International Journal of Grid and Distributed Computing
Vol. 10, No. 1 (2017), pp.63-74
http://dx.doi.org/10.14257/ijgdc.2017.10.1.07

A Novel Approach to Building Knowledge Base for Chronic Disease Management using Ontology

Wang Hui[1], Qinyan Zhang[2], Yang Huang[3], Meng Xi[3]

[1.] Inner Mongolia Medical University, 010050 Hohhot, P.R. China
[2.] Computing Center, Zhejiang University, 310058 Hangzhou, P.R. China
[3.] College of Computer Science and Technology, Zhejiang University, 310058 Hangzhou, P.R. China
Email: zqy@scezju.com

Abstract

It is convenient to use mobile device to obtain information from integrated applications over Internet. Due to the prosperity and development of networked services in our daily life, most of chronic diseases are suitable to be monitored under the Internet and Healthcare mode. However, the important issue is how to build knowledge base for Healthcare system to respond to different kinds of user queries in order to understand what user wants to search. In this paper, we focus on the hypertension management and monitoring problem of chronic diseases. First, Ontology is adopted to specify hypertension that corresponding domain concepts are organized into formal expressions and knowledge graph, which are different from the general database management since they are more flexible to implement the knowledge searching and reasoning. Second, the Chinese words segmentation method is employed to precisely analyze the searching query, and the information searching method of hypertension Ontology is proposed in order to get more accurate results. Finally, the experimental analysis are demonstrated the effectiveness and feasibility of our method.

Keywords*: Mobile Internet, Knowledge Base, Chronic Disease Management, Ontology, Chinese Words Segmentation.*

1. Introduction

With the rapid development of Internet infrastructures, the popularity of mobile devices has been promoting the new prosperity of Internet Healthcare, which enables people to use various Healthcare apps in anytime and anywhere [1]. In e-business market, chronic diseases, such as hypertension and fatty liver, have been monitored by mobile devices in our daily life, especially using the technology of IoT(Internet of Things)[2,3,4]. Mobile devices provide good opportunities to chronic diseases management and monitoring, which aims to support not only 24 hours services for hypertension patients, but also information to end user for health care, as well as preventing the future disease from happening.

As we known, knowledge base is a collection of comprehensive cognitions to a particular field, which is usually in the form of structured or semi-structured data. This information includes classifications and definitions. For example, there may have hypertension definitions of chronic diseases, relations between hypertension and other diseases, and BMI index and suggestions [5]. After that, we can carry out the information retrieval, and give decision supports. For instance, if user wants to know the hypertension definition, the diagnostic criteria can be returned by searching related keywords and similar terms. However, all these functionalities require professional knowledge base to

ISSN: 2005-4262 IJGDC

achieve the accurate searching. There is a little work to building knowledge base for the chronic disease management.

Some excellent applications are available on Internet, i.e., spring rain doctors, registration platform, Mywood health. And there are 37 intelligent devices used in the area of chronic disease, such as intelligent sphygmomanometer, MUMU sphygmomanometer, and sugar nurse [6]. Most of these products focus on communication among doctors and patients via intelligent devices. However, these applications or sphygmomanometers have defects on the accuracy and precision of information. The reason is attributed to disease management knowledge base, especially Chinese information. It is hard to describe and search because English knowledge bases are different from Chinese knowledge base. Thus, building Chinese knowledge base should consist of following research points.

1) How to build knowledge base for chronic disease. In the application level, it should take into account Chinese feature to choose the appropriate knowledge base.

2) The problem of the accuracy of the segmentation technology. The keyword should be effectively segmented and extracted from user query because the result precision is based on the efficiency of word segmentation.

Furthermore, difference between Chinese and English means that we cannot directly copy the English knowledge base. Famous medicine knowledge base UMLS (Unified Medical Language System) [7] is not suitable for the disease management because it is a desktop application which cannot be accessed by Internet. To these problems, this paper proposes a novel approach to building knowledge base of chronic disease management using Ontology, which mainly includes the constriction and information retrieval of knowledge base. In the former, it uses Ontology to describe the knowledge for chronic disease management. While in the latter, it uses the Chinese words segmentation method to handle the keyword understanding. As result, this approach can effectively manage knowledge and return suitable searching results.

