IOP Conf. Series: Materials Science and Engineering **231** (2017) 012058 doi:10.1088/1757-899X/231/1/012058

HWDA: A coherence recognition and resolution algorithm for hybrid web data aggregation

Guo Shuhang, Wang Jian, Wang Tong

Central University of Finance and Economics & School of Information, China

Abstract. Aiming at the object confliction recognition and resolution problem for hybrid distributed data stream aggregation, a distributed data stream object coherence solution technology is proposed. Firstly, the framework was defined for the object coherence conflict recognition and resolution, named HWDA. Secondly, an object coherence recognition technology was proposed based on formal language description logic and hierarchical dependency relationship between logic rules. Thirdly, a conflict traversal recognition algorithm was proposed based on the defined dependency graph. Next, the conflict resolution technology was prompted based on resolution pattern matching including the definition of the three types of conflict, conflict resolution matching pattern and arbitration resolution method. At last, the experiment use two kinds of web test data sets to validate the effect of application utilizing the conflict recognition and resolution technology of HWDA.

1 Introduction

The current distributed web data stream aggregation has become one key problem in internet information integration and the WEB3.0 future trend [1]. In terms of the conflict recognition and resolution of the hybrid web data object coherence, some researcher had already proposed that the semantic web technology will take a core role [2]. In the process of the semantic data analysis, the ontology property priority and hierarchical dependency of semantic conflict will be required to consider. The formal language reasoning rule for identification of the conflict had been mentioned in some research, which was used to contribute higher recall ratio and precision ratio, at last prompt the evolution of the conflict recognition and resolution algorithm [3]. Another, the conflict recognition ability should be with a certain self-evolution mechanism in accordance with conflict recognition in the environment of data updating, which effectively provide the ability of model self-adjust in the dynamic conflict elimination environment [4]. It has been shown that the trend of volume, huge, coherence of data will bring a more difficult situation to work on this issue. In recently, related domain researchers are trying to using a variety of machine learning, artificial intelligence methods to explore solutions to this problem.

2 Object coherence conflict recognition and resolution related work

Although the Natural Language Processing of web information aggregation domain is more and more important in recent years, but the hybrid web data object coherence recognition and resolution is still a very difficult problem. In fact, it has been widely considered to be one of the most challenging artificial intelligence problems [5]. In recent years, lots researchers are trying to apply a variety of machine learning methods to solve the problem [6] [7]. Therefore, the recognition and resolution of the hybrid web data object coherence in the Web3.0 still needs to

be further studied. The goal of object coherence research is to identify the coherence relationships among the different web object data [8].

3. Object coherence conflict recognition and resolution Algorithm

3.1. Object coherence conflict recognition and resolution framework

This paper is concerned with the heterogeneity problem of object coherence in hybrid distributed data stream. The general framework of conflict recognition and resolution is as shown in Figure 1.

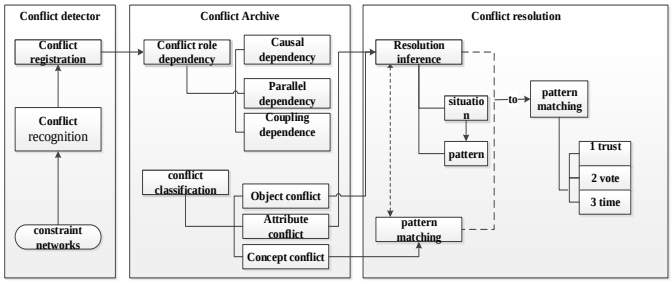

Figure 1. Process Framework for Object Coherence Conflict Recognition and Resolution

In figure 1, the process framework was divide into three phases. On the first phase, the conflict detector can use the constraint networks defined by different description rule, recognize the conflict among the object and register the conflict elements. On the second phase, the conflict elements were classified into different categories including object conflict, attribute conflict and concept conflict. On the third phase, the conflict element were resolution by pattern matching and artificial arbitration.

