$ is not $\emptyset$. It means that they are talking two different topics, and actually, they do focus on different topics although "*apple*" is mentioned in both.

Definition 5.2: The *union* operation of two texts T_a and T_b, which stands for all the information that is included in two texts, is denoted as U (T_a, T_b).

$$\begin{aligned} U(T_a,T_b) &= (\mathbf{\Phi}^0_{x_a} \wedge \mathbf{\Phi}^1_{x_a} \wedge \mathbf{\Phi}^2_{x_a}) \cup (\mathbf{\Phi}^0_{x_b} \wedge \mathbf{\Phi}^1_{x_b} \wedge \mathbf{\Phi}^2_{x_b}) \\ &= (\mathbf{\Phi}^0_{x_a} \cup \mathbf{\Phi}^0_{x_b}) \wedge (\mathbf{\Phi}^1_{x_a} \cup \mathbf{\Phi}^1_{x_b}) \wedge (\mathbf{\Phi}^2_{x_a} \cup \mathbf{\Phi}^2_{x_b}) \\ &= \mathbf{\Phi}^0_{x_\cup} \wedge \mathbf{\Phi}^1_{x_\cup} \wedge \mathbf{\Phi}^2_{x_\cup} \end{aligned}$$

where x_a is the set of terms of text T_a, x_b is the set of terms of text T_b, and $x_U = x_a \cup x_b$ is the union of x_a and x_b.

When *intersection* reasoning between two texts is not $\emptyset$, *union* operation can extract comprehensive knowledge from two texts. Similarly, we also discuss three situations.

1) If only $\mathbf{\Phi}^0_{x_\cap}$ and $\mathbf{\Phi}^0_{x_\cup}$ are not $\emptyset$, $\mathbf{\Phi}^0_{x_\cup}$ may contain the common knowledge of the two texts although some exceptions also exist, for example, as shown in Table I, $\mathbf{\Phi}^0_{x_\cup}$ cannot be used to represent the common knowledge of $text_a$ and $text_c$ as "*apple*" has different meanings in them. However, such situations occur infrequently.
2) If $\mathbf{\Phi}^0_{x_\cap}, \mathbf{\Phi}^1_{x_\cap}, \mathbf{\Phi}^0_{x_\cup}$, and $\mathbf{\Phi}^1_{x_\cup}$ are not $\emptyset$, $\mathbf{\Phi}^0_{x_\cup} \wedge \mathbf{\Phi}^1_{x_\cup}$ means that $U(T_a, T_b)$ contains not only the common knowledge of the two, but also the different aspects that belong to the same theme.
3) If $I(T_a, T_b)$ is $\emptyset, U(T_a, T_b)$ is proper to be nonsense, as T_a and T_b are discussing two different topics.

Definition 5.3: The *subtraction* operation of T_a to T_b, which stands for the information that appears in T_a but not in T_b, is denoted as $S_{a-b}(T_a, T_b)$

$$\begin{aligned} S_{a-b}(T_a,T_b) &= (\mathbf{\Phi}^0_{x_a} \wedge \mathbf{\Phi}^1_{x_a} \wedge \mathbf{\Phi}^2_{x_a}) - (\mathbf{\Phi}^0_{x_b} \wedge \mathbf{\Phi}^1_{x_b} \wedge \mathbf{\Phi}^2_{x_b}) \\ &= (\mathbf{\Phi}^0_{x_a} - \mathbf{\Phi}^0_{x_b}) \wedge (\mathbf{\Phi}^1_{x_a} - \mathbf{\Phi}^1_{x_b}) \wedge (\mathbf{\Phi}^2_{x_a} - \mathbf{\Phi}^2_{x_b}) \\ &= \mathbf{\Phi}^0_{x_a - x_b} \wedge \mathbf{\Phi}^1_{x_a - x_b} \wedge \mathbf{\Phi}^2_{x_a - x_b} \\ &= \mathbf{\Phi}^0_{x_{a-b}} \wedge \mathbf{\Phi}^1_{x_{a-b}} \wedge \mathbf{\Phi}^2_{x_{a-b}} \end{aligned}$$

where x_a is the set of terms of text T_a, x_b is the set of terms of text T_b, and $x_{a-b} = x_a - x_b$ is the *Subtraction* operation of x_a to x_b.

Similarly, $S_{b-a}(T_a, T_b)$ is the subtraction of T_b to T_a, which means that the information appears in T_b but not in T_a. *Subtraction* operation between two texts means that there are some differences between them.

The previous three types of operations are the basic components of the reasoning rules of PSR. With $U(T_a, T_b)$, we are able to find all the content of the two texts; with $I(T_a, T_b)$, we could find the shared information between the two texts; with $S_{a-b}(T_a, T_b)$ and $S_{b-a}(T_a, T_b)$, we can find the different parts from each other. Therefore, the three types of basic operations play an important role in the reasoning of power series text representation.

Based on previous three basic operations, we can calculate the similarity between two texts. In addition, there are many other reasoning rules in our model, such as "*exclusive or*" and "*inclusive or.*" However, these two reasoning rules may be deduced by the previous three types of basic operation rules. Therefore, *intersection*, *union,* and *subtraction* constitute the principal part of the reasoning rules of the PSR model.

B. General Extended Reasoning

In Section V-A, we have introduced three types of basic operations, which are mainly applied into two texts. Then, in this section, we will extend the reasoning operation from a more macroscopic and general perspective. The general extended reasoning can combine all different information from different texts, and finally forms a common knowledge of those texts or other kinds of information that are latent and can be further mined.

Based on the degree-2 hypothesis of the PSR model, text universal set, which is different from traditional universal set for its several particular elements, is defined as follows.

Definition 5.4 (Text universal set): In the process of reasoning, text knowledge is represented by *degree-2 hypothesis of* the PSR model. Herein, text universal set is defined as a set that is constructed not only by the union operation of texts but also by the association rules in the prior knowledge, which is also represented by *degree-2 hypothesis of* the PSR model. The basic

idea to build the prior knowledge is to regard the prior text set as a very long text and then to obtain its text assertions and association rules with the algorithms mentioned in Sections IV-B and IV-C. However, different from single text, when mining the association rules in the prior knowledge, each text is treated as a transaction.

As a result, for n texts $\{T_1, T_2, \ldots, T_n\}$, the universal set is $U_{\{T_1,T_2,\ldots,T_n\}}$, i.e., $U(T_1, T_2, \ldots, T_n, T^{fk_{\{T_1,T_2,\ldots,T_n\}}})$, where $T^{fk_{\{T_1,T_2,\ldots,T_n\}}}$ denotes the association rules of the prior knowledge that is contained by texts $\{T_1, T_2, \ldots, T_n\}$. That is to say

$$\begin{aligned} U_{\{T_1,T_2,\ldots,T_n\}} &= U(T_1, T_2, \ldots, T_n, T^{fk_{\{T_1,T_2,\ldots,T_n\}}}) \\ &= U(T_1, U(T_2, \ldots, T_n, T^{fk_{\{T_1,T_2,\ldots,T_n\}}})) \\ &= U(T_1, U(T_2, U(T_3, \ldots, U \\ &\quad \times (T_n, T^{fk_{\{T_1,T_2,\ldots,T_n\}}})). \end{aligned}$$

Specially, for two texts T_a and T_b

$$U_{\{T_a,T_b\}} = U(T_a, T_b, T^{fk_{\{T_a,T_b\}}}) = U(T_a, U(T_b, T^{fk_{\{T_a,T_b\}}})).$$

Text universal set is the superset of union between two texts. It does not only contain the information from two texts, but also the latent background knowledge. That is to say, it can reserve the relationship between two texts.

