



HAL
open science

Research on Information Integration Method of Agricultural Products Producing and Managing Based on Knowledge Graph

Xiang Sun, Huarui Wu, Peng Hao, Qingxue Li

► **To cite this version:**

Xiang Sun, Huarui Wu, Peng Hao, Qingxue Li. Research on Information Integration Method of Agricultural Products Producing and Managing Based on Knowledge Graph. 11th International Conference on Computer and Computing Technologies in Agriculture (CCTA), Aug 2017, Jilin, China. pp.232-243, 10.1007/978-3-030-06137-1_22 . hal-02124251

HAL Id: hal-02124251

<https://inria.hal.science/hal-02124251v1>

Submitted on 9 May 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Research on information integration method of agricultural products producing and managing based on Knowledge graph

Xiang Sun^{1,2,3}, Huarui Wu^{1,2,3} (✉), Peng Hao^{1,2,3}, and Qingxue Li^{1,2,3}

¹ Beijing Research Center for Information Technology in Agriculture, Beijing Academy of Agriculture and Forestry Sciences, Beijing, PR China

{sunx,wuhr,haop,liqx}@nercita.org.cn

² National Engineering Research Center for Information Technology in Agriculture, Beijing, PR China

{sunx,wuhr,haop,liqx}@nercita.org.cn

³ Key Laboratory for Information Technologies in Agriculture, Ministry of Agriculture, Beijing, PR China

{sunx,wuhr,haop,liqx}@nercita.org.cn

Abstract. In order to improve the integration and access efficiency of agricultural information, this paper propose an agricultural information integration framework based on knowledge graph. A knowledge graph of agricultural products producing and managing was constructed, covering the basic process of "Planting - farming - processing - quality inspection - warehousing - Transportation - Sales" and realizing the storage, mapping and inquiry of knowledge graph. Improves the method of mapping data linkage based on database mapping relation, and realizes the transformation of elements from database to knowledge graph elements. Map data link method of database based on mapping relations, realize the conversion of database elements to the knowledge graph elements, the iterative discovery of relation and pattern in text information is realized by means of weak supervised machine learning method. This method integrates the application in the Green-Cloud-Grid platform, and improves the efficiency of information source integration, correlation analysis and mining utilization under the platform.

Keywords: knowledge graph · agricultural products producing and managing · information integration · weak supervision

1 Introduction

The quality of agricultural products has attracted public attention; it covers many aspects of agricultural products production related to agricultural supply, habitat environment, disease control, production management, agricultural products processing, distribution and other links. The information characteristics of the links are real-time, distributed, heterogeneous and massive. The information integration and acquisition efficiency is low, and the information asymmetry between producers, consumers and managers is one of the important reasons for the quality and safety of agricultural products.

In recent years, many scholars have carried out a lot of research and application work for information integration, such as Liu Jiawei [1] through the domestic and foreign agricultural information related fields and knowledge graph tools, drawing

construction methods comparative analysis, to explore the knowledge graph technology applications of agricultural information service. H Xiao [2] describes the semantic heterogeneity of information system integration, introduces agricultural ontology representation, domain knowledge sharing and reuse, as well as conceptual similarity computing method. Zhao Ming [3] takes Baidu encyclopedia plant category entry text information as the corpus, and uses ontology learning method to achieve better results in the extraction of non-taxonomic relations.

Although the existing information integration technology research and application has made some progress, but the ontology information integration method is difficult to adapt to the scale of growth massive data every day, in addition to regular information extraction by Information extraction and storage is difficult to meet the requirement of real-time application in agricultural field, greatly affected the effectiveness of agricultural production diagnosis and management decision.

Knowledge graph is a massive information management and service model based on information units, which can draw, mine, analyze and display domain information and their relationship with each other [4]. As a carrier of information, it graphically represents people's information images which are difficult to understand, through the content analysis, citation analysis and visualization of information structure and its relationship with cognitive habits of human beings, the user can quickly obtain the logic and the relationship between information.

