DATA SCIENCE JOURNAL

Xie, N et al 2015 Research on an Agricultural Knowledge Fusion Method for Big Data. *Data Science Journal*, 14: 7, pp. 1–9, DOI: http://dx.doi.org/10.5334/dsj-2015-007

PROCEEDINGS PAPER

Research on an Agricultural Knowledge Fusion Method for Big Data

Nengfu Xie[1], Wensheng Wang[1], Bingxian Ma[2], Xuefu Zhang[1], Wei Sun[1] and Fenglei Guo[1]

[1] Key Laboratory of Digital Agricultural Early-Warning Technology, Agricultural Information Institute of Chinese Academy of Agricultural Sciences, Beijing 100081, China
xienengfu@caas.cn
zhangxuefu@caas.cn

[2] School of Information Science and Engineering, University of Jinan, Jinan 250022, China
ise_mabx@ujn.edu.cn

The object of our research is to develop an ontology-based agricultural knowledge fusion method that can be used as a comprehensive basis on which to solve agricultural information inconsistencies, analyze data, and discover new knowledge. A recent survey has provided a detailed comparison of various fusion methods used with Deep Web data (Li, 2013). In this paper, we propose an effective agricultural ontology-based knowledge fusion method by leveraging recent advances in data fusion, such as the semantic web and big data technologies, that will enhance the identification and fusion of new and existing data sets to make big data analytics more possible. We provide a detailed fusion method that includes agricultural ontology building, fusion rule construction, an evaluation module, etc. Empirical results show that this knowledge fusion method is useful for knowledge discovery.

Keywords: Ontology; Big data; Agriculture; Knowledge fusion; Information integration; Inconsistency

1 Introduction

Currently, most people use the Internet and the World-Wide-Web for browsing and getting information. In fact, however, you cannot obtain the complete, correct, timely information or knowledge that directly affects your judgment and decision-making in the web environment because of the heterogeneity of the information and big data scenarios. Knowledge fusion can be seen as an advanced information integration approach. Information integration focuses on how to find relevant information, but in knowledge fusion this information is merged to create knowledge that is more complete, less uncertain, and less conflicting than the input (Hu, Hu, Sekhari, Peng, & Cao, 2011). This reduces the cost of data access and enhances the value of the discovered data. The research on web-oriented knowledge fusion theory, methods, and knowledge of tools and development has become an important concern for knowledge-oriented service (Stegmaier, 2010; Wang, 2009).

In the 1960s, the international academic community began to research knowledge fusion, but early scholars did not explicitly put forward the concept of knowledge fusion. In the late 1980s, the rise of knowledge engineering increased attention to knowledge fusion. Feigenbaum (1983) put forward a "knowledge principle", in which knowledge fusion is one of the most important functional modules. Douglas Lenat's Cyc project, built upon Feigenbaum's knowledge principle, was an artificial intelligence project that attempted to assemble a comprehensive ontology and knowledge base of everyday common sense knowledge, with the goal of enabling AI applications to perform human-like reasoning (Lenat & Guha, 1990).

KRAFT (Knowledge Reuse And Fusion/Transformation) aims to develop a combination of database and artificial intelligence technology to allow scientists and engineers to find and exploit knowledge available on the Internet. KRAFT was a close collaboration between universities and industry (Preece, 2001). Based on

KRAFT, knowledge fusion has attracted many researchers. Hunter and Williams (2010) advocated a knowledge-based approach to merging semi-structured information. They used fusion rules to manage the semi-structured information that was input for merging. These fusion rules were a form of scripting language that defined how structured reports should be merged. The work assumed that structured news reports did not require natural language processing and used fusion rules to handle their inconsistencies and uncertainty. Fusionplex was a system for integrating multiple heterogeneous and autonomous information sources that used data fusion to resolve factual inconsistencies among the individual sources. To accomplish this, the system relied on source features, which were metadata, on the merits of each information source (Motro, 2004). A Dynamic Ontology Construction Method has also been proposed by analyzing knowledge requirements for more effective Knowledge Fusion (Liu, 2014).