After defining text universal set separately, the text complementary set is similar to the mathematical complementary set.

Definition 5.5 (Text complementary set): For text T_1 about the reasoning of $\{T_1, T_2, \ldots, T_n\}$, the complementary set of T_1 is

$$\begin{aligned} C_{\{T_1,T_2,\ldots,T_n\}}(T_1) &= U_{\{T_1,T_2,\ldots,T_n\}} - T_1 \\ &= S_{\{1,2,\ldots,n\}-1}(U_{\{T_1,T_2,\ldots,T_n\}}, T_1) \\ &= S_{\{1,2,\ldots,n\}-1} \\ &\quad \times (U(T_1, T_2, \ldots, T_n, T^{fk_{\{T_1,T_2,\ldots,T_n\}}}), T_1). \end{aligned}$$

Especially, for two texts T_a and T_b

$$\begin{aligned} C_{\{T_a,T_b\}}(T_a) &= U_{\{T_a,T_b\}} - T_a \\ &= S_{\{a,b\}-a}(U_{\{T_a,T_b\}}, T_a) \\ &= S_{\{a,b\}-a}(U(T_a, T_b, T^{fk_{\{T_a,T_b\}}}), T_a). \end{aligned}$$

Based on the definition of text universal set, text complementary set means the irrelative information of the certain texts.

Given a text complementary set, we can utilize the three types of basic reasoning rules to reason the inclusive set and exclusive set of texts.

Definition 5.6: The inclusive set of T_a and T_b

$$\begin{aligned} In(T_a, T_b) &= I(T_a, T_b) + I(T'_a, T'_b) \\ &= I(T_a, T_b) + I(C_{\{T_a,T_b\}}(T_a), C_{\{T_a,T_b\}}(T_b)) \\ &= I(T_a, T_b) + I(S_{\{a,b\}-a}(U(T_a, T_b, T^{fk_{\{T_a,T_b\}}}), T_a) \\ &\quad S_{\{a,b\}-b}(U(T_a, T_b, T^{fk_{\{T_a,T_b\}}}), T_b)) \\ &= \mathbf{\Phi}^0_{x_{In(a-b)}} \wedge \mathbf{\Phi}^1_{x_{In(a-b)}} \wedge \mathbf{\Phi}^2_{x_{In(a-b)}} \end{aligned}$$

where T'_a and T'_b are, respectively, the text complementary sets of T_a and T_b.

Definition 5.7: The exclusive set of T_a and T_b

$$\begin{aligned} Ex(T_a, T_b) &= I(T'_a, T_b) + I(T_a, T'_b) \\ &= I(C_{\{T_a,T_b\}}(T_a), T_b) + I(T_a, C_{\{T_a,T_b\}}(T_b)) \\ &= I(S_{\{a,b\}-a}(U(T_a, T_b, T^{fk_{\{T_a,T_b\}}}), T_a), T_b) \\ &\quad + I(T_a, S_{\{a,b\}-b}(U(T_a, T_b, T^{fk_{\{T_a,T_b\}}}), T_b)) \\ &= \mathbf{\Phi}^0_{x_{Ex(a-b)}} \wedge \mathbf{\Phi}^1_{x_{Ex(a-b)}} \wedge \mathbf{\Phi}^2_{x_{Ex(a-b)}} \end{aligned}$$

where T'_a and T'_b are, respectively, the text complementary sets of T_a and T_b.

Inclusive set and exclusive set are the extended reasoning rules. The knowledge that is obtained from inclusive set contains not only the common knowledge inside of T_a and T_b but also the common knowledge outside of them. Contrarily, the exclusive set is able to be seen as the text complementary set of the inclusive set. It contains all the knowledge that is different from each other in T_a and T_b.

C. Advanced Extended Reasoning

Last two sections have focused on the basic and general extended reasoning operations among texts, which are used to reflect the common or different knowledge between texts, and background knowledge. In this section, through advanced extended reasoning operations, we would like to discuss the knowledge inside of a text, such as whether two related knowledge units are reachable, what kind of topics are discussed in the text, and so on. Therefore, advanced extended reasoning is considered from a more microperspective.

Before introducing to the advanced extended reasoning, we will first convert[3] the simplified form of PSR, $T|_{\sum} = \mathbf{\Phi}^0_x \wedge \mathbf{\Phi}^1_x \wedge \mathbf{\Phi}^2_x$, into a $m \times m$ fuzzy matrix, shown as the following:

$$FuzzyM\left(T|_{\sum}\right) = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{pmatrix} \quad (5)$$

where r_{ij} identifies whether an association relation existing between term K_i and K_j through a weighted value during 0 and 1. If exists, r_{ij} will be set as the confidence of this association rule; otherwise, it will be set as 0. And for each diagonal element r_{ii}, the value is set as 1.

Additionally, the row i or column j refers to a term that appears in $\mathbf{\Phi}^0_x \wedge \mathbf{\Phi}^1_x \wedge \mathbf{\Phi}^2_x$ or a set of two terms that appear in the antecedent of $\mathbf{\Phi}^2_x$. Herein, we consider them as the ith and jth knowledge unit, which means the basic knowledge for us to understand. For example, given $T_a = \mathbf{\Phi}^0_x \wedge \mathbf{\Phi}^1_x \wedge \mathbf{\Phi}^2_x = \{a, b\} \wedge \{a \xrightarrow{r_1} b, a \xrightarrow{r_2} c\} \wedge \{a, b \xrightarrow{r_3} c\}$, the rows/columns

[3]It is a one-way conversion. When $T|_{\sum} = \mathbf{\Phi}^0_x \wedge \mathbf{\Phi}^1_x \wedge \mathbf{\Phi}^2_x$ is converted into a fuzzy matrix, it will lose a bit of knowledge, therefore the inverse conversion cannot be carried out without additional knowledge.

of $FuzzyM(T_a)$ are $\{a, b, c, ab\}$ and each of them is considered as a knowledge unit. For two fuzzy matrices $FuzzyM(T_a) = (a_{ij})_{m\times m}$ and $FuzzyM(T_b) = (b_{ij})_{m\times m}$, where a_{ij} and b_{ij} denote the relations in T_a and T_b, respectively, the composition of them is defined as

$$FuzzyM(T_a) \circ FuzzyM(T_b) = (c_{ij})_{m\times m},$$
$$\text{iff } c_{ij} = \bigvee_{k=1}^{m} (a_{ik} \wedge b_{kj})$$

where the operations of $\vee$ and $\wedge$ are used to get the maximum and minimum value between a_{ik} and b_{kj}.

Based on the conversion from $T|_{\sum} = \mathbf{\Phi}_x^0 \wedge \mathbf{\Phi}_x^1 \wedge \mathbf{\Phi}_x^2$ into fuzzy matrix, advanced extended reasoning definitions are given as follows.

