International Journal of Hybrid Information Technology
Vol.8, No.9 (2015), pp.225-234
http://dx.doi.org/10.14257/ijhit.2015.8.9.21

Integrate Metadata by Semantic Recommendation: A Psychology Inspired Method

Xixu Fu[1] and Yuan Ren[2]

Institute of Information and Education Technology, Shanghai Ocean University[1]
School of Computer Science, Shanghai Dianji University[2]
xxfu@shou.edu.cn

Abstract

Complicated data structure can handicap the integration of systems in enterprises and universities. It is important to find an efficient way to deal with the mass documents of these heterogeneous and distributed systems to be integrated. In this paper, a three layered data scheme is introduced for data recommendation in software engineering process inspired by spreading activation theory. Metadata can be more efficiently managed with this scheme in integration and analysis.

Keywords: *Data Integration, Recommender System, Spreading Activation, Metadata*

1. Introduction

Data integration is an essential work in system integration. However, mass heterogeneous tables and views seriously handicaps the efficiency of data integration. It is a hard task to comprehend data in old systems and utilize them effectively into new systems and platforms. Data structure of core systems in a university is shown in Table 1.

Table 1. Data Structure of Main Systems In A University

System	Database	Data warehouse	Number of Tables
Student Management	Oracle	none	306
Faculty Management	Oracle	none	183
Teaching Management	Sybase	none	167
Course Management	MySQL	none	69
Research Management	Oracle	Implemented	183
Financial Management	DB2	Implemented	137

Concerning great number of attributes and relations in these tables, metadata can be complicated for data integration in enterprises with such systems. An efficient manage method of semantics in metadata should be advanced to enhance the efficiency of metadata analysis.

Fortunately, study in psychology and knowledge systems provide a way to deal with such problems [15]. Spreading activation model is a famous theory about reasoning in human memory based on ACT-R theory [1]. The theory can explain the nature of human association and some problem solving methods [5]. It is efficient to use this theory to model knowledge systems. Metadata can also be managed use this model. With the rapid development of XML and knowledge grid [9], it is a good idea to setup a scheme to enhance the management and design of metadata in data integration and recommend data structure automatically.

ISSN: 1738-9968 IJHIT

This paper focus on building an efficient semantic recommending system for metadata based on spreading activation theory. The massive document reading task can be replaced with several efficient interacts with recommending systems. ETL process can also be built automatically.

The paper analyzed related works on psychology theory, knowledge systems and metadata management respectively. Then, a three layered scheme was advanced for data integration. After that, spreading activation and metadata recommendation was discussed. Finally, data integration process based on the model was evaluated and compared with traditional manual data integration process.

2. Related Work

2.1. Spreading Activation Theory

Spreading activation theory can explain how knowledge is applied in human minds. Models also showed the superiority of spreading activation theory in dense knowledge environment [5].

Early version of spreading activation theory was generated in the study of language and semantics [2]. In classical description of spreading activation theory, related semantic elements such as words can activate related elements to form needed scenarios. Spreading activation serves the function of quickly spreading an associative relevancy measure over declarative memory [1].

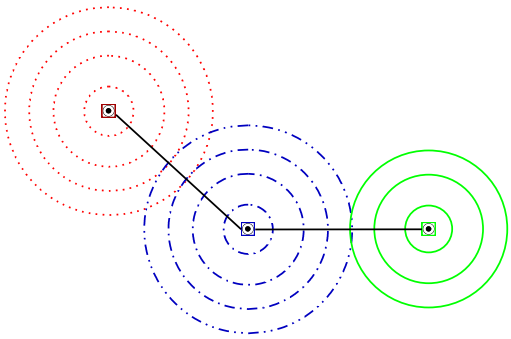

Figure 1. Soaking among Nodes

As shown in Figure 1, a node can send information to activate connected nodes by sending activation information. Activated nodes can generate new information to activate other nodes. In the activation process, heuristic information can be generated and the soaking process will be suppressed and will stop at last.

Spreading activation theory had been successfully used in recommendation systems [4] and information retrieval [3]. It is a normal way to reduce the complexity of document management with recommending systems.