In the next sections, we will discus the agricultural knowledge fusion problem and propose a general architecture for our fusion method. Finally, we will describe the knowledge fusion process in detail.

2 Related Techological Aspects

2.1 Agriculture Big Data Technologies

Agriculture big data means big data concepts, techniques, and methods practiced in the agriculture domain. In addition to having a vast body mass, modal variety and generating fast, low density value, agricultural big data are pervasive, contralateral, and have other characteristics. In the agricultural domain, agricultural production and research generate a large amount of data; in particular the application of information and communications technology (ICT) in agriculture will produce more in-depth, agricultural data soon achieving the ZB level. Integration and future mining of these data used for the development of modern agriculture will play an extremely important role. The big data technologies, including data-processing models and emerging tools, are being developed for implementation of our fusion system.

2.2 Ontologies

In general, ontology is an explicit specification of conceptualization (Stegmaier et al., 2010). Nevertheless, the term ontology has been controversial in current AI practice, and so far no formal definition exists. In our work, we have selected to use the term domain-specific ontology (DSO). In practical terms, developing an agricultural ontology (AgriOnto) includes three steps:

- building a domain-specific knowledge hierarchy;
- defining slots of the categories and representing axioms; and
- acquiring knowledge, that is to say, filling in the specific data values for slots.

2.3 Information Integration and Knowledge Fusion

Knowledge fusion appears naturally, and its related synonym is information integration. In detail, information integration focuses on how to find related information while knowledge fusion focuses on how to find accurate and complete information based on information integration. Therefore, knowledge fusion can be understood as a high-quality integration method, aimed at solving the conflicts of integration-based data; information documents can be integrated to guarantee that information is understandable by machines.

It is well recognized that information integration based on a ranking function has very limited value in selecting the correct value from diverse web resources because inconsistencies exist among information from different agricultural information sources. Our proposed approach is a six-step data flow process based information integration, called primary knowledge fusion (PKF) (**Figure 1**). First, it extracts related information from the PKF through a query. Second, the semantic analysis will be calculated if each piece of information is an instance of a concept of agricultural ontology (Agri-ontology) and the knowledge it contains. The third step annotates each instance according to the ontology. In the fourth step, the instances are clustered into different clusters by instance similarity. Next, the instances are fused according knowledge fusion rules. Finally, the fused result is evaluated and a new knowledge object (KO) produced.

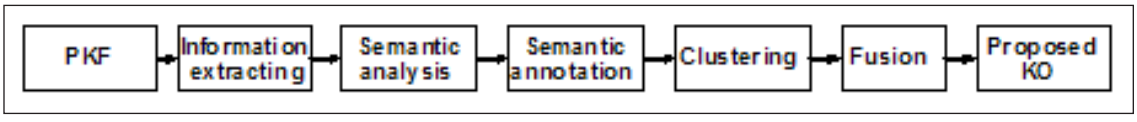

Figure 1: The six steps of our approach to knowledge fusion.

When multiple agricultural information sources provide inconsistent information, the knowledge fusion method is called upon to produce new information (knowledge) that is complete and accurate.

3 Agricultural Knowledge Fusion Model

The agricultural knowledge fusion method provides integrated knowledge and involves not only delivering available valuable information via links to users but also analyzing and merging the information results from agricultural information sources by solving result consistencies, removing duplicates, etc., based on agricultural domain ontology.

***Definition 1*:** Given a set of agricultural information sources (AISS), the PKF can be defined as a 3 triple such as PKF = (AISS, M, Q), where:

- AISS = $\{Is_1, Is_2, \ldots, IS_n\}$.
- M is the mapping between the global ontology and the ontology of AISS, defined as M = (Ω, O, g)
  - $\Omega = \{\Omega_1, \Omega_2, \ldots, \Omega_n\}$, Ω_i is the ontology of IS_i.
  - O is a global ontology.
  - $g(\Omega_i)$ is the mapping relation of Ω_i in the O.
- *Q* is the user query.