Definition 5.8: The *power operation* of a text T_a is defined as

$$\begin{aligned}&(FuzzyM(T_a))^k = (FuzzyM(T_a))^{k-1} \circ FuzzyM(T_a)\\ &= ((FuzzyM(T_a))^{k-2} \circ FuzzyM(T_a)) \circ FuzzyM(T_a)\\ &= \underbrace{FuzzyM(T_a) \circ FuzzyM(T_a) \circ \cdots \circ FuzzyM(T_a)}_{k}.\end{aligned}$$

As aforementioned, each element a_{ij} in $FuzzyM(T_a)$ means whether the ith knowledge unit in T_a can directly reach the jth knowledge unit. Moreover, it could be deduced that each element a'_{ij} in $FuzzyM(T_a) \circ FuzzyM(T_a)$ means whether the ith knowledge unit is able to reach the jth knowledge unit in two steps. Consequently, the power operation $(FuzzyM(T_a))^k$ *reflects whether two knowledge units in T_a are reachable in k steps*.

In addition, according to the theories of fuzzy mathematics [40], since each diagonal element of $FuzzyM(T_a)$ is 1, $FuzzyM(T_a)$ is a reflexive fuzzy matrix and it has such a feature as the following shows:

$$\begin{aligned}IM \subseteq FuzzyM(T_a) \subseteq (FuzzyM(T_a))^2\\ \subseteq \cdots \subseteq (FuzzyM(T_a))^k \subseteq \cdots\end{aligned}$$

where *IM* denotes the identity matrix, and the symbol $\subseteq$ means that if $A \subseteq B$, each element b_{ij} in B is no less than a_{ij} in A.

Hence, when $(FuzzyM(T_a))^k = (FuzzyM(T_a))^{k+1}$, we are able to obtain that in T_a the maximum length of reachable path between two knowledge units is k (different with the k in degree-k), which is denoted as *MaxL* in this paper. In this situation, we will know the pairs of knowledge units that are reachable and the pairs that are not reachable in text T_a. Therefore, with the help of power *MaxL* operation $(FuzzyM(T_a))^{\mathrm{Max}L}$, the coherence of the text T_a can be well reflected.

Definition 5.9: The clustering operation of a text T_a is defined as

$$Clstr_\lambda(FuzzyM(T_a)) = \left(FuzzyM(T_a) \cup FuzzyM^T(T_a)\right)_\lambda^{\mathrm{Max}L}$$

where $FuzzyM^T(T_a)$ is the transpose of $FuzzyM(T_a)$; the superscript T denotes the transpose matrix; the symbol $\cup$ means that for two matrices $A = (a_{ij})_{m\times m}$ and $B = (b_{ij})_{m\times m}$, $A \cup B = (a_{ij} \vee b_{ij})_{m\times m}$; *MaxL* is the maximum length of reachable path in text T_a as mentioned in Definition 5.8; λ is a variable on the interval [0,1]; and the matrix $A_\lambda = (a_{ij})_{m\times m}$ means that only if $a_{ij} \geq \lambda$, $a_{ij} = 1$, otherwise, $a_{ij} = 0$.

Based on the theories of fuzzy mathematics [40], it is easy to proof that $\left(FuzzyM(T_a) \cup FuzzyM^T(T_a)\right)^{\mathrm{Max}L}$ is a reflexive, symmetric, and transitive fuzzy matrix, and hence, it is a fuzzy equivalent matrix. Therefore, according to the different selections of λ, we are able to dynamically group the knowledge units into several topics. For an example, given a text T_a that is converted into the following fuzzy matrix:

$$FuzzyM(T_a) = \begin{pmatrix} 1 & 0.4 & 0.6 & 0.5 & 0.5\\ 0.2 & 1 & 0.4 & 0 & 0.4\\ 0.8 & 0 & 1 & 0.5 & 0.5\\ 0 & 0.4 & 0 & 1 & 0\\ 0 & 0 & 0 & 0.6 & 1 \end{pmatrix}$$

then $FuzzyM(T_a) \cup FuzzyM^T(T_a)$ can be obtained

$$FuzzyM(T_a) \cup FuzzyM^T(T_a) = \begin{pmatrix} 1 & 0.4 & 0.8 & 0.5 & 0.5\\ 0.4 & 1 & 0.4 & 0.4 & 0.4\\ 0.8 & 0.4 & 1 & 0.5 & 0.5\\ 0.5 & 0.4 & 0.5 & 1 & 0.6\\ 0.5 & 0.4 & 0.5 & 0.6 & 1 \end{pmatrix}.$$

Based on the previous result, we are able to find that

$$FuzzyM(T_a) \cup FuzzyM^T(T_a) = (FuzzyM(T_a) \cup FuzzyM^T(T_a))^2$$

therefore *MaxL* of T_a can be fixed as 1. Finally, we select 1.0, 0.8, 0.6, and 0.5 as the value of λ to specify the dynamic clustering results. To avoid unnecessary details, we use $Clstr_\lambda$ to abbreviate $\left(FuzzyM(T_a) \cup FuzzyM^T(T_a)\right)_\lambda$ in the following discussions.

According to the different selections of λ, we know that when λ is 1.0, text T_a is divided into five parts: $\{ku_1\}$, $\{ku_2\}$, $\{ku_3\}$, $\{ku_4\}$, and $\{ku_5\}$ (herein, the ith knowledge unit is denoted as ku_i and if ku_i is equal to ku_j, they will be attributed to one part); when λ is 0.8, text T_a is divided into four parts: $\{ku_1, ku_3\}$, $\{ku_2\}$, $\{ku_4\}$, and $\{ku_5\}$; when λ is 0.6, text T_a is divided into three parts: $\{ku_1, ku_3\}$, $\{ku_2\}$, and $\{ku_4, ku_5\}$; when λ is 0.5, text T_a is divided into two parts: $\{ku_1, ku_3, ku_4, ku_5\}$ and $\{ku_2\}$. Each part of knowledge units is probably considered as a topic in the text.

As a result, through an appropriate λ, the clustering operation $Clstr_\lambda(FuzzyM(T_a))$ is capable of discovering the topics discussed in the text. This operation makes it possible for us to

further acquire deep knowledge in text

$$Clstr_{1.0} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad Clstr_{0.8} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$Clstr_{0.6} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \quad Clstr_{0.5} = \begin{pmatrix} 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

In conclusion, in this section, we have studied three types of reasoning rules that include basic reasoning, general extended reasoning, and advanced extended reasoning. The former two reasonings mainly consider the relations among texts while the latter primarily considers the inner knowledge in a text. It is worth noting that there are many differences between our proposed model and other models with a reasoning ability as follows.

1) The reasoning process of probability-based models, LDA for an instance, often focus on computing the posterior distribution of the hidden variables [12], with the help of approximate reasoning algorithms like Laplace approximation or Markov chain Monte Carlo [41]. This kind of reasoning always involves with several assumptions, prior knowledge, and plenty of computations, including parameter estimation, etc. Therefore, probability-based reasoning that obtains a high computing complexity is not suitable to be applied to a large-scale web environment.
2) For ontology-based reasoning, such as OWL [17], it is apparent that its reasoning is capable of offering us abundant information that possess rich knowledge by dint of much human work. However, the anticipation of human kinds also brings ontology-based reasoning many restrictions, weak immediacy, and scalability for example. As a result, ontology-based reasoning is not fit for reasoning on a large-scale dataset as well as probability-based reasoning.
3) Comparably, reasoning on the *degree*-2 *hypothesis of* the PSR model is mainly based on set operations, with which it obtains a lower complicated computation, and owing to the construction of the PSR model, this reasoning have also have a good immediacy and scalability. Consequently, *although reasoning on* the *PSR model contains less knowledge than ontology-based reasoning, it obtains good immediacies and scalabilities, and can be carried out automatically with a low complicated computation*.