In order to solve the above problems, this paper constructs the knowledge graph of agricultural products production and management information integration method based on the relationship between knowledge discovery and extraction of large blocks of information value high density, to provide technical support for the integration and utilization of quality and safety data of large-scale agricultural products.

2 An Information Integration Framework for Agricultural Products Based on Knowledge graph

This paper proposes a framework of agricultural product information integration based on knowledge graph, which can be divided into three levels: agricultural products producing and managing ontology, knowledge graph database and knowledge graph application(fig.1).

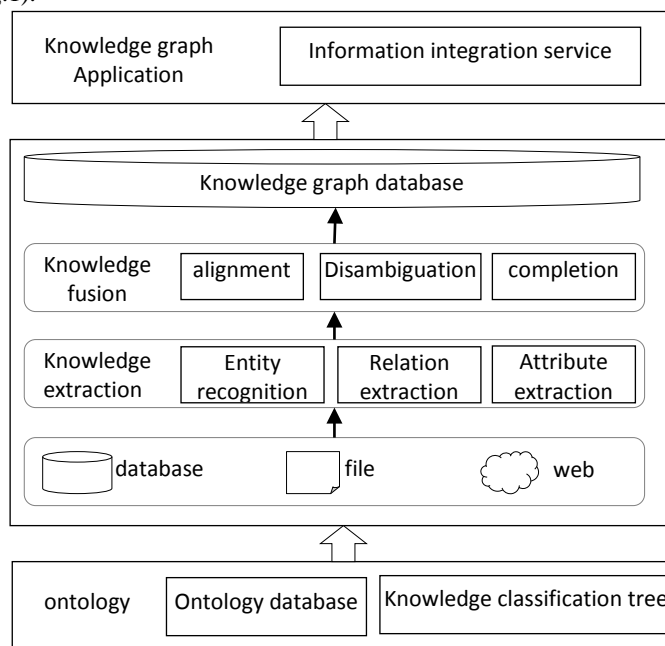


Fig. 1. Knowledge graph framework

The ontology of agricultural products producing and managing: As the foundation of establishing knowledge graph, it abstracts and constrains the concepts, attributes and relations of related information. For example, the concept of "fruit" has the characteristics of variety, picking time, fruit tree age, occurrence of diseases and insect pests, fertilizer input and growth environment; "diseases and insect pests" include the types of the invasion, the location of the injury, the period of the invasion, symptoms, prevention and treatment methods and other concepts; "diseases and insect pests" and "fruit" are related to diseases and insect pests attacked fruit trees, fruit is found in fruit trees. In addition, the following concepts of "varieties" include apples, pears, grapes, peaches, cherries and so on. Here, a knowledge classification tree is defined to describe the classification of the concepts and entities, and the relationship between the upper and lower, and fruit knowledge system is formed.

Knowledge graph database: The agricultural production management database, agricultural product quality testing database, market circulation database, technology and market data of agricultural websites as the information source, using the bottom-up approach, first through the knowledge extraction obtained the entities, attributes and relationships in the source information, and then through the knowledge fusion to alignment and matching the entities, then add the confidence extraction results to the knowledge graph, so as to complete the construction of knowledge graph database.

Knowledge graph application: Realization of whole process information integration service of agricultural products based on knowledge graph database.

As Fig.2, information integration based on knowledge graph can be formalized into the triples, such as:

$$O = \langle T, M, S \rangle \quad (1)$$

In the formula, T is the knowledge graph, that is, concept and relation description, M is semantic mapping, and S is the data source scattered everywhere.