The rest of this paper is organized as follows. In section 2, the domain Ontology of chronic disease management is discussed. In section 3, the Ontology searching based on Chinese word segmentation is introduced. In section 4, experiments and analysis are discussed. Finally, we conclude this research and discuss future works in section 6.

2. The Domain Ontology of Chronic Disease Management

2.1. Hypertension Ontology

Hypertension is considered as one of chronic diseases. We aim to build knowledge base for the hypertension management. Before building Ontology for chronic disease management, we should clearly establish the domain of Ontology. Thus, it is necessary to collect all aspects of knowledge which are related to hypertension management from professional books, such as the doctor guidelines to the prevention and treatment of hypertension. Then, it should learn the knowledge of hypertension by multiple communications to physicians, and read related books, including WHO standard and national standard, to understand the framework and process of hypertension management. Third, it should extract key hypertension concepts according to the situation of communication and learning [8,9].

The hypertension management involves hypertension diagnosis and detection, risk factors of hypertension, pushing health information, diet management, sports management, medication management, non-pharmaceutical therapy and other concepts. As result, the key contents of hypertension management are shown in Figure 1.

Figure 1. Key Concepts of Hypertension Management in Chinese

Then, it continues to find out the hierarchical relationship among different concepts. For example, if hypertension and diabetes are belonged to a specific disease of chronic disease, then chronic disease is created as a parent class for hypertension class and diabetes class. Similarly, food is considered as parent class to two sub classes that processed food and non-processed food. This process helps to clearly show the hierarchical relationship. For the hierarchical relationship, we can seek medical professionals to revise and optimize the class hierarchy until the result is consistent with the expectations.

The other work is to map relationships among concepts. The common relationship includes the equal relationship and disjoints relationship. The object attributes and data attributes associated with concept need to be identified, and the hierarchies between them need to be defined clearly. For example, it requires to list attributes, such as drug's expiration date and foods' calories. Then, we should set the limits of attributes' domain and range to build the triple relationship in the form of class-object attributes-class or class-data attributes-specific data. For instance, coconut bread (class) - calories per 100g (data property) -320KJ (data) means that each 100g coconut bread contains 320 thousand calories. Calcium antagonist (class) - non applicable disease (object property) - heart failure (class) indicates that calcium antagonists cannot be used to patients because they are occurred with heart failure and hypertension.

Finally, professionals are invited to evaluate the entire knowledge base model, including indicators that integrity, accuracy, and usability. If possible, the hypertension Ontology should be improved according to the evaluation results.

2.2 Data Attribute Processing Method

The standard language of OWL has defined *minInclusive* and *maxInclusive* as fixed vocabulary to represent the values of the lower and upper bounds. To follow these rules in storage structure, it is to translate these words into Chinese when displaying information in external. We create a static class *TranslateMap* to store the mapping relationship between English and Chinese vocabulary, by which the public method *translate*() is used to provide services. In class, member variables are stored in static map, English terms are stored in the key domain, and the corresponding Chinese terms are stored in the value domain. Due to map is static, it doesn't need to create instance repeatedly, which saves the time and memory. Considering the frequent problem, the hash index will be used to improve the execution efficiency. In the best condition, the time complexity is worked at constant level.

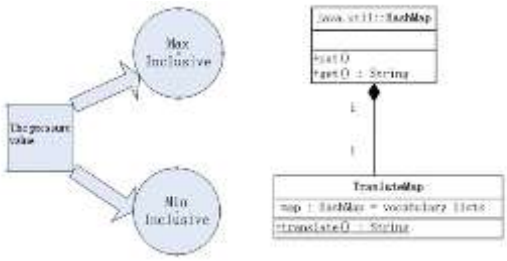

Figure 2. Definition of Specific Range of Data Attributes

As shown in Figure 2, it gives implementation details. In the process of practical operation, *TranslateMap* is invoked to display Chinese characters. If the corresponding dictionary does not exist, then it returns the original string.