VI. PSR Model Extension Based on Degree-2 Hypothesis

According to the analysis of the distribution of all the association rules' degrees that are discussed in Section IV-D, we proposed a degree-2 hypothesis, which can simplify our PSR model as $T|_{\sum} = \mathbf{\Phi}_x^0 \wedge \mathbf{\Phi}_x^1 \wedge \mathbf{\Phi}_x^2$.

This hypothesis will not only make the PSR model of text knowledge that is obtained under a much lower complexity, but will also give us an opportunity to improve the PSR model and to put more knowledge on this model.

As aforementioned, each text can be represented by two parts: text assertions and text association rules. Text assertion is a kind of commonsense terms, while text association rule describes the association relationship among terms. Both of them will make us better understand the text. However, association rules still seem to be not sufficient to express the deeper knowledge, for they only tell us the fact that there exists a relationship among the terms, of which an association rule consists, (considering to change this sentence) nevertheless, we cannot understand what kind of relations they are. For example, through the association rule $governments \rightarrow plan$, what we shall learn is that the "*governments*" has some kind of relation with the "*plan*" while the specific information, such as its attitude to this "*plan,*" is unknown.

As a result, with the help of *degree*-2 *hypothesis*, we shall use some appointed terms, verbs for example, to specify the relations among the terms. In other terms, the relations will be further explained by several specific terms. Still taking $governments \rightarrow plan$ as an example, a verb "*agree*" will be used to explain the relation between the "*governments*" and the "*plan*." Then, the extension of text association rule $governments \rightarrow plan$ will be represented as $governments \xrightarrow{\{agree\}} plan$. From such a further specification, we shall know that the governments tend to agree with this plan, which can not be learnt from the association rule $governments \rightarrow plan$. In the consideration of that, the formal representation of a text association rule with semantic explanation is shown as follows:

$$a, b, c, \cdots \xrightarrow{\{verb_1, verb_2 \ldots, verb_s\}} x$$

where $a, b, c, \ldots, x$ belong to a set consisting of nouns and noun phrases that are extracted from the text. The verbs upon the arrow are obtained based on the general association rule $a, b, c, \cdots \rightarrow x$.

Therefore, the general extension of the PSR model constructed by text assertions and text association rules with semantic explanation is formalized as

$0 : \mathbf{\Phi}_x'^0 = \mathbf{\Phi}_x^0$(text assertions)
$1 : \mathbf{\Phi}_x'^1 = \left\{ a \xrightarrow{v_1, v_2 \ldots} b, \cdots \right\}$
$\ldots$
$k : \mathbf{\Phi}_x'^k = \left\{ \overbrace{c, d, \ldots, f}^{k} \xrightarrow{v_i, v_{i+1} \ldots} y, \cdots \right\}$
$\ldots$
$D-1 : \mathbf{\Phi}_x'^{D-1}$(text association rules of the largest degree)

where $\mathbf{\Phi}_x'^k$ denotes the set of degree-k association rules with semantic explanation.

However, the higher the degree, the harder the explanation obtained. At this moment, degree-2 hypothesis plays an important role to solve this problem for it can not only simplify the

TABLE II
BRIEF COMPARISON OF A REASONING ABILITY, AUTOMATIC CONSTRUCTION ABILITY, COMPLICATED COMPUTATION, AND SEMANTIC RICHNESS AMONG THE VSM, EFCM, PSR, LSI, LDA, ATM, CTM, AND OWL

Representation models		VSM	EFCM	PSR	LSI	LDA, ATM, CTM	OWL
Reasoning	Union	√	√	√	×	×	√
	Intersection	√	√	√	×	×	√
	Subtraction	√	√	√	×	×	√
	Inclusive	√	√	√	×	×	√
	Exclusive	√	√	√	×	×	√
Automatic construction		√	√	√	√	Semi-Automatic	Semi-Automatic
Computing complexity		Very low	Low	Low	High	Very High	Very High
Semantic richness		Poor	Medium	Rich	Rich	Very Rich	Very Rich

PSR model but also simplify its general extension. Therefore, based on degree-2 hypothesis, each text will be represented as

$$T|_{\sum} = \mathbf{\Phi}_x'^0 \wedge \mathbf{\Phi}_x'^1 \wedge \mathbf{\Phi}_x'^2 \tag{6}$$

which can be acquired under a lower complexity but with more knowledge.

The extension of the PSR model based on degree-2 hypothesis can also be constructed automatically, but moreover, it contains more knowledge than the original PSR model and has excellent compatibility with other text representation models, such as the VSM, EFCM, and resource description framework (RDF) [42]. Herein, we take RDF as an example.

As we know, a RDF statement consists of three individual parts that are called the subject, the predicate, and the object. In our PSR extension model, without consideration of text assertions, a resource can also be represented by three parts: the antecedent, the descendent, and the verb set used to specify the relation between the antecedent and the descendent. As a result, it is obvious that there exists a one-to-one correspondence between the RDF model and ours. Therefore, the PSR extension model can be converted into the RDF model under a given condition, which is to not take Degree-2 association rules into account, whereas the conversion cannot be reversed since the RDF model contains nothing about text assertions.

Through the analysis of the previous examples, the fact can be learnt that it is convenient to convert a text which is represented by the PSR extension model to the one represented by others because of its good compatibility. Moreover, drawing on the experience of this advantage, it is also possible to apply the PSR extension model to the applications based on other text representations that PSR can be converted to.

VII. COMPARING PSR WITH OTHER TEXT REPRESENTATION MODELS

In Section V, we have studied the reasoning operations of the PSR model. In order to verify that the PSR model has better performances than other models in representing text knowledge, we compare our model with them in this section.

A. Generally Comparing PSR With Other Models

In this part, we generally compare the PSR model with current knowledge representation models in reasoning ability, complicated computation, automatic construction ability, and semantic richness. The general comparison results are shown in Table II. This table discusses statistics models, cognition-based models, probability topic models, ontology representation models, and our PSR model. From this table, we can find that PSR, LSI, LDA ATM, CTM, and OWL are all of rich knowledge, but only PSR can be constructed automatically with a lower complicated computation. In the next section, we will focus on the comparison among the VSM, EFCM, PSR, LDA, and OWL through theoretical analyses.

B. Theoretical Analyses and Comparisons

Currently, as the number of web pages is growing much faster than before, the ability of automatic construction has become an increasingly important feature of text representation models. Therefore, in this paper, automatic construction ability is taken into account. Simultaneously, herein, we compare our proposed PSR model with these four kind of models, which are statistics model VSM, cognition-based model EFCM, probability topic model LDA, and ontology-based model OWL, respectively. And the comparative features include three abilities: knowledge representation ability, reasoning ability, computing complexity and automatic construction ability.

The comparison results are shown in Tables III–V in detail. As shown in those tables, it can be concluded that among all of the five representation models, LDA and OWL models obtain the richest knowledge in the abilities of text representation and reasoning; however, as LDA needs prior knowledge and OWL involves with huge human efforts, they obtain a high complexity and are not able to be constructed automatically. On the contrary, although the VSM has the lowest complexity and is able to be automatically constructed, it contains the least knowledge and weakest reasoning ability. EFCM possesses more knowledge than the VSM and is also able to be automatically constructed with a low complexity. However, EFCM is still weaker in the ability of knowledge representation than OWL and PSR.