***Definition 2*:** Given PKF = (AISS, M, Q), the agricultural knowledge fusion problem is defined as AKF = (PKF, f, FR), where:

- PKF is the primary knowledge fusion.
- f is the operating function as $f(PKF) = \{\omega_1, \omega_2, \ldots, \omega_n\}$, and ω_i is the information instance annotated by the ontology.
- $FR = \{fr_1, fr_2, \ldots, fr_n\}$ is a set of knowledge fusion rules for attributes in agricultural ontology.

***Definition 3*:** Given AKF = (PKF, f, FR), the solution K satisfies:

- $\forall s \in slot(O)$, if $\exists fr \in FR$, then $K \cdot s = fr(s,\omega)$.
- $K \models Q$ means K is the answer to the Q. In this paper, K is the knowledge object and is described as $K = (K\_Name, ((s_1, v_1), (s_2, v_2), \ldots, (s_n, v_n)))$. We call (s_i, v_i) a knowledge unit. s_i is the slot attribute of a concept in ontology, and v_i is the value of s_i of an instance.

The above illustrates the agricultural knowledge fusion model in detail and gives a formal description of how to find a solution that merges the information from multi agricultural information sources into consistent knowledge that will answer users' queries.

4 Agricultural Knowlege Fusion Architecture

In this paper, we propose a general agri-ontology-based knowledge fusion architecture as shown in **Figure 2**. The architecture consists of three main aspects: 1) agricultural ontology and fusion rules are the cornerstones of the convergence of agricultural knowledge; 2) agricultural ontology-based knowledge representation and matching, as well as mining and automatically selecting fusion rules based on the property of concept, are the key components in knowledge fusion; 3) in order to find more accurate knowledge to satisfy users' queries, assessment of the fusion results is necessary to enhance knowledge fusion. All these parts form a complete system of knowledge fusion.

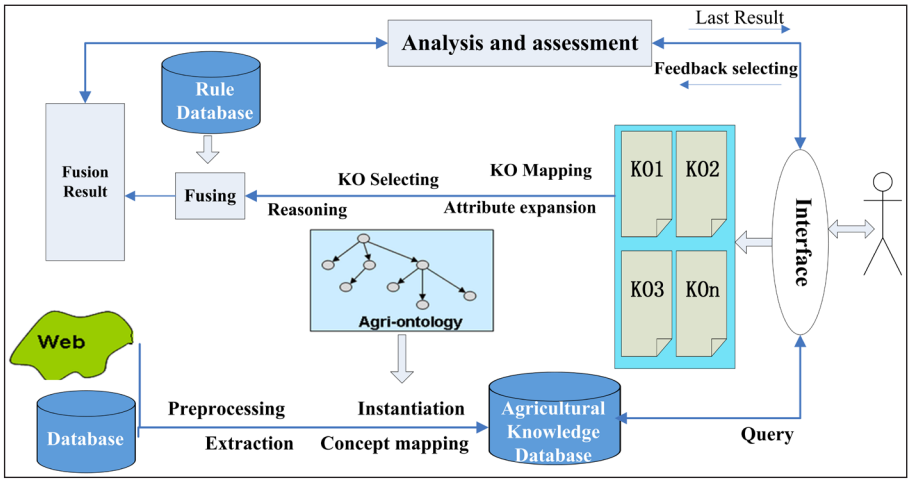

Figure 2: The agricultural knowledge fusion architecture.

4.1 AgriOnto

AgriOnto is the formal definition of agriculture and its relationships (see **Figure 3**). The definition and relationships form an integrated hierarchy of agriculture. With the labor object as the center of the agriculture hierarchy, we divide agriculture knowledge into seven taxa: labor object, production process, production technology, agriculture engineering, agriculture branch, agriculture environment, and agriculture regulation. Putting the labor object as the center of the agriculture knowledge hierarchy aims to aid those users who want labor object knowledge to access related knowledge of other taxa.

