Study on Semantic Heterogeneity Elimination of Agricultural Product Price Information in Multi-source Network

Jing Zhang, Guo-Min Zhou, Jian Wang, Jie Zhang, Fangli Xie

Study on Semantic Heterogeneity Elimination of Agricultural Product Price Information in Multi-Source Network

Jing Zhang[1,a], Guomin Zhou[1,b], Jian Wang[1,c], Jie Zhang[1,d], Fangli Xie[1,e]

[1]Agricultural Information Institute of CAAS, Beijing 100081, China
[a]zuer0101@163.com, [b]zhouguomin@caas.cn, [c]wangjian01@caas.cn, [d]tulaapple@163.com, [e]fangli_x@163.com

Abstract. With development of web information technologies, internet applications have come into burst growth, which brings different kinds of standard, database and description methods. This condition causes the "information islands" that make a lot of trouble in data sharing. The root cause of these problems is information source heterogeneity, which consists of four levels, and among them the semantic heterogeneity is the most difficult issue. The reasons of semantic heterogeneity existing are pointed out in this article and the research status of heterogeneity elimination is introduced. At last the solutions of agricultural product price information in multi-source network are summarized and the research emphasis in future is prospected.

Keywords: network information technology, semantic heterogeneity, agricultural product price information

1 Introduction

Network information technology has penetrated into human production and other aspects of our life, and so as agricultural sector as a basis for human survival and development. At present, agriculture website is developing rapidly, but the content is very complex. It's a great difficult for the peasantry and agriculturists to get the information conveniently and effectively. Therefore, the development of agriculture search engine has become their urgent needs. Agricultural product price information is the most important part of the rural information service system. However, the existence of heterogeneity led to disconnect and asymmetries appearance in prices of agriculture search engine [1]. In order to make agriculture search engine more accurate and effective, it is imperatively to eliminate semantic heterogeneity.

This article explained the reason of semantic heterogeneity, discussed the treating mode of semantic heterogeneity which have three levels (schematic heterogeneous, context heterogeneous, individual heterogeneous), and then discussed semantic heterogeneity of agricultural product price information in multi-source network, at last summarized the research progress of the price information semantic heterogeneity's eliminating.

2 Generation and Research Progress of Semantic Heterogeneity

Information sources heterogeneity can be divided into four levels: system, syntax, structural and semantic, during these the most difficult to eliminate is semantic heterogeneity [2]. To some extent, systems, syntax and structure heterogeneity has been resolved through the traditional information integration. Therefore, the identification and elimination of semantic heterogeneity has become a difficulty and then the research hotspot.

The generation of semantic heterogeneity has been caused by different types of semantic conflict [3]. Park and Ram (2004) proposed classification of semantic heterogeneity conflicts: scheme-lever conflict and data-lever conflict [4]. Scheme-lever conflict is the conflicts caused by the logical structure that used by the same concepts in different information sources. Data-lever conflict is reflected in same aspects, such as naming identifiers, precision, representation, reliability, etc., which is due to the different perception of the same concept [5]. Zhou Jianfang et al (2008), from Huazhong University of Science and Technology, proposed semantic heterogeneity between the data sources that are not independent have interrelationship with each other. Semantic heterogeneity is manifested as schematic heterogeneity, context heterogeneity, and individual heterogeneity [6, 11]. In order to enable users to obtain efficient and accurate data, all of the three levels of conflicts need eliminating.

(1) Schematic heterogeneous elimination

Schematic heterogeneity is characterized by differences in logical structures and/or inconsistencies in metadata (i.e., schemas) of the same application domain [4]. At present, schematic heterogeneous data sources have been widely studied. The main eliminating method is to use the existing metadata and a small amount of users' intervention to achieve the automatic schema mapping. Mediator Environment for Multiple Information Sources (MOMIS) [3], Developed by Bergamaschi et al in 1998, is one of the typical representatives.