According to Tables III–V, comparing with the other models, PSR obtains more knowledge than the VSM and EFCM but lower knowledge and complexity than LDA and OWL. And it also has a good reasoning ability that consists of interaction, union, subtraction, etc. Therefore, the PSR model can leverage the contradiction between the carrying rich knowledge and the automatic construction complexity in the text knowledge representation process.

With respect to the rapid growth of a web environment, we would like to compare the proposed model PSR with the models that can be automatically constructed in the following section. In

TABLE III
KNOWLEDGE REPRESENTATION ABILITIES OF THE VSM, EFCM, PSR, LDA, AND OWL

Model	Knowledge Representation Ability
VSM	It just can represent the weight of terms, so this model has poor knowledge representation ability.
EFCM	Co-occurrence among terms can be represented, and *degree-1* relationships can be expressed by the model. So, this model has rich knowledge than VSM.
PSR	This model contains the text assertion, *degree-1* text association rules and *degree-2* text association rules; therefore it can express rich semantic relationship among texts. So this model has high rich knowledge representation ability.
LDA	This model is a generative probabilistic model of a corpus. Texts in this model are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. Therefore, this model has moreknowledge than VSM and less knowledge than PSR.
OWL	Texts represented by OWL often contain these components that concern classes, properties, instances of classes and relationships between these instances. These components enable OWL to obtain rich knowledge from texts, but at the same time, they also request its contruction to be involved with human efforts.

TABLE IV
REASONING ABILITY OF THE VSM, EFCM, PSR, LDA, AND OWL

Model	Reasoning Ability and Computing Complexity
VSM	There is low reasoning ability and very low complicated computation since this model does not take relation into consideration.
EFCM	This model can be formed into a matrix, so it can carry out simple numerical reasoning when the pages and terms are not very huge. And the relations between terms can be druggedout from original text contents.
PSR	Based on *degree-2 hypothesis*, union, intersection, subtraction, universal, complementary, inclusive, exclusive, power and clustering operators can be processed. With these operators, the knowledge inside or outside of texts can be well reasoned.
LDA	The key reasoning problem in LDA is that of computing the posterior distribution of the hidden variables, which is intractable for exact reasoning. So a wide of approximate reasoning algorithms are considered for LDA, including Laplace approximation, variational approximation and Markov chain Monte.
OWL	With the help of human efforts, reasoning in OWL can offer us rich knowledge, such as subclass-of, or subproperty-of relations, etc. But there are many restrictive syntactic constrains resulting in its inconvenient construction.

TABLE V
AUTOMATIC CONSTRUCTION ABILITY AND COMPLICATED COMPUTATION OF THE VSM, EFCM, PSR, LDA, AND OWL

Model	Automatic Construction Ability
VSM	This model mainly considers weights of terms from texts, so it can be constructed automatically by the known term frequency-inverse document frequency (TF-IDF).
EFCM	This model mainly expresses the co-occurrence among terms, so similarly the model can be built through data mining to obtain the *degree-1* relations between terms. Therefore, this model's complicated computation is medium.
PSR	From our paper, this model can be constructed automatically by the mature algorithms of data mining, based on the idea of generating implications among concepts. Because the model only condiser the *degree-1* relations and *degree-2* relations as well as the text assertion, this model's complicated computation is not very high.
LDA	According to the assumptions of LDA, the number of topics in the corpus represented by LDA won't be changed as long as it is constructed. And its construction involves with a lot of prior knowledge offered by humans. We know that the variational approximation and Markov chain Monte have very high complicated computation, so this model has a very high complicated computation.
OWL	The components in OWL, referring to classes, properties, instances of classes and relationships between these instances, are hard to be obtained without human efforts. So it is obvious to get that the construction of OWL model is semi-automatic.

another word, models of LDA and OWL will not be discussed in the rest of this paper. Additionally, in our preliminary work [43] on PSR done before, we have also compared PSR with the VSM and EFCM through some case studies, with which the concept of PSR will be understood more deeply.

C. Verification by Experiments

With the rapid development of the web, more and more information is added to it every day; accordingly, the way to acquire knowledge is mainly depending on search engines. Therefore, at the present, the task of information retrieval is becoming increasingly important to us. In this section, we will compare PSR with EFCM and the VSM in the view of information retrieval since those models do not need prior knowledge and can be constructed automatically while the others (e.g., ATM, CTM, OIL, and OWL) cannot be automatically constructed. Therefore, in this paper, we present an experiment in the view of information retrieval to show the benefits of PSR in a dataset that included 684 documents.

It was said by Salton *et al*. [5] that in document retrieval, it appeared that the best indexing (property) space was the one where each entity (i.e., document) lied as far away from the others as possible; in particular, retrieval performance might correlate inversely with space density.

In this paper, we are going to use a function F, which is designed to reflect the total similarity of a document space, to represent the space density. And the following gives its

definition:

$$F = \sum_{i=1}^{n} \sum_{j=1 \wedge j \neq i}^{n} s(D_i, D_j) \tag{7}$$

where $s(D_i, D_j)$ denotes the similarity between the pair of documents D_i and D_j, and n is the size of document collection.

Herein, in order to make the results more general, we adopt three different methods to compute the similarities between pairs of documents, which are, respectively, cosine, Jaccard, and Dice similarities. Cosine similarity is defined as the following:

$$Cosine\_s(D_i, D_j) = \frac{\sum_{k=1}^{s} (wf_{ik} \times wf_{jk})}{\sqrt{\sum_{k=1}^{s} (wf_{ik})^2} \times \sqrt{\sum_{k=1}^{s} (wf_{jk})^2}} \tag{8}$$

where wf_{ik} is the weight of the kth feature of document D_i, and s denotes the total number of all the features.

In addition, Jaccard and Dice can be generalized by Tversky's ratio model [44], which is defined as follows:

$$RatioModel\_s(D_i, D_j) = \frac{f(D_i \cap D_j)}{f(D_i \cap D_j) + \alpha f(D_i - D_j) + \beta f(D_j - D_i)} \tag{9}$$

where $f(D_i \cap D_j)$ represents the features that documents D_i and D_j have in common; $f(D_i - D_j)$ represents the features that document D_i has but D_j does not, and $f(D_j - D_i)$ represents the features that document D_j has but D_i does not; α and β are the parameters to specify the importance of the components. For $\alpha = \beta = 1$, the similarity measure is the Jaccard coefficient and for $\alpha = \beta = 0.5$, the similarity measure is the Dice coefficient. Both of them are given as follows:

$$Jaccard\_s(D_i, D_j) = \frac{f(D_i \cap D_j)}{f(D_i \cup D_j)} \tag{10}$$

where $f(D_i \cup D_j)$ represents all the features that document D_i or D_j has

$$Dice\_s(D_i, D_j) = \frac{f(D_i \cap D_j)}{f(D_i) + f(D_j)} \tag{11}$$

where $f(D_i)$ represents the features that D_i has.

Moreover, in this paper, the function f in (9) to (11) is defined as

$$f(D) = \sum_{k=1}^{s} wf_k \tag{12}$$

where wf_k is the weight of the kth feature of document D, and s denotes the total number of all the features.

According to the equations from (9) to (12), the total similarity of a document space based on function F in (7) are able to be obtained. Obviously, when the value returned by the function F is small, the differences between pairs of documents are big; thus, this document space density is small. As aforementioned, retrieval performance correlates inversely with space density [5], therefore the smaller the F, the better the retrieval performance.

