Copyright © 2012 American Scientific Publishers
All rights reserved
Printed in the United States of America

SENSOR LETTERS
Vol. 10, 1–4, 2012

Rule-Based Agricultural Knowledge Fusion in Web Information Integration

Xie Nengfu*, Wang Wensheng, Yang Xiaorong, and Jiang Lihua

Agricultural Information Institute, The Chinese Academy of Agricultural Sciences, Beijing, P. R. China;
Key Laboratory of Digital Agricultural Early-Warning Technology, Ministry of Agriculture, P. R. China

(Received: 30 June 2011. Accepted: 20 September 2011)

In traditional Web information integration systems, answers are provided as a large relevant information entity set stratifying a user' question, which make a user to browsing the set for the final answer that may not exits. Deeper information integration, called knowledge fusion (KF), which provides a more integrated answer, involves not only delivering the answer information available via the links to user, but also analyzing, and merging the information results coming from agricultural information sources by solving the result consistencies, removing duplicates, etc based on agricultural domain ontology. In the paper, we give a detail about the knowledge fusion Method, and a KF-based information access interface. Many experiments prove that the method is effective.

Keywords:

RESEARCH ARTICLE

1. INTRODUCTION

The Web was designed as an information space by Tim Berners-Lee, with the goal not only that it should be available for human reading, but also that machine would be able to participate in and help users to communicate with each other. More information on the Web needs to be in a form that machines can "understand" rather than simply display.[1, 3] So in order to explain the knowledge fusion research, we will discuss the knowledge fusion based on XML sources,[9, 10] Generally, in traditional Web information integration systems, answers are provided as a large relevant information set stratifying a user' question, which make a user to browsing the set for the final answer that may not exits because information content is not deeply analyzed, and processed, especially in consistency with no semantic understanding. In the paper, we look knowledge fusion as an extension of information integration, which aims to fuse the information answers from different sources to an integrated answer by analyzing, processing information incompleteness, consistency and redundancy. The paper will focus on information content conflict, information extension conflict, which is solved with fusion rules and data quality according to knowledge fusion model.[10] Generally, information extension conflict resolving methods can be sum up as the followings:

- Sort Method. The method will provide an information set related with user query. In the set, elements will sorted by similarity value between query and elements.
- Random Method. The method will select an answer to user from the query answers randomly.
- Preference Method. It will give an information answer for a user query with high similarity value from the result of Sort Method.
- Fusion Method. An integrated information entity will be merged according to fusion rules from a information set of information integration for a user query.

Currently, most of search engines and information integration systems adopt the first method. We are more interested in knowledge fusion to provide an fusied answer for agricultural users. In database area, some related research were open, such as Fuisonplex,[2] FraQL.[6] The paper will discuss the knowledge fusion method in detail on the basis of knowledge fusion model.[10]

The remainder of the paper is organized as follows: the Section 2 presents related work. In Section 3 we discuss the answer fusion rules in more detail. A description of answer fusion method is proposed in Section 4. Section 5 concludes the paper and discusses the future work.

*Corresponding author; E-mail: nfxie@caas.net.cn

2. RELATED WORK

What a current information retrieval system or search engine can do is just to retrieve documents ranked by a certain algorithm of computing a similarity degree between the question (query) and each answer. That is to say, given some keywords it will only return the relevant documents that contain the keywords.[12] A user needs to rummage about the result until the correct document is found (if he/she is lucky). In many cases, a user may have to accept that the correct documents may not be contained in the result. However, what a user really wants is often a single and precise answer to a question.[10] For instance, given the question "what's the wheat's cultivated technology?," the user may only want to know the wheat's cultivated technology, but in fact, in order to find the exact answer by himself, he has to read through lots of documents returned by the search engine, which contains the key words in the question.