4.2 Fusion Rules

Each fusion rule, such as Min, Max, and Avg, can be looked as an aggregation function in the database (Xie et al., 2012). We divide fusion rules into two types: the single data fusion rule and the multi data fusion rule.

Definition 4: The single data fusion rule (SFR) is a type of aggregation function such that:

$$f : D1 \times D2 \times , \ldots , \times Dn \rightarrow D$$

where D_i is the value domain that has been unified as a domain so $D_1 = D_2 = ,\ldots, = D_n$. Given $v_i \in D$ $(i = 1,2,\ldots, n)$, $f(v_1, v_2,\ldots, v_n) = v$, $v \in D$. In this paper, the SFR includes Majr(Majority rule), Max, Min, Avg, Minr (Min-Priority rule), etc.

Definition 5: The multi data fusion rule (MFR) is a type of aggregation function such that:

$$f : D_1 \times D_2 \times \ldots \times D_n \rightarrow 2^D$$

given $v_i \in D$ $(i = 1, 2,\ldots, n)$, $f(v_1, v_2,\ldots, v_n) = D'$, $v_i \in D$, $D' \subseteq D$. The MFR includes CInt (Interval Rule), Or, and And.

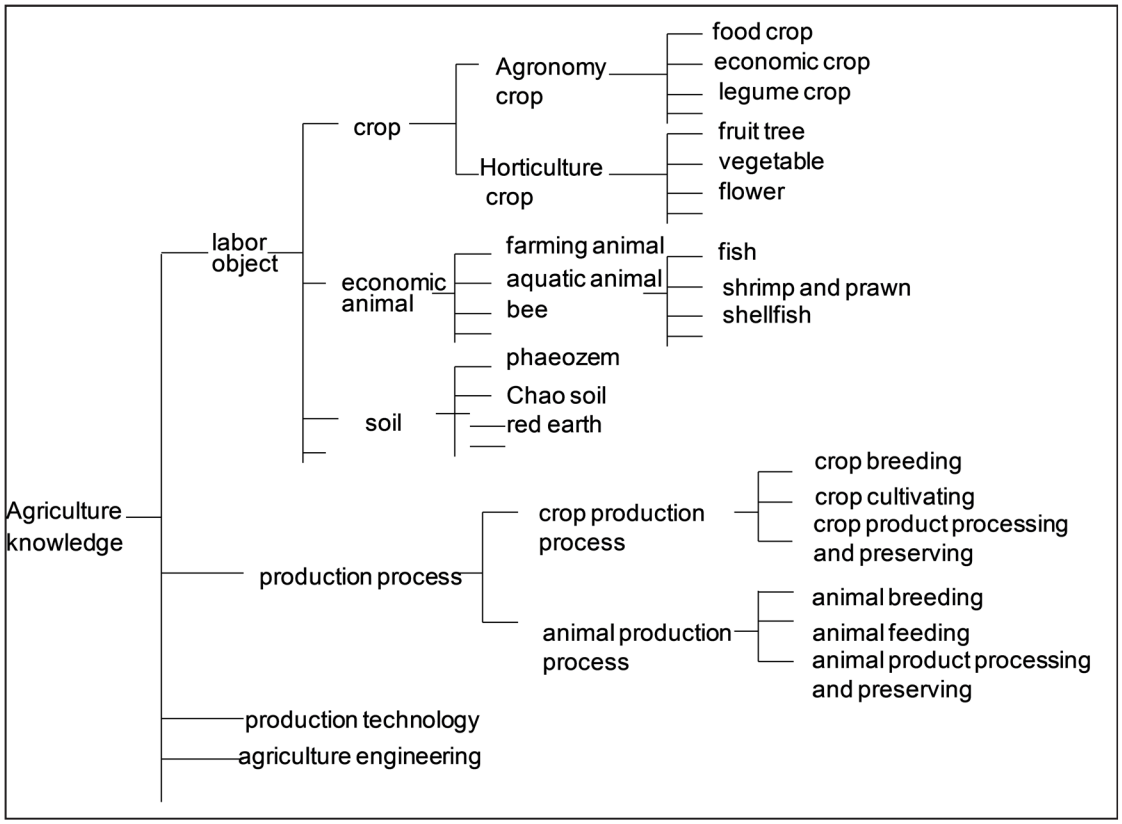

Figure 3: Some parts of the agriculture knowledge hierarchy. A favorable hierarchical hierarchy of agriculture knowledge is very useful to building an agriculture ontology. Our AgriOnto is built on this hierarchical structure.

In general, the single data fusion rule and the multi data fusion rule cannot be applied to an information set. Instead, we must analyze the query and answer type and then define a combination of fusion rules. However, usually the user participates in the rules selection to finish the knowledge fusion process. We have defined 13 fusion operator rules based on global ontology. For example, a closed interval operator is a fusion operator whose definition is as follows:

Definition 6: Given a domain D and possible values on it $D' = \{v_1', v_2', \ldots v_n'\}$, the closed interval operator(CInt) satisfies:

$$CInt(D') = [v_i, v_j], \text{ if } \forall\ v_i' \in D', \text{ then } v_i' \in [v_i,\ v]$$

Example 1: If there exist three possible tuples: v_1 = (Wang da hong; age; 12), v_2 = (Wang da hong; age; 13), and v_3 = (Wang da hong; age; 15), then we will get CInt ($\{v_2, v_2, v_3\}$) = (Wang da hong; age; [12–15]).

In our Fusion rule selection, each rule will be limited to some condition that can be deduced by a rule character and a query that can be defined:

***Definition 7*:** Given query ontology Ω, a knowledge fusion query can be formally defined:

$$o \cdot \{(s_1, fr_1) = ?, \ldots, (s_n, fr_n) = ?\} \mid cnt, o \cdot \{(s_1, fr_1) = ?, \ldots,$$