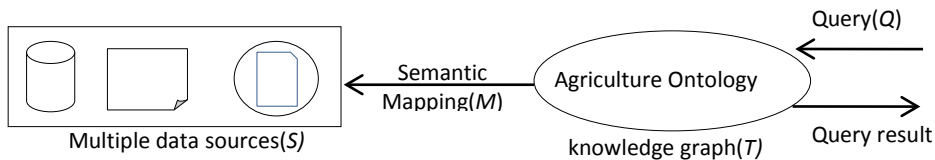


Fig. 2. Information integration process

3 Construction of Knowledge Graph of Agricultural Products Producing and Managing

Knowledge graph are used to describe the relationships between agricultural information entities. As a liaison network, knowledge graph can realize the mutual exchange of information, databases, file resources and systems related to agricultural products producing and managing.

The agricultural management department needs to manage the information of agricultural enterprises within its jurisdiction. The traditional data management model includes relational model, hierarchical model and mesh model, and the above model is difficult to realize application semantic extension. Based on the relational model, semantic model can express the complex structure and rich semantic of agricultural information more naturally.

Agricultural products producing and managing covers the basic process of "Planting - farming - processing - quality inspection - warehousing - Transportation - Sales" and realizing the storage, mapping and inquiry of knowledge graph, first defines the top class for the semantic model according to these processes. Secondly, the method of "up and down" is used to define the concept of generalization in the field, and then gradually refine the description, determine the attribute information and the relationship between the knowledge objects.

AS Fig. 3, the semantic model of agricultural products producing and managing uses

the attributes of `isComposedOf(Agro-operationflowEntity)` and `isMemberOf(Agro-operationflowEntity)` to construct model hierarchy, Uses the attributes of `Contains(CriticalinformationEntity)` and `containedBy(CriticalinformationEntity)` to construct the inclusion relation among structures. `Manages(roleEntity)`, `consumers(roleEntity)`, `farmer(roleEntity)`, `experts(roleEntity)` and `Technician(roleEntity)` are used to distinguish between different roles. `HasMeasurement(handleEntity)` and `associatedTo(handleEntity)` are used to establish the extended relationship between attributes under different levels of structure.

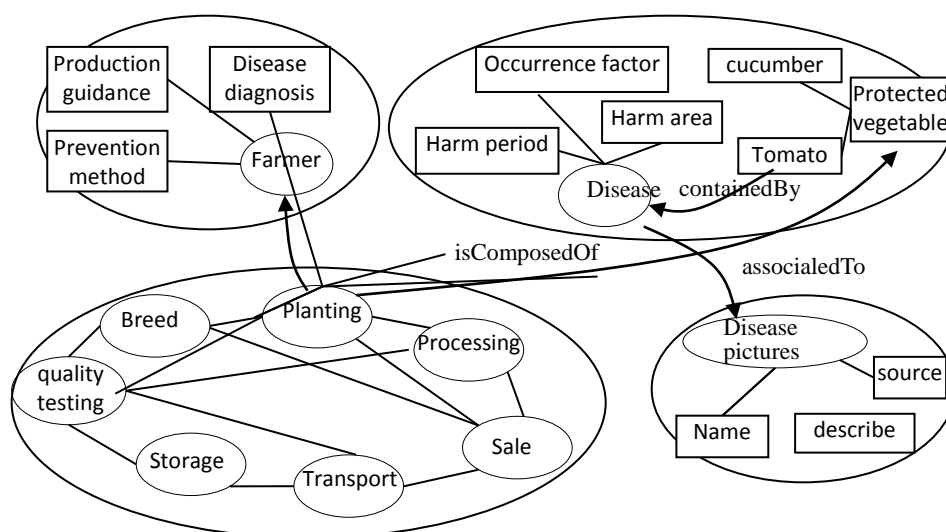


Fig. 3. Knowledge graph semantic association model

Based on the above knowledge graph semantic association model, the knowledge graph is constructed based on the following key steps:

- Determine the range of knowledge graph. Through the knowledge graph, we can establish the correlation between the information of the production and management of agricultural products. A farmers' professional cooperatives have more than one agricultural base for example, each base has a plurality of production plots, each production block information involves planting varieties, planting history, soil and environmental data, real-time video, harvest time and yield; unified purchasing, processing and marketing cooperatives after harvest of agricultural products, information includes pesticide detection, packing date, order information, logistics data etc.. Using the semantic model of map to organize information, the query feedback can be carried out quickly through the semantic relation of various kinds of information.
- Define the classification of information and integrate the terms in Agricultural Thesaurus. To describe the production and operation of agricultural products "planting aquaculture processing quality inspection warehousing transportation sales" of the important link hierarchy, use classification information to describe the concept of entity and to describe the hierarchy between concepts with attributes.
- Through knowledge learning methods, the data instances of knowledge graph are enriched, and the semantics of sub class relations are embodied. If the class F is a subclass of class G, the properties of all instances of class G, as well as instances of class F, are also available. Therefore, attributes can be defined by class inheritance, and they are constantly adjusted and updated dynamically.

4 Intelligent Extraction Method of Knowledge Graph

Select the typical concepts and characteristics of agricultural products producing and managing, establish the core ontology, and then extend the core ontology based on machine learning method to extract entities, attributes, values and relationships in the knowledge source. Since there are many sources of knowledge and many types of information, we adopt different extraction methods for different information sources.

Agricultural databases are structured knowledge, and D2R

(Relational Database to RDF

mapping methods are used to translate agricultural databases into link data in knowledge graph[5]. Semi-structured knowledge such as tables in web pages is extracted using wrapper based (Wrapper) extraction methods [6]. For unstructured text knowledge, an incremental iterative extraction method with weakly supervised machine learning is adopted.

4.1 Mapping data linking method based on database mapping relation

Analysis of agricultural database structure, through analyzing the database name, table name, column name, column values, records, table constraints, establish the knowledge classification tree and obtain knowledge graph concept, entity, attribute and attribute value, realize the database elements to the knowledge graph converting.

Table 1. Transformation mapping

Database	Knowledge classification tree	Knowledge graph
Database name(D)	Primary classification	Upper layer concept(C)
Table name(T)	secondary classification	concept(C)
Column name (L)	Three stage classification	attribute(A)
Column value(L _v)	/	Attribute value(V)
Record(B _i)	/	Entity(E)
Table constraint	/	Relationship(R)

As you can see from the previous table, a record B_i in table T, conceptually belongs to T and defines C(T) as a conceptual node, E(B_i) as the entity corresponding to the recorded data, and the conceptual node C(T) pointing to the entity node E_i(B). The entity E_i(B) contains the column names A_i(L) and column value V_i(L_v) of table T.

A constraint relationship is defined between tables in a database. The column M in table Ta and the column Tb in table N are foreign key constraints, that is, when Ta_i(M) = Tb_i(N), then E(a_i) and E(b_i) have conceptual relationships R(a,b).

The table Tc in the database stores table Ta, a record corresponding table, Tb multiple records, the establishment of table Ta, the entity E_i(a) and table Tb, the entity E_i(b) constructs the concept relation R(c).

As Fig.4 shows the three relationships, Ta_diagnose, Tb_diseases, and Tc_symptom; Dise_id and Sym_id are the primary keys of the relationship; Ta_diagnose, the foreign key Dise_id, refers to the Dise_id in Tb_diseases.