Now, many research works are paying attention to give a concise answer including fusion method and Preference Method. Fusionplex is a system for integrating multiple heterogeneous and autonomous information sources that uses data fusion to resolve factual inconsistencies among the individual sources. To accomplish this, the system relies on source features, which are meta-data on the merits of each information source.[10] Hunter etc. advocates a knowledge-based approach to merging semi-structured information. He uses fusion rules to manage the semi-structured information that is input for merging. Fusion rules are a form of scripting language that defines how structured reports should be merged. The work assumes that structured news reports do not require natural language processing and uses fusion rules to handle the inconsistencies and uncertainty of news reports.[4] Question answering has recently received attention from the communities of information retrieval, information extraction, machine learning, and natural language processing.[5] The goal of a question answering system is to return a concise answer to a question rather than a list of documents as most information retrieval systems currently do. The Text Retrieval Conference (TREC) Series has greatly motivated Question answering research in the recent years. In TREC8, RECT9, TRECT10, a question answering system is required to return 5 ranked answers evaluated by MRR metric for each test question.[7,8] In TRECT11, a question answering system is required to return only one extract answer for each test question, text strings consisting of a complete answer and nothing else.[12] In the Web environment, information is inconsistent, uncertain, incomplete, imperfect, and dynamic. If two or more different answers exist for the same question, how to select the final answer from the two candidate answers is a big challenge? Often the number of the candidate answers is very large, and there exist inconsistencies among them due to autonomous information sources.

A general model is implemented with fusion rule in our knowledge fusion method in which we also take data quality into account, more detail in Section 3.

3. FUSION RULE

3.1. Fusion Rule Define

Each fusion rules can be looked as an aggregation function in database, such as Min, Max and Avg.[9,10] We divide fusion rule into two types: single data fusion rule and multi data fusion rule.

DEFINITION 1. Single data fusion rule (SFR) is a kind of aggregation functions like:

$$f\colon D_1 \times D_2 \times \cdots \times D_n \to D \tag{1}$$

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

DEFINITION 2. Multi data fusion rule (MFR) is a kind of aggregation functions like:

$$f\colon D_1 \times D_2 \times \cdots \times D_n \to 2^D \tag{2}$$

Give $\mathrm{v}_i \in D$ $(i = 1, 2, \ldots, n)$, $f(v_1, v_2, \ldots, v_n) = D'$, $v_i \in D$, $D' \subseteq D$. MFR includes CInt (Interval Rule), Or, and And.

3.2. Fusion Rule Analysis

Generally, Single data fusion rule and Multi data fusion rule can not be applied into an information set, and we need analyze the query and answer type, and then define the combination of fusion rules, but usually, a use participates in rules selecting to finish the knowledge fusion process. We have defined 13 fusion operator rules based on the global ontology. For an example, a closed interval operator is fusion operator whose definition is as following:

DEFINITION 3. Given a domain D and possible values on it $D' = \{v'_1, v'_2, \ldots v'_n\}$, closed interval operator (CInt) satisfy:

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

EXAMPLE 1. If there exists 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 will get CInt $(\{v_2, v_2, v_3\}) =$ (Wang da Hong; age;[12–15]).

In our Fusion rule selecting, each rule will be limited to some condition that can be deduce by a rule characters and a query which can be defined:

DEFINITION 4. Given a query ontology Ω, a knowledge fusion query can be formally defined:

$$o\cdot\{(\mathrm{s}_1, \mathrm{fr}_1) =?, \ldots, (\mathrm{s}_n, \mathrm{fr}_n) =?\}|\mathrm{cnt},$$

$$o\cdot\{(\mathrm{s}_1, \mathrm{fr}_1) =?, \ldots, (\mathrm{s}_n, \mathrm{fr}_n) =?\}$$

RESEARCH ARTICLE

represent query object, and cnt is a set of constraint condition. O is a concept or instance in Ω, si is a slot (attribute) of o, and fri is a fusion rule. If fri is omitted, the query will be changed into a general query in traditional information integration.

EXAMPLE 2. A query = Potato·(price, Avg), the knowledge fusion system should provide an average price of a price set of the Potato returned by information integration. If the Avg is NULL, then knowledge fusion system will return the potato price like most of traditional information integration. Often, a user can select a rule according his preference.

In a query ontology Ω, we define a default rule for each slot of a concept, involving two slot type: a meta-slot and composite-slot. A meta-slot is a slot can not be divided semantically, and composite-slot can be divided into many meta-slots. For example, a slot Identity No of a concept person is a meta-slot, but Name is a compos-ite-slot including a meta-slot first-name and a meta-slot last name usually. A fusion rule for meta-slot is always predefined according to meta-slot definition, but a composite-slot usually need concatenate rule. In order to acquire a high quality answer, we need extend the slots of a concept for filtering some unuseful information. The slots also are called data quality sot 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 term of time.
- *Completeness(DQc)* The degree to which all data relevant to an application domain has been recorded in an information sources.