2.2. Knowledge Grid and Software Design

XML is a popular technology for knowledge representation and management. Knowledge grid [9] and data grid [10] have been implemented for software design and distributed data integrating. It is reasonable to use XML to manage metadata. Metadata of documents can also be built with XML.

2.3. Metadata and Data Integration

Management of metadata plays an important role in the design and maintenance of software and database. Management of metadata becomes more and more explicit

since year 2000 [7]. Generating of metadata of web pages has been advanced in year 2004 [8]. Distributed metadata management [6] and mining of metadata [11] have been advanced in recent time.

Data integration also becomes a popular topic in research and engineering. Thomo advanced a new method for XML data [12]. A framework on data integration, data mining and decision support had been advanced in recent time [13]. Service oriented distributed and heterogeneous data integration [14, 18] had been advanced too. Integration of different type of data becomes matured for realizing.

Integration of different data models have been discussed in recent works too. Exchange from XML to relation model [16] and their equivalence in FDs [17] have been discussed in Davidson and Vincent's paper respectively in year 2007.

2.4. E-R graph and Data Models

E-R graph is an important meta data representation method [19]. Methods have been advanced to recognize entities and relations from data [20, 21]. Different models such as object-oriented models can be represented in E-R diagrams easily [23, 24]. It is good to extend E-R graph to represent meta data. Searching based on E-R graph have been advanced too [22].

3. Three Layered Data Scheme & Metadata Unification

For the convenience of data integration, data can be divided into three layers. Data layer consists of data which is stored in various types of media and forms. Meta layer consists of metadata including data models, data structures and related explains. Semantic view layer is the interface layer for data integration which consists of semantic elements which can be used for integration. Figure 2 shows the three-layered scheme for data integration and metadata management.

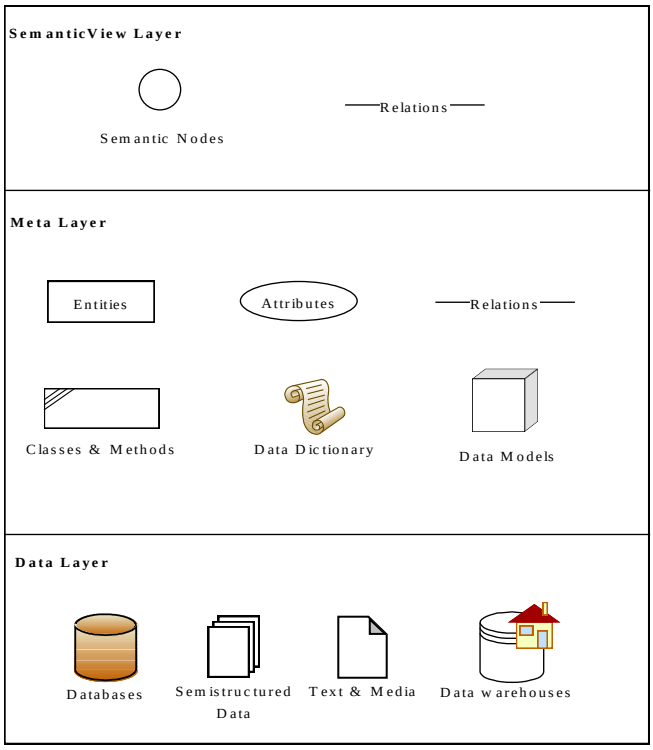

Figure 2. Three Layered Data Integration

Heterogeneous data can be divided into three layers. The semantic view layer unified different data structure and data model into semantic nodes and relations. Heterogeneous data and metadata should be unified to support the integration.

3.1. E-R graph and Data Models

Data layer is a set of data in all systems. Metadata needs to be extracted from data in this layer. Data layer can also provide references for the unification of metadata. For example, text or XML data without specified field data type may gain data type from data.

3.2. Unifying Meta Layer

Different data models and structures can be implemented in the meta layer. However, entity relation (E-R) model can interpret these models and structures with entities and relations. Objects can be implemented to extend the expressive power of E-R model.

According to the E-R model, a more materialized data scheme can be defined to unify the meta layer.