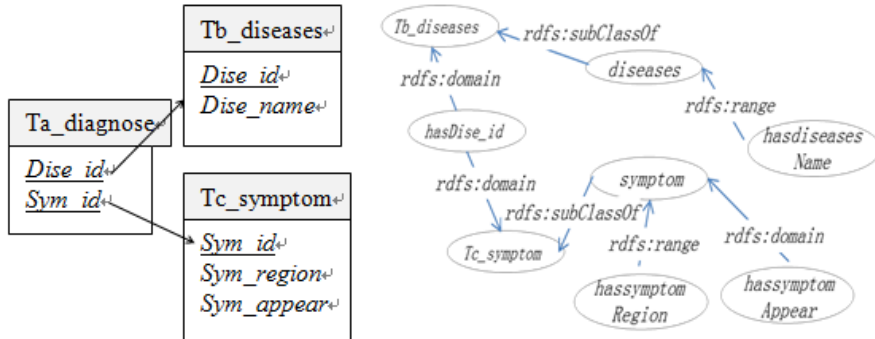


Fig. 4. Relational schema transformation

4.2 Incremental iterative extraction of knowledge graph relation for weakly supervised machine learning

This paper analyzes the unstructured text agricultural information, and realizes the transformation from nonstructural agricultural information elements to knowledge graph elements. The method of supervised machine learning is adopted in this paper. A small amount of tagged corpus is used to define the relational schema and relational instance as the initial seed set, and new relational templates and instance are discovered by machine learning [7, 8].

Weak supervision is also known as remote supervision. Extracting an entity relationship in this way generally involves two procedures: first look for entity statements that contain known relationships, and then use these statements as a training set for classification.

The earliest weak supervisor method is the DIPPRE method [9]. Using the correspondence between the entities and the relationship between the description relationships, starting from a set of seed relations, the context in which these seeds appear is found, and then the correspondence is generated from these contexts, Then use these patterns to find more relational instances from the text, and then select the new seed set from these relational instances, iteratively derive the corresponding relationships and patterns from the text.

In this paper, "planting", "breeding", "processing", "quality inspection", "storage", "transportation" and "sale" as the first level classification labels are adopted to determine the contextual entries in the text through the crawler. The main processes include:

- The choice of agricultural science thesaurus and small part of typical agricultural products and the characteristics of terms and words into words dictionary, a table of the data stored in the data segment all results, calculation of each word feature vector based on the thesaurus;
- Set up the part of speech in the corpus, store the corresponding part of speech in the corpus, and calculate each word based on the feature vector of the part of speech;
- Establish vocabulary to store these entries, and obtain the feature vectors of each word based on the vocabulary. Pressure
- Establish the feature vectors of context, window, word list and part of speech of each word, and realize semantic annotation of other data.

5 Information Retrieval Based on Knowledge Graph

After the extraction and recognition of the semantic relation between the entity level relation and the domain entity in the field of agricultural product production and management, the domain knowledge graph can be stored, drawn and searched.

Gayley tools were used to build our knowledge graph. Gayley uses MongoDB NoSQL database persistent storage, imports N-Quads file format stored knowledge graph. The N-Quads file determines which category is represented by the "subject predicate object", identifies the triples, and then converts each relation in the knowledge graph into triples, and serves as a record, stored as a N-Quads file in line [10, 12].

We use the Gremlin language interface in the Gayley service to query the knowledge graph [13, 15]. The node query is done by traversing the edge pointing relation and the edge name. As shown in table 2:

Table 2. Example of Gremlin query

Query requirement	Sentence	Result
Small fruit disease	g.Q('Small fruit'). Out(disease)	Iron deficiency, nitrogen deficiency, manganese deficiency.....
Symptoms of manganese deficiency	g.Q('manganese deficiency').Out(Symptoms)	Among the veins of pale green spots or spots, the softening of the fruit and leaf symptoms in the night side, small fruit tree

It is obvious that the knowledge contained in relational databases can be effectively organized and displayed by constructing relational databases into knowledge graph. At the same time, through the knowledge graph and query engine, we can use simple and flexible query statements to efficiently query knowledge. Users need to write information requirements into a specific format of the query, in order to obtain the appropriate knowledge, it requires users to have professional knowledge and training.

This paper supports users to query using natural language, and the steps include: factor detection, template matching, query generation.

factor detection: First, the natural language query is defined as the four elements of

attribute value (V), relation (R), entity (E) and concept (C), then, the lexicon are established according to the names of concepts, entities, and relationships in the knowledge graph.

When a user enters a natural language query, the natural language query is segmented into words and phrases using a natural language segmentation algorithm based on the lexicon.

For example, "what are the symptoms of manganese deficiency?" The result of the participle is: < what, V>, <Mn, C> < lack, R> < symptoms, E>.