where $(s_n, fr_n) = ?\}$ represents query objects, and cnt is a set of constraint conditions. O is a concept or instance in Ω, s_1 is a slot (attribute) of o, and fr_1 is a fusion rule. If fr_1 is omitted, the query will be changed into a general query in traditional information integration.

***Example 2*:** Given a query = Potato · (price, Avg), the knowledge fusion system should provide an average price of a price set of potatoes returned by information integration. If Avg is NULL, then the knowledge fusion system will return the potato price in a way similar to traditional information integration. Often a user can select a rule according his preference.

In query ontology Ω, we define a default rule for each slot of a concept, involving two slot types: meta-slot and composite-slot. A meta-slot is a slot that cannot be divided semantically while a composite-slot can be divided into many meta-slots. For example, slot IdentityNo of a concept person is a meta-slot, but Name, usually, is a composite-slot including a meta-slot first-name and a meta-slot last name. A fusion rule for meta-slot is always pre-defined according to the meta-slot definition, but a composite-slot usually needs a concatenate rule. In order to acquire a high quality answer, we need to extend the slots of a concept to filter out useless information. The slots also are called data quality slots including:

- **Authority (DQa)** The data quality authority is used to measure the probability of information correctness in information sources.
- **Timeliness (DQt)** Timeliness presents a means to estimate the goodness (or badness) of information in information sources in terms of time.
- **Completeness(DQc)** The degree to which all data relevant to an application domain have been recorded in an information source.

Therefore, given a concept and its slot set $\{a_1, a_2, \ldots, a_n\}$, the extensional slot set will be $\{a_1, a_2, \ldots, a_n, DQ_a, DQ_t, DQ_c\}$.

4.3 Knowledge Inconsistency Problem Analysis

In general, knowledge consistency means a judgment is in accord with both historical judgments and current facts. On the other hand, inconsistency means a contradiction between the historical judgments and current facts. From the aspect of ontology, consistency means that the logic relationships of the terminology are consistent while inconsistency means conflicts exist between some parts of the ontologies. For example, we define grain crops and cash crops as disjoint classes that do not have the same instances. If the class wheat belongs to both grain crops and cash crops, an inconsistency will occur.