In this section, we crawled 8 classes, totally 684 documents, downloaded from www.reuters.com, which are alum (58 documents, abbreviate as docs), bop (101 docs), carcass (74 docs), cocoa (68 docs), coffee (143 docs), copper (77 docs), cotton (62 docs), and cpi (101 docs). The documents in each class are, respectively, represented by PSR, EFCM, and the VSM, and then we compute the total similarity (function F) for each class according to cosine, Jaccard, and Dice similarities. The results are shown in Figs. 6–8, from which, we are easy to obtain a consistent result that for all of the three different similarity methods, the VSM has the biggest similarity and PSR gets the smallest similarity. In other words, PSR possesses a better retrieval performance than the VSM and EFCM since retrieval performance correlates inversely with space density. As a result, in the view of information retrieval [5], the experimental result shows that the proposed PSR model is better than EFCM and the VSM.

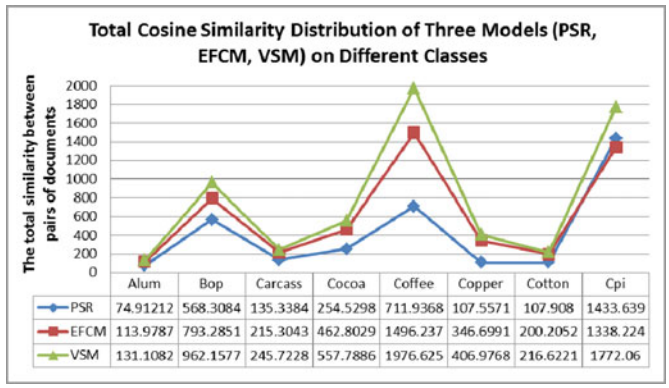


Fig. 6. Total cosine similarity distribution of three models (PSR, EFCM, VSM) on different classes.

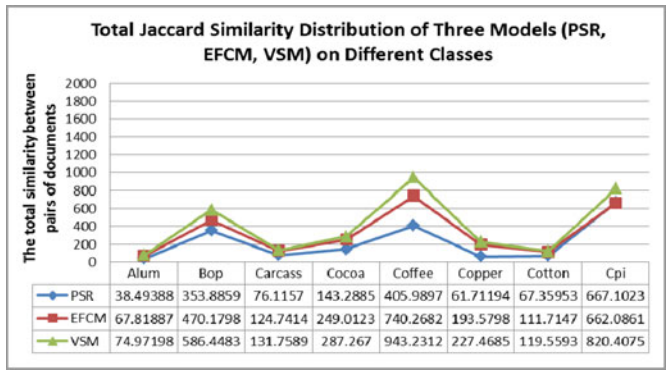


Fig. 7. Total Jaccard similarity distribution of three models (PSR, EFCM, VSM) on different classes.

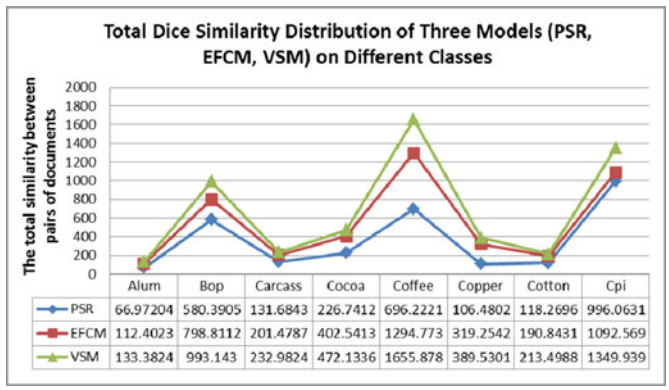


Fig. 8. Total dice similarity distribution of three models (PSR, EFCM, VSM) on different classes.

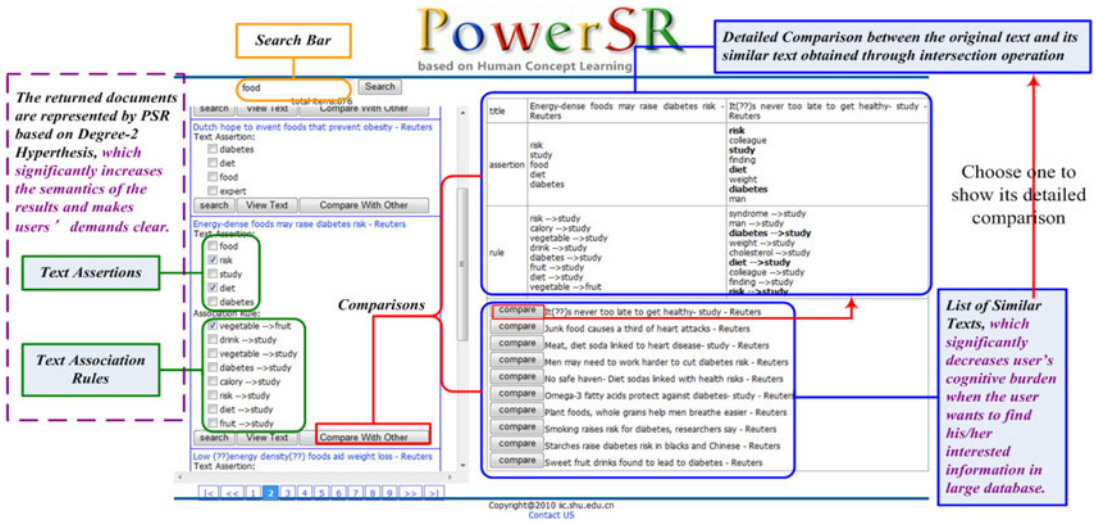


Fig. 9. Interface of a text search engine based on PSR degree-2 hypothesis.

From Sections VII-A to VII-C, we not only give the summary comparisons and theoretical analyses between the PSR model and other models, but also present verifications by experimental results, which show the benefits of our proposed model. Moreover, we will carry out a text search engine based on the PSR model to further express its virtue in semantic search in the next section.

VIII. Text Search Engine Based on PSR Degree-2 Hypothesis

In Section V, we have systematically studied the reasoning characters of the PSR model. Based on the model, three types of basic reasoning rules are developed, which are the operations of *intersection*, *union*, and *subtraction* between two texts represented by *degree*-2 *hypothesis* of PSR.

In this section, a text search engine based on the *degree*-2 *hypothesis* of PSR is given for searching suitable texts for users. An experimental search engine is built to explain the application of the PSR model. In this search engine, 14 627 pieces of news are downloaded from www.reuters.com as the dataset, which belongs to *health news*. The search engine (http://iic.shu.edu.cn/psr/) can push suitable texts to users under certain requirement, when some operations have been done in the engine by users.

Taking "*food*" as an example, when user enters a word "*food*" in the search engine, the results will be listed in the left part of the interface that is shown in Fig. 9. Each returned result is composed of text assertions, degree-1 and degree-2 text association rules. The returned result can provide rich knowledge for users, and the users can gain new query term from the returned result. Therefore, our search engine can provide users a new way to modify user's requirements and obtain new query terms through the selection of text assertions and text association rules, which make user clear his/her requirements effectively.