So given a concept and its slot set $\{a_1, a_2,\ldots,a_n\}$, the extensional slot set will be $\{a_1, a_2,\ldots,a_n, \mathrm{DQ}_a, \mathrm{DQ}_t, \mathrm{DQ}_c\}$.

4. FUSION RULES-BASED KNOWLEDGE FUSION

In sector 3, we have discussed fusion rules, but the rule use only is applied to the same entities, equivalent entities (EE) called in the paper. The knowledge fusion is proposed because of the following two factors:

- Equivalent entities exist in different information source, and may contain conflict data.
- An information source is an incomplete information carrier, that is to say, any information source can not contain information for any users' query.

So we need solve data confliction between equivalent entities from different information sources, and aim to provide uses a more complete answer in which data may coming from different entities.

4.1. Equivalent Entity Distinguishing

Equivalent entity distinguishing use clustering algorithm to classify the same entities into a category by identity slots (IS), that is to say, if IS(entiy1) = IS(entiy2), then entiy1 is equivalent to entiy2 in view of entity (entiy1 $\approx$ entiy2). We also think that the two entities have different description about an object. From the equivalent entity definition, we can conclude the following propositions: Proposition 1: if $E1 \approx E2 \bigwedge E2 \approx E3$, then $E1 \approx E3$ Proposition 2: if $E1 \approx E2 \bigwedge E2 \neq E3$, then $E1 \neq E3$ Proposition 3: if $E1 \approx E2$, then $E2 \approx E1$.

In order to determine two entities is equivalent, we need analyze the identity slots' value:

- *Abbreviation.* An abbreviation is a shorter way to say something, for example, Massachusetts = Mass.
- *Synonym.* Given two words that are synonyms, they represent the same entity or concept, for instance,.
- *Prefix and Suffix.* It is one kind of abbreviation is when you say the first or last letter of each word, for example, IM = Instant Messaging.

If data in identity slot have pre-processed and IS (entiy1) = IS (entiy2), then entiy1 $\approx$ entiy2.

4.2. Fusion Method

The aim of knowledge fusion is to select fusion rules for a slot in term of its constraints and user preference. In our knowledge fusion system, we define default fusion rules for each slot constraint. For instance, if data type of a slot is numeric, then we may use Max, Min, CInt and etc. rules applied to a slot data type, and think another complex fusion rule under considering the distributing of all values in a slot. Given the values of a slot $V = \{280, 230, 231, 231, 234, 235, 450\}$. Obviously, value 450 is exclusive value, and if Max is applied in V, the fused answer may be not correct. In the paper, we suppose the values obey normal distribution:

$$\varphi_{a,\sigma} = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\alpha)^2}/2\sigma^2$$

where α is mathematical expectation, σ is standard deviation. We use median-number rule to get α_0 as experience value of α. Because σ is unknown, statistical control $t = \sqrt{n-1}(x-\alpha_0)/s$ obey $t(n-1)$ distribution, so under confidence level, confidence interval of α is $[\overline{X} - (s/\sqrt{n-1})t_{a/2}(n-1),\ \overline{X} + (s/\sqrt{n-1})t_{a/2}(n-1)]$. If confidence is 0.95 and α_0 is in confidence interval, we determine that α_0 is estimated value of α, so we use CInt rule to get an fused answer $[\alpha_{0-s}, \alpha_{0+s}]$, or we delete the data from the two end of V and compute again until the fused is available.

The above rule is only default rule for numeric data type of a slot, but in fact, the user like to select rues for his interest. In our fusion rule section, we define default fusion

RESEARCH ARTICLE

rule for each slot in a concept. For example, the data type of a slot is a string type, which is very different in rule selection compared with number data type. Generally, we use the following methods to obtain the fusion answers.

• *Data quality-based fusion.* The data quality attributes are dynamically added to the answer description.[9, 10] Each data quality attribute is a float value between 0 and 1. The overall DQ of the answer can be computed by the following formula.

$$dq = \sum w_i DQ_i$$

Where w_i is the weight of the ith data quality attribute, $\Sigma w_i = 1$. In the method, we will choose the final answer whose d_q is the highest. Note that these data quality values are also frequently changed.