In this paper, agricultural ontology consistency includes consistency between the ontology definition and the knowledge based on the ontology. This means that we cannot obtain conflicting knowledge from the knowledge base. Generally, when a knowledge base exists, conflict knowledge depends on the following conditions:

1) The consistency of concept defining. That is to say, the formal definition contains the same meaning as the informal one. Take the concept "dogs" as an example. If the formal definition of dogs is the same as that of the concept cats, inconsistency exists.
2) The consistency of concept extension. In terms of formal or non-formal concept definitions, conflict knowledge can exist through concept explanation (including reasoning). For example, cats can catch mice, but we cannot say that mice can catch cats.
3) The consistency of axioms. The axiom system will not produce conflict knowledge.

From the viewpoint of knowledge application, the knowledge base can guide users to make correct decisions and ensure that no confusing conclusions arise. In brief, consistency is an important criterion with which to evaluate an ontology-based knowledge base. Knowledge inconsistency will lead to unreliable service, which threatens knowledge correctness. This paper proposes a method with which to check ontology consistency.

Definition 8: Given knowledge base K, the knowledge inconsistency problem is a 3 triple KI = (K, Y, Q), which satisfies :

- Y = {y1, y2, …, yn} is a knowledge operation set.
- Q is a given knowledge query.

Definition 9: Given knowledge inconsistency problem KI = (K, Y, Q). If a knowledge conflict exists in K, it satisfies the following conditions:

- $\exists\, k, k_{11}, k_{22}, \ldots, k_{1j} \in K, y_{11}, y_{12}, \ldots, y_{1j} \in Y, \sum_{1}^{j} y_{1i}(k_{1i}) \models k \wedge k \rightarrow Q$. The symbol |= indicates "reason out" and → represents "can satisfy".

- $\exists k, k_{11}, k_{22}, \ldots, k_{1m} \in K, y_{21}, y_{22}, \ldots, y_{2j} \in Y, \sum_{1}^{j} y_{2l}(k_{2l}) \models \neg k \wedge \neg k \rightarrow Q$.

From the above definitions, we see that the knowledge base has inconsistency if there are two pieces of contradictory knowledge. It is very import to find a mechanism or method to check this knowledge base inconsistency (Xie, 2012).

5 Agrionto-Based Knowledge Fusion

5.1 Equivalent Entity Distinguishing

Equivalent entity distinguishing uses a clustering algorithm to classify the same entities into categories using identity slots (IS); that is to say, if IS(entiy1) = IS(entiy2), then entiy1is equivalent to entiy2, from the entity viewpoint (entiy1 ≈ entiy2). We also think that the two entities have different descriptions of an object. From the equivalent entity definition, we can conclude the following propositions: Proposition1: if E1 ≈ E2 ∧ E2 ≈ E3, then E1 ≈ E3; Proposition2: if E1 ≈ E2 ∧ E2 ≠ E3, then E1 ≠ E3; Proposition3: if E1 ≈ E2, then E2 ≈ E1

In order to determine whether two entities are equivalent, we need to analyze the identity slots' values:

- **Abbreviation.** An abbreviation is a shorter way to say something, for example, Massachusetts = Mass.

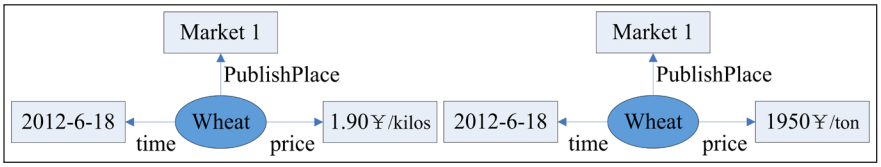

Figure 4: The extracted information fragments of two instances.