3.2. Object coherence conflict recognition role description and traversal

The system of conflict reasoning based on description logic must consider the priority and dependence between the conflict recognition roles. In the domain of object coherence conflict, an analysis study shows that the conflict dependency relationship exists different layers. The layer dependency relationship between conflicts will be used to define the dependent relationship between the conflict recognition roles, furthermore, determine the priority between the conflict recognition roles. When conflicts gradually derive, the role of conflict recognition will continuously extend and gradually form a better conflict recognition role model. From the perspective of graph theory, the recognition model is a directed acyclic graph, as shown in Figure 2. The node is the conflict recognition role of conflict recognition description logic. If there is a dependency between two roles of conflict recognition, we will create one directed edge which indicates one dependent role and one being dependent role.

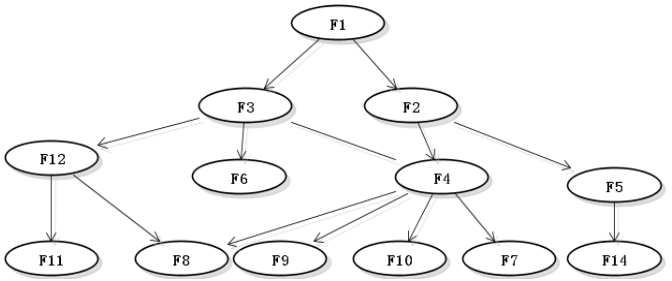

Figure 2. Conflict Recognition Role Dependency Directed Acyclic Graph

Algorithm1 $CIGRSearch(Smodel, G)$

Begin

Step1. $V := 1$.

Depth first principle is used to traverse the search of the current V node of the unchecking node sets, the output is $Evnext[]$.

$f(V.evNext[] \neq null)$

$foreach(Vnext \in evNext[] \cap NoneChecked(Vnext))$

$CIRGSearch(GetSubTree(Smodel, Vnext), G)$

Recursive loop to determine, from the leaf node to start.

Step2. $for(e \in Evnext[])$

$f(e.node \notin W \cap NoneChecked(e.node) \cap e.node \neq V)W$

$= inlcudeNode(e.node, W)$

If the $e.node$ don't belong to W and is not checked, then it will be merged into the set of nodes W to be checked. If there is no such side, that is, the nodes associated with each side have been detected and go to Step4.

Step3. For the set of nodes to be checked, begin to perform the following operations:

$for(v = GetNextNodeSetfrom(W))\{$

$currentConflict = DetecConflict(Smodel, GetNodeRole(v))\}$ Loop using the conflict recognition role of the node W to recognize conflict among them. $Smodel.conflictSet = conflictSet \cup currentConflict$

Add the conflict to the conflict Set.

$endfor$

Step4. Determine the edge that point to node V in the set T $U \mapsto V$

To the edge to return to the node U,and set $V := U$,to Step2.

If no edge $U \mapsto V$ then stop.

Step5. Output conflict set conflict Set in Smodel.

End

3.3. Object coherence conflict resolution process

From the semantic web and ontology theory, the concept conflict is because that these inconsistency of domain object is described by different terminology concepts. The property is a basic element of ontology concepts. One ontology is composed of different property. Therefore, the property conflict would inevitably lead to an inconsistency between ontology concepts. In this paper, we propose definition of conflict resolution algorithm $CRBGA$, specific as shown in algorithm 2.

Algorithm2 Object coherence conflict resolution algorithm $CRBGA$

Begin

Step1.Initial object coherence $ConceptConflictset$, $PropertyConflictSet$, $IndividualConflictSet$;

Step2. Get Conceptual conflict

$ConceptConflictset = getConceptConflictBasedG(confilctSet, G)$;

Step3. Get Property conflict

$PropertyConflictset = getPropertyConflictBasedG(confilctSet, G)$;

Step4. Get Object conflict

$IndividualConflictset = getIndividualConflictBasedG(confilctSet, G)$;

Step5. Conducting pattern matching and the resolution for individual conflicts by looping.

$For(eachConflict \in ObjectConflict)\{$

$boolResult = isExistedPattern(eachConflict, patternDB)$;

$selectedPattern = selectPattern(eachConflict, patternDB);$

$if\ (boolResult == true)$
$executePattern(selectedPattern, eachConflict)$;

$\}$

Step6. Conducting pattern matching and pattern resolution for property conflict by looping.

$For(eachConflict \in PropertyConflict)\{$

$boolResult = isExistedPattern(eachConflict, patternDB)$;

$selectedPattern$
$= selectPattern(eachConflict, patternDB);$

$if\ (boolResult == true)$
$executePattern(selectedPattern, eachConflict)$;

$\}$

Step7. Conducting pattern matching and pattern resolution for property conflict for the conceptual conflict by looping.