Template matching: according to the segmentation results obtained in step 1, the natural language elements model is obtained.

For example, the element mode of "what are the symptoms of manganese deficiency?" is <V><C><R><E>. By querying the template library, you can implement the mapping of the element schema to the query schema.

Query generation: fill in the query model obtained in step 2) with the element values obtained in step 1), and generate knowledge graph query statements. For example, "what are the symptoms of manganese deficiency?" after step 2, the query pattern is g.Q(E).In(R).And(g.Q(C).Out()). Factor values obtained in step 1) such as < what, V>< Mn, C>< Lack, R>< Symptoms, E> are filled into query mode, get the query: g.Q (symptoms).In (lack).And (g.Q (Mn).Out), execute the query can be obtained all the symptoms of manganese deficiency entity.

6 Applications and Result Analysis

The knowledge graph in this paper is integrated in the Green Cloud Grid Platform. The green cloud platform has brought together the Internet of things data, production history and market circulation information of the more than 560 agricultural production bases in the country, and the daily price data of agricultural products in more than 1000 wholesale markets of agricultural products. With the help of knowledge graph, the platform realizes the semantic mapping between the map and various data sources, forms the data organization network of the platform, and finds the knowledge contained in it through the association analysis among the knowledge objects.

For example, in the retrieval of an expected production base of agricultural products, agricultural products can also get the planting time, fertilizer input, disease and medication, time to market information, and analysis for the same period surrounding the prices of agricultural products, yield and growth period of meteorological data, so as to arrange the picking plan and production under the guidance of a wheel.

The platform of knowledge graph covering vegetable, aquatic products, livestock, crops and more than 120 varieties, establishing 3489 information template, including 114 thousand ontology instance of planting, breeding, processing, quality inspection, warehousing, transportation, sales and other aspects seven links, extended resources including text, pictures, video, file, database etc. The amount of resources reached 5.2TB, improve the agricultural products producing and managing information integration and utilization efficiency of intelligent service.

Fig.5 is the process of cucumber disease diagnosis, which uses the knowledge graph established in this paper to correlate all kinds of resources needed in the decision-making process, and improves the accuracy of decision-making.

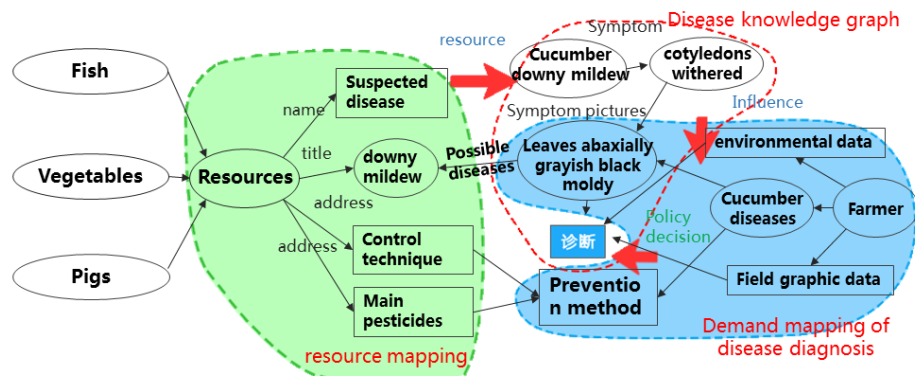


Fig. 5. Semantic relatedness of multiple resources related to cucumber diseases



Fig. 6. Cucumber diagnostic interface of the platform

7 Conclusions

The rapid development of information technology in agriculture has led to the overload of information caused by the massive growth of data, and reduced the efficiency of agricultural engineers to obtain decision-making knowledge from information.

This paper presents an information integration framework for agricultural products production and management based on knowledge graph, and realizes the automatic construction of knowledge graph. It integrated applications in the green cloud platform, provides a technical support for the integration, association analysis and mining of various information sources.