2.3 Processing of Attribute Characteristics

The standard language of OWL defines 6 attributes (predicates) characteristics, including transfer characteristic, symmetrical characteristic, function characteristic, inverse function characteristic, reflexive characteristic and reciprocal characteristic. To this point, we consider whether the specific attributes have these six characteristics when creating them, which plays an important role in the accuracy of Ontology definition [10,11].

2.3.1 Transfer Characteristic

For any resource: X, Y and Z, P is attribute that acts on X, Y and Z. If there exists P(X, Y) and P (Y, Z), then there is P (X, Z). For example, the class chronic disease we created is the subclass of root class *Thing*. And chronic disease contains the sub class hypertension. Hypertension should also be a subclass of Thing. Thus, the attribute that X is a subclass of Y should be defined to have transfer characteristic. In the hypertension management ontology, it is needed to record the user's address. For example, someone's location is in Hangzhou, meanwhile the location of Hangzhou is in Zhejiang Province. As result, it is clear that someone's location can also be Zhejiang Province. Hence, the attribute location *lacateIn* also should be defined to have transfer characteristic.

2.3.2 Function Characteristic

For any resource: X, Y and Z, P is attribute that acts on X, Y and Z. If there exists P(X, Y), P (X, Z), and Y=Z then it is function characteristic. In hypertension ontology, the normal systolic pressure class has the attribute describing the range of blood pressure. In the actual scene, the value of this attribute must have a fixed range from 20 to 139 *mmgh*, which is considered as health. Thus, the range attribute of blood pressure should be defined to have function characteristic.

2.3.3 Inverse Function Characteristic

For any resource: X, Y and Z, P is attribute that acts on X, Y and Z. If there exists P(Y, X), which make P (Z, X) to contain Y=Z, it is inverse function characteristic. In hypertension ontology, once a specific range of blood pressure value is fixed, such as 140-400*mmgh*, which means that the corresponding value in definition domain is high systolic blood pressure. Thus, the range value of blood pressure has both function characteristic and inverse function characteristic. It means that elements in the domain should be the on-to-one mapping relation.

2.3.4 Reflexive Characteristic

For any resource: X, P is an attribute that acts on X. P has reflexive characteristic is equivalent to P (X, X). For example, coffees have the same color, which means that the same color is a reflexive attribute. In hypertension Ontology, the parent class is defined as reflexive characteristic, which means that a class is also its parent class.

2.3.5 Reciprocal Characteristic

For any resource: X and Y, P1, P2 are attributes that act on X and Y. If P1 is an inverse attribute of P2, for all X and Y: P1 (X, Y) holds if and only if P2 (X, Y) holds. For example, A is a subclass of B, which is equivalent to the fact that B is the parent class of A. Therefore, possessing parent class and possessing subclass are two reciprocal attributes. In the diet management of hypertension ontology, French manufactory fabricates wine is equivalent to the manufacturer of wine is French manufactory.

2.4 Naming rules

In the construction process of hypertension Ontology, the sentences are divided into three blocks that subject, predicate and object, which cannot express a clear semantic meaning respectively. Hence, there is a need for each block to make corresponding naming rules so that a single subject (predicate or object) can express clear semantic meaning. This problem can refer to the English Ontology. However, the differences between Chinese and English include grammar, which calls for adopting the appropriate solution according to the characteristics of Chinese grammar.

2.4.1 Class and Instance Naming Rules

Class and instance can be used as subject or object, where both of them can express the complete semantics. However, due to the differences in English and Chinese grammar, English is always able to express the attribute of the word, while Chinese does not have this ability. For example, words ‘Hypertension’ can clearly show that this is the hypertension in chronic diseases.