Additionally, according to the PSR reasoning operations, the engine will offer users a comparison with other texts as shown in the right side of Fig. 9. The search engine will return the top ten texts that have the biggest intersection values with the corresponding text. Through the comparison, users can learn much more about their requests and significantly decrease user's cognitive burden when the user wants to find his/her need information in large database. For instance, if the user finds that the text of "*Energy-dense foods may raise diabetes risk*" is very close to his/her interest and want to get more similar texts to find out the most related one, he/she can click the button of "*Compare With Other*" to get the right-side results shown in Fig. 9, which are composed of two parts: 1) one is a detailed comparison between the original text and its most similar text obtained through *intersection operation*; 2) the other is a list of the top ten texts similar to the original one. In this case, we obtain that the text of "*It's never too late to get health*" is the most similar one to the text of "*Energy-dense foods may raise diabetes risk*" through the basic reasoning of *intersection*. Besides, the user perhaps is able to find his/her interests in the list of similar texts, then the user can click the button of "*compare with other*" to get the detailed information of those texts he/she interested in, which is shown in the right of the interface.

From the previous discussion, we know that our PSR-based search engine has three contributions compared with the traditional search engine: 1) it can provide a rule-based search experience while traditional search engines are only term based, which can give a new response that is much closer to user's demands; 2) users are enabled to choose their interested information that is represented by text assertions and text association rules, which can make user clear his/her demands and significantly increase the knowledge of the results in the search process; 3) the engine can return the top ten texts that have the biggest intersection values with the corresponding text, which can effectively decrease the cognitive burden when user wants to find suitable text.

The previous three contributions illustrate that a PSR-based search engine has a good prospect on the semantic search area. As a result, with the help of the PSR-based search engine, users can not only get the specific information of a text, but can also find out his/her interest in a much less interaction steps comparing with traditional search engines. In addition, this demo can

be found at http://iic.shu.edu.cn/psr/. Moreover, some search engine-based text analyzing approaches, such as measuring the semantic similarity between words [45], can also be largely improved through our proposed PSR-based search engine.

IX. Conclusion

In order to leverage the contradiction between the carrying rich knowledge and the automatic construction with low complicated computation in the text knowledge representation process, a novel text knowledge representation model, PSR, has been proposed based on the human concept learning, which makes some contributions to the development of the text knowledge representation as follows.

1) The PSR model of text knowledge has been proposed based on the human concept learning. Comparing with other models, PSR contains more knowledge than LSI and (EFCMs, and is in consistence with the human cognitive process. Additionally, the establishment of the PSR model is automatic while the construction process of OWL is semiautomatic.
2) *Degree*-2 *hypothesis* has been proposed in order to simplify the PSR-based text knowledge representation. Based on the *sliding window*, the mining algorithm of high-degree text association rules obtains a high complexity that makes PSR hard to be applied into a large database. During such a situation, *degree*-2 *hypothesis* is proposed to point that as degree-1 and degree-2 text association rules contain more than 90% knowledge of the text, the PSR model can be simplified as $\mathbf{\Phi}_x^0 \wedge \mathbf{\Phi}_x^1 \wedge \mathbf{\Phi}_x^2$. As a result, the simplified model largely reduces the constructional complexity of PSR.
3) Based on the *degree*-2 hypothesis, we have studied the reasoning rules of the PSR model, which performs a flexible reasoning ability than LDA and OWL. Through the implementation process of the reasoning rules, we can find the common information, differences between texts, the domain knowledge of all the texts, and the inner knowledge in the text, all of which supply the theoretical foundation for discovering deep knowledge in texts.
4) The PSR-based text semantic search gives a novel search way, rule-based search, to further specify a user's search goal that enables the responses to be much closer to the ones the user expects. Besides, through intersection operation, the search engine can also provide a comparison between the similar texts, which can significantly decrease the cognitive burden when user wants to find a suitable text and effectively make users clear their requirement.

In our future work, we will apply the PSR model of text knowledge to e-learning [8], information retrieval, and semantic link network-based semantic search [3], etc.

Acknowledgment

The authors would like to thank the editors and all the reviewers for their detailed comments and suggestions that help them to significantly improve the quality of this paper.

References

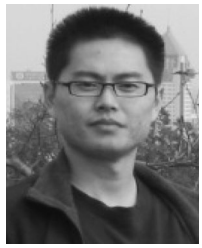
[1] Q. Li, J. Wang, Y. P. Chen, and Z. Lin, "User comments for news recommendation in forum-based social media," *Inf. Sci*., vol. 180, no. 24, pp. 4929–4939, 2010.
[2] A. Liu, Q. Li, L. Huang, and M. Xiao, "FACTS: A framework for fault-tolerant composition of transactional web services," *IEEE Trans. Serv. Comput*., vol. 3, no. 1, pp. 46–59, Jan./Mar. 2010.
[3] X. Luo and Z. Xu, "Building association link network for semantic link on web resources," *IEEE Trans. Autom. Sci. Eng*., vol. 8, no. 3, pp. 482–494, Jul. 2011.
[4] B. Heitmann, R. Cyganiak, C. Hayes, and S. Decker, "An empirically grounded conceptual architecture for applications on the web of data," *IEEE Trans. Syst. Man Cybern., Part C: Appl. Rev*., vol. 42, no. 1, pp. 51–60, Jan. 2012.
[5] G. Salton, A. Wang, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 8, no. 11, pp. 613–620, 1975.
[6] S. Deerwester, S. T. Dumais, G. W. Furnas, and T. K. Landauer, "Indexing by latent semantic analysis," *J. Am. Soc. Inf. Sci*., vol. 41, pp. 391–407, 1990.
[7] K. Perusich and M. D. McNeese, "Using fuzzy cognitive maps for knowledge management in a conflict environment," *IEEE Trans. Syst. Man Cybern., Part C: Appl. Rev*., vol. 36, no. 6, pp. 810–821, Nov. 2006.
[8] X. Luo, X. Wei, and J. Zhang, "Guided game-based learning using fuzzy cognitive maps," *IEEE Trans. Learn. Technol*., vol. 3, no. 4, pp. 344–357, Oct./Dec. 2010.
[9] Y. Wang, S. Patel, and D. Patel, "A layered reference model of the brain (LRMB)," *IEEE Trans. Syst. Man Cybern., Part C: Appl. Rev*., vol. 36, no. 2, pp. 124–133, Mar. 2006.
[10] J. Feldman, "Minimization of Boolean complexity in human concept learning," *Nature*, vol. 407, no. 5, pp. 630–633, 2000.
[11] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Annu. Int. SIGIR Conf. Res. Dev. Inf. Retrieval*, 1999, pp. 50–57.
[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res*., vol. 3, pp. 993–1022, 2003.
[13] M. Rosen-Zvi, T. Griffths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proc. 20th Int. Conf. Uncertainty Artif. Intell*., 2004, pp. 487–494.
[14] A. McCallum, A. Corrada-Emmanuel, and X. Wang, "The author-recipient-topic model for topic and role discovery in social networks: Experiments with Enron and Academic email," *J. Artif. Intell. Res*., vol. 29, pp. 249–272, 2007.
[15] D. M. Blei and J. D. Lafferty, "Correlated topic models," in *Advances in Neural Information Processing Systems 18*. Cambridge, MA, USA: MIT Press, 2006.
[16] R. Fikes and D. L. McGuinness, "An axiomatic semantics for RDF, RDF-Schema, and DAML-ONT," Knowledge Systems Laboratory, Stanford Univ., Stanford, CA, Tech. Rep., Dec. 2000. [Online]. Available: http://www.ksl.stanford.edu/people/dlm/daml-semantics/abstract-axiomatic-semantics.html
[17] OWL Web Ontology Language Overview (2004). [Online]. Available: http://www.w3.org/TR/owl-features/
[18] V. Milea, F. Frasincar, and U. Kaymak, "tOWL: A temporal web ontology language," *IEEE Trans. Syst. Man Cybern., Part B: Cybern*., vol. 42, no. 1, pp. 268–281, Feb. 2012.
[19] (2000). [Online]. Available: http://www.cs.umd.edu/projects/plus/SHOE/
[20] X. Tao, Y. Li, and N. Zhong, "A personalized ontology model for web information gathering," *IEEE Trans. Knowl. Data Eng*., vol. 23, no. 4, pp. 496–511, Apr. 2011.
[21] N. Fang and X. Luo, "Measuring textual context based on cognitive principles," *J. Softw. Sci. Comput. Intell*., vol. 1, no. 4, pp. 61–89, 2010.
[22] J. Feldman, "An algebra of human concept learning," *J. Math. Psychol*., vol. 50, pp. 339–368, 2006.
[23] J. Feldman, "A catalog of Boolean concepts," *J. Math. Psychol*., vol. 47, pp. 75–89, 2003.
[24] D. L. Medin and P. J. Schwanenflugel, "Linear separability in classification learning," *J. Exp. Psychol.: Human Learn. Memory*, vol. 7, no. 5, pp. 355–368, 1981.
[25] J. P. Minda, J. D. Smith, and M. J. Morgan, "Straight talk about linear separability," *J. Exp. Psychol.: Learn. Memory Cognit*., vol. 23, pp. 659–680, 1997.
[26] S. A. Sloman, B. C. Love, and W. K. Ahn, "Feature centrality and conceptual coherence," *Cognit. Sci*., vol. 22, pp. 189–228, 1998.
[27] J. Feldman, "The simplicity principle in human concept learning," *Curr. Directions Psychol. Sci*., vol. 12, no. 6, pp. 227–232, 2003.