(2) Context heterogeneous elimination

Context heterogeneity is the different data sources that have the same semantic information (including entities and attributes). Data interpretation deals with the existence of heterogeneous contexts, whereby each source of information and potential receiver of that information may operate with a different context which leads to large-scale semantic heterogeneity [7]. At present, among all the methods of context heterogeneous elimination based on the context mediation, the typical representative is COIN project team in the MIT [8]. The project solves four kinds of context heterogeneities.

(3) Individual heterogeneous elimination

Individual heterogeneity mainly refers to individual identification in different data sources [6]. After schematic and context heterogeneity elimination had been implemented, the same individual in different data sources would still use different forms of representation and description. Thus, the traditional method to exactly match the property value comparison can't eliminate individual heterogeneity. Active Atlas system is a typical representative, which has been developed by Tejada et al. on University of Southern California [9-10]. However, the solutions in the system can only applied to string matching without other types of data matching. What is more, this solution does not apply to Chinese.

Zhou Jianfang proposed a general framework for semantic information integration, which is used to solve the three levels semantic heterogeneity [11]. In the framework,

query engine, context mediation and results processing components are the core processing components. They work together in order to solve schematic, context and individual heterogeneity. However, each core processing component needs to reference the relevant metadata. The acquisition of this metadata requires the domain experts and system designers to work together.

3 Semantic heterogeneity of the agricultural product price information

3.1 Influence and relative research status

Nowadays, with the rapid development of Chinese agricultural internet, there have been more than 30,000 agricultural website. Many problems would occur if researchers just simply collect price information in multi-source network without eliminating its semantic heterogeneity:

(1) Information structure is inconsistent. The different sources release price information in different formats. Some only publish average price and specifications while some else provide wholesale market, quality and other related content as well. Therefore, agricultural search cannot be used if we do not solve this kind of semantic heterogeneity.

(2) With information's complexity and redundancy, it is difficult to meet all the needs of users. Due to geographical and cultural differences, the same kind of agricultural commodities have different titles, such as 'huanggua', also known as 'qinggua' in Guangzhou. The release of this kind of price information will bring about unnecessarily redundant information if researchers do not eliminate such a semantic heterogeneity.

(3) Actually practical value is reduced. The agricultural product price information without eliminating its semantic heterogeneity lacks unified comparison and overview. The price information cannot either help majority famers to understand the market or be able to give a supporting guidance to agricultural-related policy.

Through above analysis, the identification and elimination technology of semantic heterogeneity of agricultural product price information in multi-source network is of vital importance to improve agricultural information services' quality and provide information that is more consistent with the needs of farmers and related agricultural workers. Related research about the technology to eliminate the semantic heterogeneity of agricultural product price information in the multi-source network is limited currently; it is University of Science and Technology of China that mainly conducted some research. According to the characteristics of agricultural product price information given by the Internet, Lei Ying (2010) uses a seven-tuple (price, unit, name of agricultural products, source, products trading market, province, transaction date) to represent the properties of agricultural product price information, so as to solve semantic heterogeneity [12]. But not all sources are able to meet the seven-tuple. As many sources just provide products trading market, it is necessary to get specific province information. He Huang (2010) treated the market name of agricultural product price information by dividing them into two cases, and establish a

uniform location data table to solve spatial information heterogeneous through a combination of automatic annotation of spatial properties using administrative divisions indexing library and extracting location information for labeling using the general search engines [13]. What's more, he eliminating the data reliability conflicts through the establishment of abnormal data detection system and redundant data processing methods based on the semantics which could develop the quality and availability of data. With further research, the elimination of semantic heterogeneity in order to clean a lot of redundant data comes true with the system's combination with agricultural classification ontology, geographic name resolution system, unit conversion rules, and several other modules [14].

The current study of the semantic heterogeneity of agricultural product price information in multi-source network does not have a comprehensive summary and classification, what the elimination technology could solve is only parts of semantic phenomenon. What the problem that needs to be focused on next step is that to explicit the classification of semantic heterogeneity of agricultural product price information through a lot of research and the participation of experts so as to gradually improve the semantic heterogeneity elimination.