Table 1. Examples of Naming Class and Instance

English	Chinese Meaning
Hypertension	高血压慢病
Drug Therapies	药物治疗方法
Overweight	超重的状态

While in Chinese, three words ‘高血压’ cannot express its attributes clearly. It can be a chronic disease named hypertension. It may refer to the status of blood pressure. Therefore, it is necessary to emphasize the attributes of Chinese words, such as “高血压慢病”. Typical examples are shown in Table 1.

2.4.2 Attribute Naming Rules

Attributes are often considered as predicates. Due to the connection characteristics of predicates, it requires a combination of subject and object to express complete semantics. So naming a predicate for complete semantics means high complexity. UMLS (Unified Medical Language System) is in the form of verb and noun, such as *hasMaker*, *hasBloodPressureRange*, *isSubClassOf*, etc. In order to express the semantic meaning of the attribute clearly, it is required to add verb or other definite noun to accurately express the semantic meaning. Generally, the prefix “has” needs to be expressed as having XX or

owning XX. And the prefix “is” needs to be expressed as the entity XX. In general, texts containing “XX from” express as the object made up of the latter XX. Typical examples are shown in Table 2.

Table 2. Typical Examples of Naming Attribute

English Term	Chinese Meaning
hasMaker	拥有制造者
hasBloodPressureRange	具有血压值的范围
isSubClassOf	是后者的子类
madeFromGrape	是由葡萄加工制成的

2.4.3 Literal Constant Naming Rules

In standard language of OWL, the literal constant is only used as object. It is equivalent to the edge node in Ontology. This kind of literal constant is relatively simple, where we need to pay enough attentions to the integrity of expression.

3. The Ontology Searching Based on Chinese Word Segmentation

Word segmentation technique is the cornerstone of the full text search query, which impacts on the Ontology searching. In this paper, the query module refers to the analyzer interface of Lucene [12,13] to achieve word segmentation in the field of chronic disease management.

3.1 The Thesaurus Construction

The word segmentation needs a thesaurus to work, which is a collection of meaningful phrases to guide system to know the query intention. During the word segmentation, we should put a sentence into the correct phrases reasonably. For example, "高血压的不可变危险因素" in Chinese, which can be decomposed into phrases that "高血压", "的","不可变","危险","因素". If the result is that "高血","压","的不","可变因","素", it has no specific meaning[14,15].

There are a lot of general open source thesauruses on Github[16]. The most famous projects are IK and stuttering. In the former, there are more than 270 thousand phrases. In the latter, there are more than 340 thousand. In order to reduce errors, it is necessary to take the union of two thesauruses.

If the thesaurus needs to update dynamically, the time-consuming is too large to be acceptable. We consider using the hash function for each phrase. In the matching process, we directly compare the hash function value, which only costs 340 thousand on average. The operation in the same test machine only takes 8ms. Obviously, the optimized merging algorithm improves the performance. As shown in Figure 3, it gives the merging algorithm for Chinese Thesaurus.

```
public String[] Merge(String[] A,String[] B)
{
    Set<String> set=new HashSet<String>();

    for(String s:A)
    {
        set.add(s);
    }
    for(String s:B)
    {
        if(!set.contains(s))
        {
            set.add(s);
        }
    }
    return (String[]) set.toArray();
}
```

Figure 3. Chinese Thesaurus Merging Algorithm

The open source thesauruses lacks of professional vocabulary in the field of hypertension management. Therefore, we take the Ontology established in Section 2 as the data source, and extract all the key words to build the professional thesaurus of hypertension management, including class, instance, object attribute, and data attribute.

3.2. Building Word Searching Tree

The segmentation thesaurus stored in the form of text sequence, which needs to be loaded into memory to be used and read by the matching algorithm. However, the efficiency will be low, which contributes to accelerate the searching speed.

The English words take the simple dictionary tree data structure. In this way, each node has 26 links, each link corresponds to the 26 letters from A to Z, as shown in figure 4. Each node can contain a value which is associated with a keyword, or it can simply be a node without content. As result, the word searching tree is usually built with a large number of empty links, which are always ignored when drawing a dictionary tree.