[28] B. Hayes-Roth and F. Hayes-Roth, "Concept learning and the recognition and classification of exemplars," *J. Verbal Learn. Verbal Behavior*, vol. 16, pp. 321–338, 1977.
[29] J. P. Minda and J. D. Smith, "Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation," *J. Exp. Psychol.: Learn. Memory Cognit.*, vol. 28, pp. 275–292, 2002.
[30] Y. Yao, "Interpreting concept learning in cognitive informatics and granular computing," *IEEE Trans. Syst. Man Cybern., Part B: Cybern.*, vol. 39, no. 4, pp. 855–866, Aug. 2009.
[31] Y. Wang, "Cognitive informatics models of the brain," *IEEE Trans. Syst. Man Cybern., Part C*, vol. 36, no. 2, pp. 203–207, Mar. 2006.
[32] Y. Wang, "The theoretical framework of cognitive informatics," *Int. J. Cognit. Inf. Nat. Intell.*, vol. 1, no. 1, pp. 1–27, 2007.
[33] Y. Wang, "The OAR model of neural informatics for international knowledge representation in the brain," *Int. J. Cognitive Inf. Nat. Intell.*, vol. 1, no. 3, pp. 64–75, 2007.
[34] K. Rayner, "Eye movements in reading and information processing: 20 years of research," *Psychol. Bulletin*, vol. 124, pp. 372–422, 1998.
[35] W. Ni and J. D. Fodor, "Anomaly detection: Eye movement patterns," *J. Psycholinguistic Res.*, vol. 27, pp. 515–539, 1998.
[36] X. Luo, Z. Xu, Z. Xu, Q. Li, Q. Hu, J. Yu, and X. Tang, "Generation of similarity knowledge flow for intelligent browsing based on semantic link networks," *Concurrency Comput.: Pract. Experience*, vol. 21, no. 16, pp. 2018–2032, 2009.
[37] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. VLDB Conf.*, Santiago, Chile, Sep. 1994, pp. 487–499, (Expanded version available as IBM Research Report RJ9839, 1994).
[38] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psychol. Rev.*, vol. 63, pp. 81–97, 1956.
[39] R. Knauf, S. Tsuruta, and A. J. Gonzalez, "Toward reducing human involvement in validation of knowledge-based systems," *IEEE Trans. Syst. Man Cybern., Part A: Syst. Humans*, vol. 37, no. 1, pp. 120–131, Jan. 2007.
[40] L. Zadeh, K. S. Fu, K. Tanaka, and M. Shimura, *Fuzzy Sets and Their Applications to Cognitive and Decision Processes*. New York: Academic Press, 1975.
[41] M. I. Jordan, *Learning in Graphical Models*. Cambridge, MA: MIT Press, 1999.
[42] Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation (1999). [Online]. Available: http://www.w3.org/TR/1999/ REC-rdf-syntax-19990222/
[43] X. Luo, C. Cai, and Q. Hu, "Text knowledge representation model based on human concept learning," in *Proc. 9th Int. Conf. Cognit. Informat.*, 2010, pp. 383–390.
[44] A. Tversky, "Features of similarity," *Psychol. Rev.*, vol. 84, no. 4, pp. 327–352, 1977.
[45] D. Bollegala, Y. Matsuo, and M. Ishizuka, "A web search engine-based approach to measure semantic similarity between words," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 7, pp. 977–990, Jul. 2011.

Xiangfeng Luo received the Master's and Ph.D. degrees from the Hefei University of Technology, Hefei, China, in 2000 and 2003, respectively.

He is a Professor with the School of Computers, Shanghai University, Shanghai, China. He is currently a Visiting Professor with Purdue University, West Lafayette, IN. He was a Postdoctoral Researcher with the China Knowledge Grid Research Group, Institute of Computing Technology, Chinese Academy of Sciences, from 2003 to 2005. He has authored or coauthored more than 50 publications and his publications have appeared in the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS-PART C, and IEEE TRANSACTIONS ON LEARNING TECHNOLOGY. His main research interests include web wisdom, cognitive informatics, and text understanding.

Dr. Luo has served as the Guest Editor of *ACM Transactions on Intelligent Systems and Technology*. He has also served on the committees of a number of conferences/workshops, including Program Co-Chair of the International Conference on Web-based Learning (ICWL 2010) (Shanghai), International Conference Web Information Systems and Mining (WISM 2012) (Chengdu), International Workshop on Cognitive-based Text Understanding and Web Wisdom (CTUW 2011) (Sydney), and more than 40 PC members of conferences and workshops.



Jun Zhang received the Bachelor's degree from Shanghai University, Shanghai, China, in 2008, where he is currently working toward the Ph.D. degree in the School of Computers.

His main research interests include online word relation discovery, knowledge representation, topic detection and tracking.

Feiyue Ye received the Master's degree from Shandong University, Shandong, China, in 1995, and the Ph.D. degree from the China University of Petroleum, Shandong, in 2000.

He has published more than 50 papers indexed by EI or SCI, including one treatise and three teaching materials. His research interests include information retrieval, database and mobile computing, etc.

Dr. Ye has received more than 10 scientific awards including two awards for the advanced science and technology.

Peng Wang received the Ph.D. degree from Tsinghua University, Beijing, China, in 2006.

He is currently with the National High-Performance IC Design Center, Shanghai, China. From 2008 to 2010, he was a Postdoctoral Researcher in the Jiangnan Institute of Computer Technology. His main research interests include SoC design and verification and cognitive computing.

Chuanliang Cai received the Bachelor's degree from the Wuhan University of Science and Technology, Wuhan, China, in 2008, where he is currently working toward the Master's degree in the School of Computers.

His main research interests include text representation and complexity measure.