3.2 Study on the semantic heterogeneity in the agricultural product price information

3.2.1 Classification of semantic heterogeneity of agricultural product price information

In order to get the semantic heterogeneity of agricultural product price information existing in the current agricultural sources, the survey from a large of agricultural network have many specification and standard information about the price information. The multi-source network mainly includes China Agricultural Information Network (http://www.agri.gov.cn) which provides regional agricultural websites of wholesale market, and provincial and municipal government networks which provides agricultural product price information. Through sampling survey and combination with the basic principles of semantic heterogeneity, the semantic heterogeneity of the agricultural product price information in multi-source network has been divided into three categories: semantic heterogeneity based on schematic, context data, and abnormal individual data.

(1) Semantic heterogeneity based on schematic

① The semantic homogeneous conflict is a same concept in different sources has different set of attributes. For instance, with regard to price of a certain agricultural product, some sources only provide basic information of agricultural products about varieties, prices, units and wholesale. However, other sources include origin, highest price, lowest price, specifications and other information.

② The aggregate conflict is some sources use a concept, which the others of the same kind use a set of concepts. For instance, in the description of a kind of agricultural products, product name and specifications would be compressed into a single attribute in some sources.

③ Ancestor or descendant conflict is that the same concept is a superclass in one source and a subclass in another source. For instance, ‘suan’ is a subclass in some sources while it is a superclass for ‘qingsuan’, ‘dasuan’ in other source webs.

④Classification conflict is that a same concept in different sources belongs to different superclass. For instance, some sources would be divided into two categories of fresh meat and eggs, and in the other sources meat and egg would be placed in the one category.

⑤Naming conflict is that a conflict resulting from a same concept subjectively determined by designers of different fields or from different sources. This is more common in semantic heterogeneity, which in the price information is often highlighted in the two properties: the name of agricultural products and the wholesale market.

(2) Semantic heterogeneity based on context data

① Data value conflict is a same data in different sources to explain inconsistencies because of the different data representation, scales or coding. In different sources, the same price data have a different meaning for wholesale prices or retail prices.

② Data presentation conflict is a same concept with different data types or different formats. For instance, representation of date may be conflict on different systems, such as ‘yyyy-MM-dd’ or ‘MM/dd/yyyy’.

③ Data unit field conflict is to use of different units of measurement resulting from the conflict. The conflict is particularly prominent in the agricultural product price information. For instance, some sources use ‘kg’ as the unit of measure and some use ‘jin’.

④ Data Accuracy conflict is the concept resulting from a same concept described in different precisions or the different domain and range. For instance, price information from different sources has different data accuracy, such as ‘yuan’ or ‘fen’.

(3) Semantic heterogeneity based on abnormal individual data

① Individual description conflict is that different systems use different methods to describe the same individuality. As price information of agricultural product consists of many kinds of characterized properties, different describing methods for them are involved and so conflicts are inevitable.

② Data reliability conflict is that different reliable levels are assigned to a certain conception. This comes from possible incorrect data filling of information collection sources and it is necessary to correct the data without semantic heterogeneity.

The sorting of semantic heterogeneity on agricultural product price in multi-source network has been listed above. The conflicts need cleaning in sequence by three levels. At first, semantic heterogeneity based on schematic is the most important and influencing most among all the conflicts. And its elimination is the prerequisite of solving semantic heterogeneity of context data and abnormal individual data. Secondly, the context data heterogeneity brings incorrect calculations and comparison, which make it necessary to be dealt with based on the cleaning of schematic heterogeneity before solving abnormal individual data heterogeneity. In the third level of priority, schematic and context data heterogeneity has been solved while individual abnormal data still remains in information which will cause semantic heterogeneity. Therefore to obtain the final valuable price information of agricultural products, a

certain individuality data from different sources need to be intelligent calculated including combining and duplicates eliminating.