Figure 4. Dictionary Tree Node

Suppose we want to query a word. Each character in this word will be used as a guide. Each node in the dictionary tree contains links to all characters that may appear next. We need to start from the root node and match them in turn. The first work is the link corresponding to the first letter of the given word. Then, it works to search the second letter of the given word. We redo this process until meeting one of the following three cases:

1) The given word has reached the ending, and the value of the corresponding node in the dictionary tree is not empty. This indicates that the searching is hit. The result is the value of the node corresponding to the end character.

2) The given word has reached the ending, but the value of the corresponding node in the dictionary tree is empty. This indicates that the result does not exist. Thus, the searching mishit.

3) The searching does not reach the end of the given word when it is encountered with an empty link. This indicates that the searching mishit.

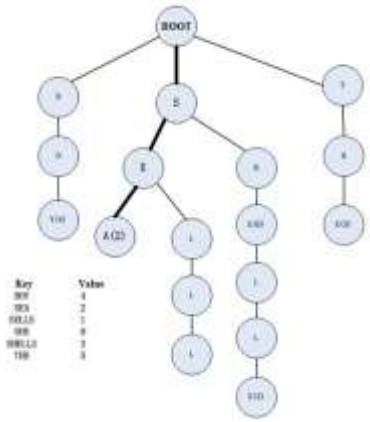

Figure 5. Search that Hit at the First Time

An example of the hit searching is as shown in Figure 5. In this case, the key word is "SEA". The searching process executes along with the bold black line. The node corresponding to the end character "A" has a value of 2, which indicates that this is a hit searching.

4. Experimental Analysis

4.1 Ontology View

As shown in figure 6, nine categories in the class view of protégé [17] include diagnostic methods, drugs, treatment methods, personal status, chronic disease, nutrition, blood pressure, food ingredients and food recommendation.

Figure 6. Ontology of Chronic Disease in Chinese

Considering the scalability, we choose "drugs" as the first class, and denote "the therapeutic drugs of hypertension" as its sub class, rather than making "the therapeutic drugs of hypertension" as the first class. In this case, the therapeutic drugs of other chronic diseases can be added subsequently, such as adding "the therapeutic drugs of diabetes" as the brother node of "the therapeutic drugs of hypertension". As shown in figure 7, it gives the reduced graph of Ontology structure in Chinese.

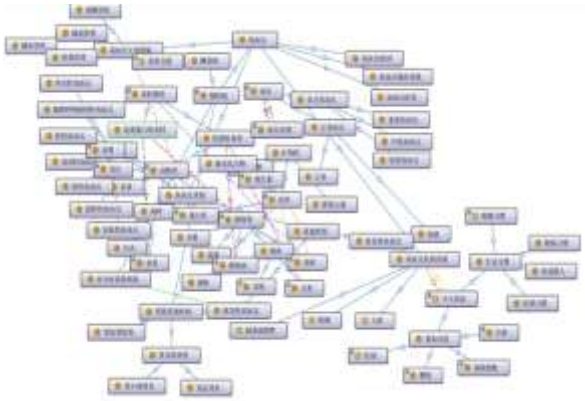

Figure 7. Reduced Graph of Ontology Structure in Chinese

4.2 Data Attributes View

Data attributes connect a class and specific data, or connect individual instances and specific data. According to the characteristics of the knowledge of chronic diseases, the data attributes include: "the year and date as", "the range value of", "value collection contains", "meaning is", and "percentage is".

Table 3. Data Attributes of Ontology

Data Attributes	Introduction
The Range Value of	Represent having the range of values
The Year and Date as	To represent values associated with year and date
Value Collection Contains	To show the contain specification when data type is collection
Percentage Is	To show the percentage of each element
Meaning Is	To explain the specific definition and meaning, etc.