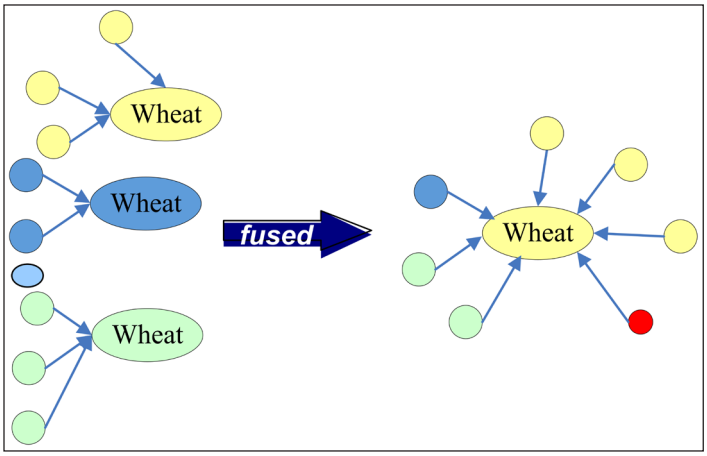

Figure 5: The instance fusion process.

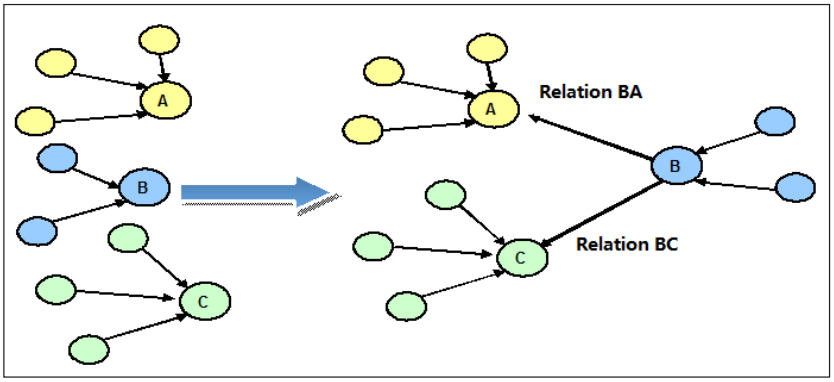

Figure 6: The concept fusion process.

- **Synonym.** Given two words that are synonyms, they represent the same entity or concept, for instance, corn and maize.
- **Prefix & Suffix.** An abbreviation using the first or last letter of each word, for example, IM = Instant Messaging.

If data in the identity slot are pre-processed and IS(entiy1) = IS(entiy2), then entiy1 ≈ entiy2.

5.2 Fusion Method

In our research, we define fusion rules at attribute granularity. Each fusion rule can be looked at as an aggregation function in the database, such as Min, Max, and Avg. In general, single data fusion rules and multi data fusion rules cannot be applied to an information set. Instead, we need to analyze the query and answer type and then define the necessary combination of fusion rules. Usually, however, a user needs to participate in rule selection to finish the knowledge fusion process. Generally, the attribute constraint determines the rule selection that is affected by the query. We divide knowledge fusion into attribute fusion, instance fusion, and concept fusion.

• Attribute fusion

Attribute fusion merges the different values at an attribute, for example (see **Figure 4**), "What price is the wheat at market 1?" The information fragments of two equivalent instances are extracted from information sources. In this case, the two values of the price are inconsistent so the last fused price will be "1.925¥/kilo" using the Avg rule. This is especially useful when the price value is an editing error.

• Instance fusion

Instance fusion merges equivalent instances that have different descriptions of the same object (see **Figure 5**). Because most information sources describe a part of an object, the fused result is the union of the equivalent instances based on the attribute fusion.

• Concept fusion

Concept fusion takes into account the correlations among equivalent instances by combining different instances that are divided into different sets of equivalent instances by the cluster algorithm (see **Figure 6**).