3.2.2 Expectations for sematic heterogeneity research of agricultural product price information

(1) Improve the sematic heterogeneity correcting system for agricultural product price information

Remaining problems will be resolved in sequence by levels according to sorting results of semantic heterogeneity research.

① Semantic heterogeneity based on schematic will be resolved through completing domain ontology and reducing naming conflicts of product types and markets.

② Context mediation technology will be involved to remove semantic heterogeneity based on context data and accurate, unified and precise data is expectable.

③ Individual conflicting information detecting and removing system will be actualized based on data conflict detecting and duplicates eliminating.

④ System efficiency will be considered in completing the whole system to achieve its fast running.

(2) Achieving personal service

The building up of semantic heterogeneity eliminating system for agricultural product price is connected with improving price showing system of agricultural product and different requests of users for data access and query are expectable. Personal service is important along with common usage. Therefore it comes into a promising research filed in building personalized data space and pushing characterized information.

4 Conclusions

Based on semantic heterogeneity research and analysis of agricultural product price information, the semantic heterogeneity is sorted into three levels as schematic, context data and abnormal individual data. Accordingly it is proposed that the heterogeneity will be eliminated in sequence of the three priorities. As a result, it is necessary to build up a complete semantic heterogeneity eliminating system to obtain the ideal information with high quality for users. Furthermore, efficient combination of the heterogeneity eliminating system and professional agricultural search engine will dramatically improve the information provided for peasantry and agriculturists.

Acknowledgment

The research is funded by the National High Technology Research and Development Program of China (2013AA102405) and the basic scientific research special fund of nonprofit research institutions at the central level (2012-J-07).

References

1. YU Jia-Xin. The Study on the Outline Model of the Agriculture Product Price Information Service [J]. Agriculture Network Information, 2004, (02)
2. Sheth A P. Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics [M]. Boston: Kluwer Academic Publishers, 1999.
3. S.Bergamashi, S.Castano, S.De Capitani di Vimercati et al. An intelligent approach to information integration [C]. Formal Ontology in Information Systems. Italy: IOS Press, 1998.
4. J.Park, S.Ram. Information Systems Interoperability: What Lies Beneath? [J]. ACM Trans on Information Systems, 2004, 22(4):595-632.
5. Li Yanxia. Research on Semantic Heterogeneity Elimination for information integration [D]. Northwest Normal University, 2009
6. ZHOU Jian-fang, XU Hai-yin, LU Zheng-ding. Research of semantic heterogeneity in information integration [J]. Application Research of Computers, 2008, (08):2349-2351
7. F.AYKUT. Information Integration Using Contextual Knowledge and Ontology Merging [D]. Cambridge: Massachusetts Institute of Technology, 2003.
8. Hongwei ZHU, S.E.MADNICK. Context Interchange as a Scalable Solution to Interoperating Amongst Heterogeneous Dynamic Services [C]. Proc of the 3rdWorkshop on eBusiness. Washington DC: [s.n.], 2004: 150-161.
9. S.Tejada, C.A.Knoblock, S.Minton. Learning Object Identification Rules for Information Integration [C]. Special Issue on Data Extraction, Cleaning, and Reconciliation, Information Systems Journal, 26(8), 2001.
10. S.Tejada, C.A.Knoblock, S.Minton. Learning Domain-Independent String Transformation Weights for High Accuracy Object Identification [C]. Proc of the 8th ACM SIGKDD International Corference on Knowledge Discovering and DataMining. New York: ACM Press, 2002: 350-359.
11. ZHOU Jian-fang. Research on Context Mediation Based Semantic Information Integration Method [D]. Huazhong University of Science & Technology, 2009
12. Lei Ying. Abnormal data detection in agriculture search engine [D]. University of Science and Technology of China, 2010
13. He Huang. Complex Adaptive Agriculture Vertical Search Model and its Implementation [D]. University of Science and Technology of China, 2010
14. Yimin Hu. Semantic Research and Implementation on Agricultural Vertical Search Engine [D]. University of Science and Technology of China, 2012