3.3 Performance Comparison

In the experiment, we carry out comparison among different word thesaurus, mainly the dictionary tree with sequential order, the traditional dictionary tree, and the improved dynamic dictionary tree. The experiment observes the time and memory consuming. The running environment is based on a computer with Intel core i5 3470 x86 system. The operating system is windows10 professional version 64. JDK version is 1.8. Tomcat version is Tomcat 7.

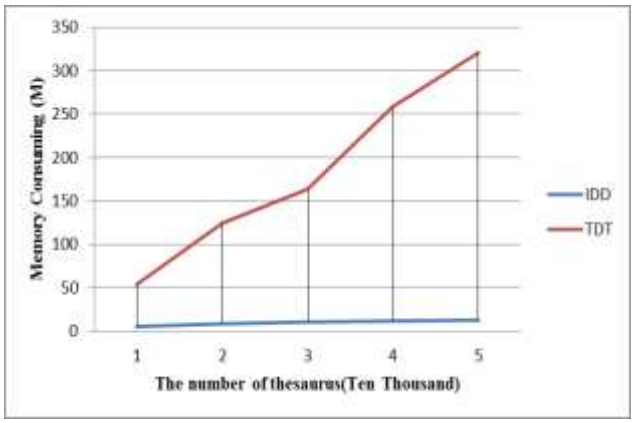

Figure 8. Memory Consuming Comparison

The experiment result of Figure 8 shows the memory consuming comparison between the traditional dictionary tree (TDT) and the improved dynamic dictionary tree (IDD). Both two methods increase linearly with the increasing of the number of word thesaurus. However, the traditional dictionary tree has bigger intercept and slope. Under the condition the number of words in thesaurus is 10 thousand, the memory consuming of the former is 12 times larger than the latter. Under the condition the number of words is 50 thousand, the memory consuming of the former is more than 16 times larger than the latter. It is clear that the improved dynamic dictionary tree has more effective searching, comparing with the original structure of the traditional dictionary tree.

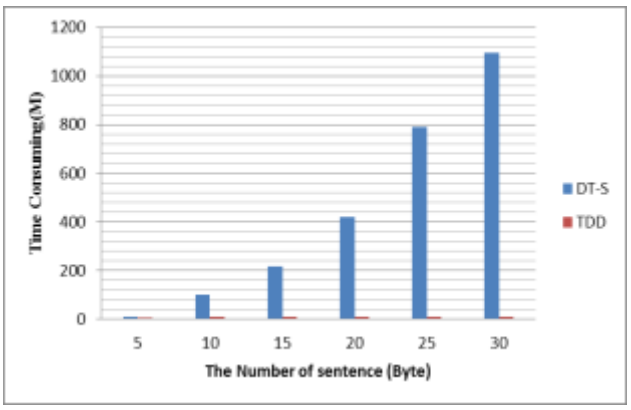

Figure 9. Time Consuming Comparison

As shown in Figure 9, it gives time consuming among the dictionary tree with sequential order (DT-S) and the traditional dictionary tree(TDD). The dictionary tree with sequential order shows parabolic increasing with the number of sentence length. However, the traditional dictionary tree does not show obvious changes with the increasing of sentence length. There is a big time gap between them. Thus, there is a great improvement on the time complexity in dynamic dictionary tree storage structure, comparing to the linear structure.

5. Conclusion and Future Works

The work of this paper includes Ontology construction and Ontology searching. The former is to build knowledge base in the field of chronic disease management of hypertension, and the latter is to use Chinese word segmentation to improve the query effectiveness. Finally, the implementation of prototype system is discussed, and experiments are carried out to demonstrate the effectiveness and feasibility of our method.

In the next step, we will take load balancing and optimization into consideration in order to make sure the system is reliable in the case of high concurrent operations.

Acknowledgement

This work was supported by Zhejiang Provincial Natural Science Foundation of China under grant No.LY15F02007.