Based on the knowledge graph of agricultural information integration and mining decision-making has a wide range of application scenarios, can support the agricultural production, agricultural market analysis, agricultural enterprise management and other fields of auxiliary decision-making applications, follow-up in the heterogeneous data knowledge graph construction and user search intent to understand and other aspects of further research.

Acknowledgements. This work was supported by Beijing Natural Science Foundation (4172026), Natural Science Foundation of China (61771058), Innovation Capability Project of Beijing Academy of Agriculture and Forestry (KJCX20170706), and Beijing Natural Science Foundation Key Program (4151001)..

References

1. LIU Jia-wei, LIU Bo, SHEN Yue.: The Application of the Knowledge Mapping Based on Agricultural Information Services. *SOFTWARE*. vol. 3, 26-30 (2015).
2. Xiao H, Qiu T, Zhou P.: Integration of heterogeneous agriculture information system based on interoperation of domain ontology. In: 2th International Conference on Agro-Geoinformatics IEEE, pp.476-480, Fairfax (2013).
3. Zhao Ming Du Yaru Du Huifang Zhang Jiajun Wang Hongshuo Chen Ying.: Research on Ontology Non-taxonomic Relations Extraction in Plant Domain Knowledge Graph Construction. *Transactions of the Chinese Society for Agricultural Machinery*, vol. 47(9), 278-284(2016).
4. SERRA I, GIRARDI R, NOVAIS P.: Evaluating techniques for learning non-taxonomic relationships of ontologies from text. *Expert Systems with Applications*, vol. 41(11), 5201-5211(2014).
5. Jin Guiyang Lü Fuzai Xiang Zhanqin.: Enterprise information integration based on knowledge graph and semantic web technology. *Journal of Southeast University (Natural Science Edition)*, vol. 44(2), 250-255(2014).
6. HUANG Feng, WU Hua-rui.: An Automatic Generating Method of Adaptable Web Information Extraction Rule. *Journal of Guangxi Normal University: Natural Science Edition*, vol. 28(1), 127-130(2010).
7. HUAI Bao-Xing, BAO Teng-Fei, ZHU Heng-Shu, LIU Qi.: Topic Modeling Approach to Named Entity Linking. *Journal of Software*, vol. 25(9), 2076-2087(2014)
8. JIA Zhen, YANG Yufei, HE Dake, LIU Shengjiu, YIN Hongfeng.: Attribute and Attribute Value Extracted from Chinese Online Encyclopedia. *Acta Scientiarum Naturalium Universitatis Pekinensis*, vol. 50(1), 41-47(2014).
9. Brin S.: Extracting Patterns and Relations from the World Wide Web. *International Workshop on the World Wide Web and Databases*, pp.172-183, Springer Berlin Heidelberg (1998).
10. Xu Jian, Zhang Zhixiong, Wu Zhenxin.: Review on Techniques of Entity Relation Extraction. *New Technology of Library and Information Service*, vol. 24(8), 18-23(2008).
11. Guo Jian Yi, Li Zhen, Yu Zheng-Tao, Zhang Zhi-Kun.: Extraction and relation prediction of domain ontology concept instance, attribute and attribute value. *Journal of Nanjing University(Natural Science)*, vol. 48(4), 383-389(2012).
12. Albitar S, Fournier S, Espinasse B.: An effective TF/IDF-based text-to-text semantic similarity measure for text classification. *Web Information Systems Engineering*. Springer International Publishing, pp.105-114, Thessaloniki Greece (2014).
13. Christian Bizer, Jens Lehmann, Georgi Kobilarov, et al. DBpedia – A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 7(3), 154-165(2009).
14. Hu F, Shao Z, Ruan T.: Self-supervised Chinese ontology learning from online encyclopedias. *The scientific world journal*, vol. 2(11), 1-13(2014).
15. Guo J Y, Li Z, Yu Z T, et al.: Extraction and relation prediction of domain ontology concept instance, attribute and attribute value. *Journal of Nanjing University*, vol. 48(4), 383-389(2012).