$For(eachConflict \in ConceptConflict)\{$

$boolResult$
$= isExistedPattern(eachConflict, patternDB)$

$selectedPattern$
$= selectPattern(eachConflict, patternDB);$

$if\ (boolResult == true) executePattern(selectedPattern, eachConflict)$;

$\}$

Step8. Remaining conflicts $otherConflict$ that unmatched results conduct the trust arbitration resolution

$if\ (otherConflict \neq null) otherConflict$
$= SrcIndexArbitrate(Smodel, otherConflict);$

Step9. Remaining conflicts $otherConflict$ after the trust arbitration resolution conduct the vote arbitration resolution

$if\ (otherConflict \neq null) otherConflict$
$= CountArbitrate(Smodel, otherConflict);$

Step10. Remaining conflict $otherConflict$ after the vote arbitration resolution conduct the time arbitration resolution

if (*otherConflict* ≠ *null*)*otherConflict*
= *TimeArbitrate*(*Smodel*,*otherConflict*);
Step11. Return the result of conflict resolution
Return(Smodel,otherConflict) .
End

4. Experimental and Analysis

Object coherence conflict resolution experiment will be explained in this paper by dividing the contents of this paper into two parts, object coherence conflict recognition and object coherence conflict resolution.

4.1. Experimental on the hybrid web data object coherence conflict recognition

The evaluation method mainly use two indicators include the recall rate and the precision rate of the query and statistical perspective, the recall rate describes the proportion of conflicts already found account for the total conflicts; the precision ratio is the proportion of true conflicts that has been identified in the conflict account for the conflicts that has been identified.As a result, *CIRGSearch* is more efficient.

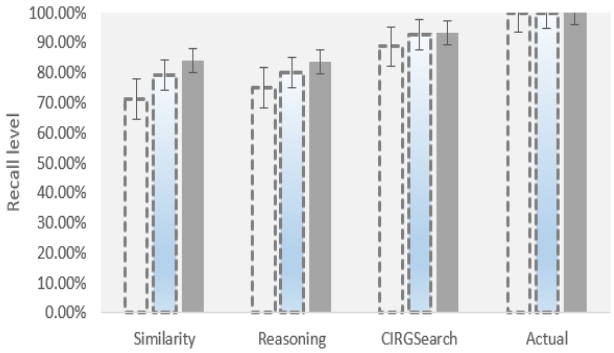

Figure 3. Contrast of Similarity, Reasoning and CIRGSearch Algorithm

From the experimental results shown in Figure 3 above, the CIRGSearch algorithm in this paper, has the highest conflict recall rate, which is obviously superior to the conflict recognition algorithm based on similarity and the conflict recognition algorithm based on rule reasoning. Using the method of logical reasoning alone is better than the single use of the similarity algorithm. This phenomenon is that the calculation of similarity degree between singer and special is difficult thus lead to some conflicts not found. Because the rule reasoning method utilizes constraint logic relationship to complete reasoning, conflict contradiction will be outstanding. It is clear that conflicting recall ratio in the third group is higher than the other two groups.

4.2. Experimental on the hybrid web data object coherence conflict resolution

After resolution of the hybrid web data object coherence conflict, the conflict resolution effect can be defined by two indicators include that the conflict resolution density index and the conflict resolution complete rate index. we select randomly 8 batches conflict from above experiment result, and statistics the conflict resolution density index, the conflict resolution complete rate can obtain results as shown in the following:

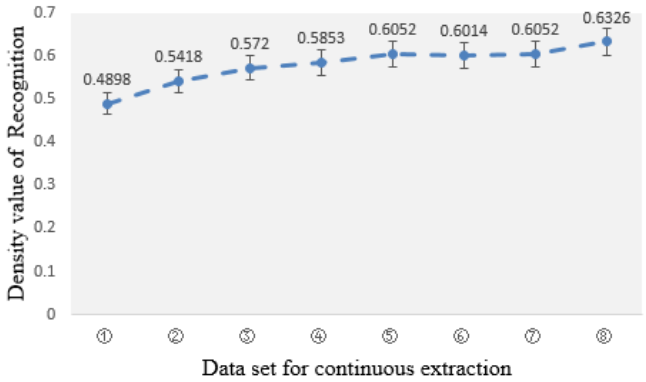

Figure 4. Density Value of Conflict Recognition in Data Sets

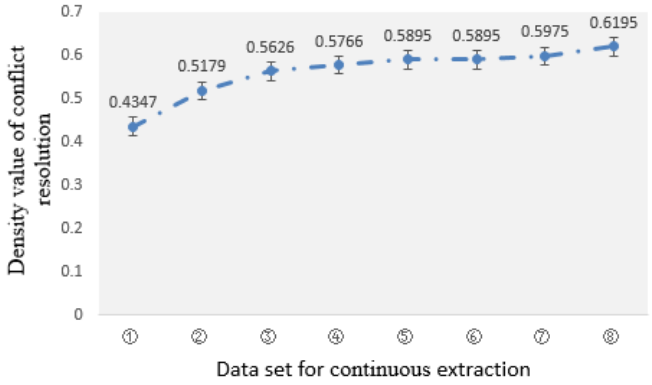

Figure 5. Density value of conflict resolution

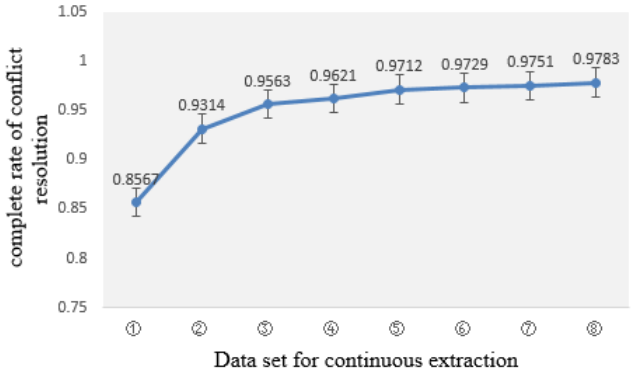

Figure 6. The Change of The Complete Rate of Conflict Resolution

As shown in the figure above, conflict recognition density index have better stability in algorithm *CRBGA*, the density index also has a certain stability; it has a good effect though the resolution of completeness index to reflect conflict recognition.

5. Conclusions

The coherence conflict resolution caused by heterogeneous between web distributed data stream is key problem in web information aggregation. Under the environment of the hybrid distributed data stream, the highly efficient heterogeneous object coherence recognition and resolution of multi-source distributed data stream will face up with great challenges in big data era. Object coherence recognition and resolution of heterogeneous big data will be a hot research direction in the field of big data aggregation.

References

[1] Engelen R V, Zhang W, Govindaraju M. Toward remote object coherence with compiled object serialization for distributed computing with xml web services [J]. In in the proceedings of Compilers for Parallel Computing (CPC）, 2010:441-455.

[2] Zhang H, Yan Z, Sun C, et al. Based on Entities Behavior Patterns of Heterogeneous Data Semantic Conflict Detection[C]. Web Information System and Application Conference. IEEE Computer Society, 2015:169-174.

[3] Ge J, Qiang B, Chen Z. Design and Application of Heterogeneous Data Semantic Integration System based on Domain Ontology [J]. International Journal of Advancements in Computing Technology, 2012, 4:260-267.

[4] Joshi A, Finin T W, Mathews M L. SYSTEM AND METHOD FOR SEMANTIC INTEGRATION OF HETEROGENEOUS DATA SOURCES FOR CONTEXT AWARE INTRUSION DETECTION, US20140337974 [P]. 2014.

[5] Nath R P D, Seddiqui H, Aono M. An Efficient Method for Ontology Instance Matching[C]. The, Conference of the Japanese Society for Artificial Intelligence. 2012.

[6] Nath R P D, Seddiqui H, Aono M. Ontology Instance Matching for Semantic Data Integration [M]. LAP LAMBERT Academic Publishing, 2014.

[7] Nath R P D, Seddiqui H, Aono M. An Efficient and Scalable Approach for Ontology Instance Matching [J]. Journal of Computers, 2014, 9(8):1755-1768.

[8] Pankowski T. Semantics Preservation in Schema Mappings within Data Exchange Systems [M]. Knowledge Engineering, Machine Learning and Lattice Computing with Applications. Springer Berlin Heidelberg, 2013:88-97.