References

[1] P. A. Laplante and N. Laplante, "The Internet of Things in Healthcare: Potential Applications and Challenges", IT Professional, vol. 18, no. 3, **(2016)**, pp. 2-4.
[2] M.-H. Kim and S.-C. Chang, "A consumer transceiver for long-range IoT communications in emergency environments", IEEE Transactions on Consumer Electronics, vol. 62, no. 3, **(2016)**, pp. 226 – 234.
[3] A. Gyrard, P. Patel, A. Sheth and M. Serrano, "Building the Web of Knowledge with Smart IoT Applications", IEEE Intelligent Systems, vol. 31, no. 5, **(2016)**, pp. 83 - 88.
[4] V. Angelakis, I. Avgouleas, N. Pappas, E. Fitzgerald and D. Yuan, "Allocation of Heterogeneous Resources of an IoT Device to Flexible Services", IEEE Internet of Things Journal, vol. 3, no. 5, **(2016)**, pp. 691-700.
[5] N. J. Cleven, J. A. Müntjes, H. Fassbender, U. Urban, M. Görtz, H. Vogt, M. Gräfe, T. Göttsche, T. Penzkofer, T. Schmitz-Rode and W. Mokwa, "A Novel Fully Implantable Wireless Sensor System for Monitoring Hypertension Patients", IEEE Transactions on Biomedical Engineering, vol. 59, no. 11, **(2012)**, pp. 3124 -3130.
[6] The Report of China Medical and Health Care, http://www.moh.gov.cn/ewebeditor/uploadfile/2013/01/20130103003113472.pdf.
[7] Unified Medical Language System (UMLS), https://www.nlm.nih.gov/research/umls/.

[8] K. Silachan and P. Tantatsanawong, "Domain Ontology Health Informatics Service from Text Medical Data Classification", Annual SRII Global Conference, **(2011)**.
[9] R. Benlamri and L. Docksteader, "MORF: A Mobile Health-Monitoring Platform", IT Professional, vol. 12, no. 3, **(2010)**, pp.18 – 25.
[10] R. V. Nava, V. H. M. Dominguez and J. G. Montalvo, "A Document Recommendation System Using a Document-Similarity Ontology", IEEE Latin America Transactions, vol. 14, no. 7, **(2016)**, pp. 3329 – 3334.
[11] Y.-W. Zhao, D. Sanán, F.-Y. Zhang and Y. Liu, "Formal Specification and Analysis of Partitioning Operating Systems by Integrating Ontology and Refinement", IEEE Transactions on Industrial Informatics, vol. 12, no. 4, **(2016)**, pp. 1321 – 1331.
[12] Y. Zhou, X.-Q. Wu and R.-Y. Wang, "A semantic similarity retrieval model based on Lucene", IEEE 5th International Conference on Software Engineering and Service Science, **(2014)**.
[13] J.-X. Chen, W. Wu and C.-Z. Wang, "A mobile phone information search engine based on Heritrix and Lucene", 7th International Conference on Computer Science & Education (ICCSE), **(2012)**.
[14] J. Tang, Q. Wu and Y.-H. Li, "An Optimization Algorithm of Chinese Word Segmentation Based on Dictionary", International Conference on Network and Information Systems for Computers, **(2015)**.
[15] C. Xiu and R. Song, "Study on the Influencing Factors of Chinese Word Segmentation", International Conference on Asian Language Processing, **(2012)**.
[16] Github. www.github.com.
[17] H. Knublauch, R. W. Fergerson, N. F. Noy and M. A. Musen, "The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications", Third International Semantic Web, vol. 3298, **(2004)**, pp. 229- 243.

Authors

Wang Hui, received his Master Degree in Computer Science from Inner Mongolia University of Technology, China, in 2013. His research interests include service computing, and cloud computing.

Qinyan Zhang, received her Master Degree in Computer Science from Zhejiang University, China, in 2016. Her research interests include service computing, process mining.

Yang Huang, received his Master Degree in Computer Science from Zhejiang University, China, in 2016. His research interests include service computing, and cloud computing.

Meng Xi, studies his B.E. Degree in Software Engineering from Zhejiang University, China. His research interests include service computing.