• *Content rule-based fusion.* We have defined 11 rules to solve the answer inconsistencies, such as Min, Max, Majr, and etc. Generally, a rule should be chosen by the users since we believe the user has the final right to determine if the answer is right. The general expression is

(a) for attribute a Content_Rule fusion_rueles

(b) Content_Rule fusion_rueles is the rule chosen by the user.

• *Mixed method-based fusion.* The mixed method-based fusion considers the fusion process from data quality and content rule-based fusion.

4.3. Experimental Result and Evaluation

Our knowledge fusion method has been used in practical system, such as in QA system,[11] and multi-sources Information fusion system.[9] In QA system, the experiment result indicates that the fusion method enhances not only the user satisfaction, but also the accuracy. We also used it in agricultural domain for agricultural product price information service. If we do not use the answer fusion, the accuracy will be reduced to 31%, which only tests factoid questions, such as "How much are the potato this month?," "Which wholesale market the price of potato is most cheap?" In Information fusion stem, the result proves that the answer fusion is very effective to enhance answer correctness and user satisfaction.

5. CONCLUSION

In the paper, we have proposed a generic knowledge fusion method for knowledge fusion model in Ref. [10]. It focuses on content fusion based on fusion rules, which combines many answers from different information source including search engines to form a single answer. In our experimental test, the result proves that the answer fusion is very effective to enhance answer correctness and user satisfaction, especially in agricultural product price information service system. In the future, we will make further study on the answer consistency issue, and give more tests in the open agricultural knowledge domain.

Acknowledgments: This work is supported by supported by the National Science and Technology Major Project of the Ministry of Science and Technology of China (Grant No. 2009ZX03001-019-01), Special fund project for Basic Science Research Business Fee, AIIS and The National Science.

References and Notes

1. A. Y. Halevy, Z. G. Ives, P. Mork, and I. Tatarinov, Piazza: Data manage¬ment infrastructure for semantic web applications, *Proceedings of the Twelfth International World Wide Web Conference*, **(2003)**, pp. 556–567.
2. Anokhin and A. Motro, Fusionplex: Resolution of data inconsistencies in the integration of heterogeneous information sources, technical report ISE-TR-03–06, Information and Software Engineering Department George Mason University, Fairfax, Virginia, **(2003)**.
3. J. Bosak and T. Bray, XML and the Second-Generation Web, Scientific American, **(1999)**, pp. 89–93.
4. S. Dumais et al., Web question answering: Is more always better? SIGIR **(2002)**, pp. 291–298.
5. A. Hunter and W. Liu, Measuring the quality of uncertain information using possibilistic logic, *Proceedings of ECSQARU'05*, LNCS, Springer, (in press) **(2005)**.
6. K.-U. Sattler, S. Conrad, and G. Saake, Adding Conflict Resolution Features to a Query Lan¬guage for Database Federations, *Proc. 3nd Int. Workshop on Engineering Federated In¬formation Systems*, Dublin, Ireland **(2000)**.
7. E. M. Voorhees and D. K. Harman, Overview of the eighth text retrieval conference (TREC-8), *Proceedings of the Eighth Text Retrieval Conference (TREC-8)* **(2000)**.
8. E. M. Voorhees, Overview of the TREC-9 question answering track, *In The Ninth Text Retrieval Conference (TREC 9)* **(2000)**, pp. 71–80.
9. N. F. Xie, Knowledge Fusion and Synchronization Methods Based on Semantic Web Tech¬nologies, Ph. Dissertation of Institute of Computing Technology, Chinese Academy of Science **(2005)**.
10. N. F. Xie and C. A Cao, Knowledge fusion model for web information, *To Appear in Proceed¬ings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)* **(2005)**.
11. N. F. Xie and W. Y. Liu, An answer fusion model for web-based question answer¬ing, *Accepted as a Research Paper by 1st International Conference on Semantics, Knowl¬edge and Grid*, Beijing, November **(2005)**.
12. D. Zhang, Web based question answering with aggregation strategy, *Proceedings of the 6th Asia Pacific Web Conference (APWEB2004)*, Hangzhou, China, April **(2004)**.

RESEARCH ARTICLE