6 Conclusion

Data have become strategic resources as important as natural resources and human resources with an implied great value and have caught the attention of both the scientific and business communities. With the recent rapid growth in the amount of data, existing data processing technology has great difficulty in meeting the large demand placed on it, and the data are very difficult to mine. In this paper, we propose a generic agricultural knowledge fusion method to fuse information from diverse information sources, such that a more comprehensive basis can be obtained for data analysis and knowledge discovery for agricultural big data. In recent years, information systems integration or business integration have received much attention (Xie & Wang, 2010; Xie, 2012). Now we must pay attention to the integration of agricultural data in the area of big data because once the data are gathered and stored in an integrated database, they will have new value. This paper describes how to make full use of agriculturual information from the aspect of knowledge fusion technology, which will accelerate the correct use of agricultural knowledge and give a knowledge basis for big data mining. In the future, we will further study data consistency, ontology-based rules, and fusion algorithms and conduct more application tests under the open agricultural big data environment.

7 Acknowledgments

This work was supported by key projects of the Ministry of Agriculture on the cultivation of new varieties of genetically modified organisms (No. 2014ZX0801101B) and CAAS Agricultural Science and Technology Innovation Program.

8 References

Feigenbaum, E. & McCorduck, P. (1983) The fifth generation: artificial intelligence and Japan's computer challenge to the world. Reading: Addison-Wesley Publishing Company.

Hu, X., Hu, J., Sekhari, A., Peng, Y.H., & Cao, Zh.M. (2011) A Fuzzy Knowledge Fusion Framework for Terms Conflict Resolution in Concurrent Engineering. *Concurrent Engineering: R&A (CERA) 19*(1), pp 71–84.

Hunter, A. & Williams, M. (2010) Qualitative Evidence Aggregation using Argumentation. *COMMA 2010*, pp 287–298.

Lenat, D.B. & Guha, R.V. (1990) *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Addison-Wesley Publishing Company, Inc.: CA.

Li, X., Dong, X. L., Lyons, K. B., Meng, W., & Srivastava, D. (2013) Truth finding on the deep web: Is the problem solved. *PVLDB 6*(2).

Liu, J., Xu, W., & Jiang, H. (2014) Research on Dynamic Ontology Construction Method for Knowledge Fusion in Group Corporation. *Advances in Intelligent Systems and Computing 278*, pp 289–298.

Motro, A. & Anokhin, P. (2004) Utility-based Resolution of Data Inconsistencies. In *Proc. International Workshop on Information Quality in Information Systems 2004*, Paris, France, pp 35–43.

Preece, A., Hui, K., Gray, W., & Marti, P. (2001) Designing for Scalability in a Fusion System. *Knowledge Based Systems, 14*(3-4), pp 173–179.

Stegmaier, F., Bürger, T., Döller, M., & Kosch, H. (2010) Knowledge Based Multimodal Result Fusion for Distributed and Heterogeneous Multimedia Environments: Concept and Ideas. *Adaptive Multimedia Retrieval*, pp 61–73.

Wang, C.Y., Hu, B., & Li, P. (2009) Empirical Study of Knowledge Fusion Process within Chinese High-Tech Industry Clusters Based on Information Fusion Method. *JIKM 8*(4), pp 353–361.

Xie, N. (2012) Research on the Inconsistency Checking in Agricultural Knowledge Base. In *Proc. CCTA (1),* pp 290–296.

Xie, N. & Wang, W. (2010) Research on 3G Technologies-Based Agricultural Information Resource Integration and Service. *CCTA 2009*, pp 114–120.

Xie, N., Wang, W., Yang, X., & Jiang, L. (2012) Rule-Based Agricultural Knowledge Fusion in Web Information Integration. *SENSOR LETTERS 10*, pp 1–4.

How to cite this article: Xie, N, Wang, W, Ma, B, Zhang, X, Sun, W and Guo, F 2015 Research on an Agricultural Knowledge Fusion Method for Big Data. *Data Science Journal,* 14: 7, pp. 1–9, DOI: http://dx.doi.org/10.5334/dsj-2015-007

Published: 22 May 2015