Definition 1. An E-R graph is a 4-tuple g(**T,A,R,I**). **T** is a set of table consists of attributes. **A** is a set of attributes. **R** is a set of relations which relate the tables. A relation can be represented as r(a1,a2). **I** is a set of inclusion among tables and attributes.

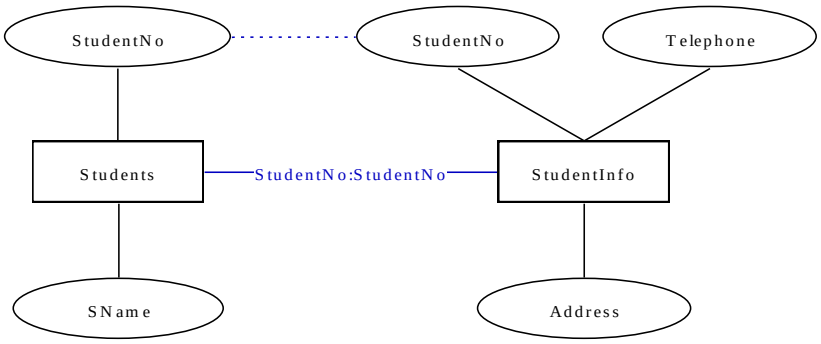

Figure 3. An E-R Graph in a Visual Mode

Figure 3 shows an entity relation graph. There are two tables named students and *studentinfo* respectively. The two tables are related with the attribute named *StudentNo*. Entities and relations should be extracted from data layer and meta layer to form E-R graphs.

3.3. Generating Data View Layer

Data view layer provides interfaces for data integration and provides data extraction method through meta layer.

Data view layer consists of E-R graphs which share data structure and semantics for integration and configurations of related systems. Schemes of views should be stored for integration. Configure information should be stored for every system.

4. Semantic Recommendation

To find attributes and relate them into a view suitable for data integration, semantic recommendation must be carried out based on the E-R graph in meta layer.

4.1. Spreading Activation on the E-R Graph

Attributes are regarded as basic elements which carry semantics. The objective of semantic recommendation is to generate views which relate these elements together. All attributes needed in integration should be indicated as target nodes from the E-R graph in the meta layer.

Spreading activation process can be divided to two stages. The first stage is the semantic searching process in which a graph consist all needed semantic elements is built. The second stage is the pruning process. In this stage, redundant attributes and relations are erased to provide a simple view.

An E-R graph can be regarded as a graph consists of two kinds of edges and two different granularities of nodes. In the semantic searching process, it can be regarded as a graph consists of attributes and relations. All attributes are regarded as nodes of the graph. All relations are regarded as edges of the graph. In the pruning process, tables and relations are the basic elements considered in order to make the recommendation more adaptive.

Figure 4 shows the semantic searching process which generate the initial E-R graph to create needed view for integration.

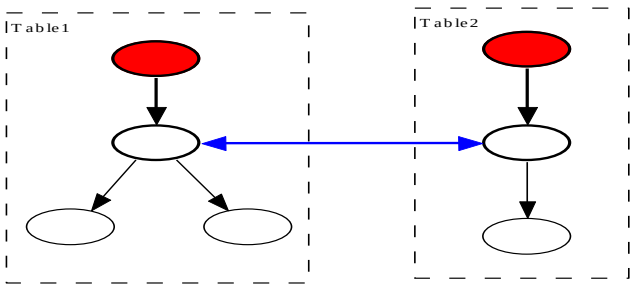

Figure 4. Semantic Searching Processes

The figure shows semantic searching in two tables. Attributes are represented as ellipses. Red ellipses represent attributes required in integration which are regarded as the initial nodes. By soaking from the initial nodes, a graph connecting the initial nodes to a view is generated.

Figure 5 shows the pruning process. Tables provide no useful attribute and relations are regarded as redundant table.

Definition 2. Let t be a table. Let A be the attribute set of t. Let At be the set of attributes needed in integration. Table t is a redundant table if:

- The table is not related to two or more non redundant tables and
- Not exist an attribute a $\in$ {a| a $\in$ A_t $\cap$ A, a is not a foreign key related to non redundant tables}

Erasing all redundant tables can make the view simple and clear.

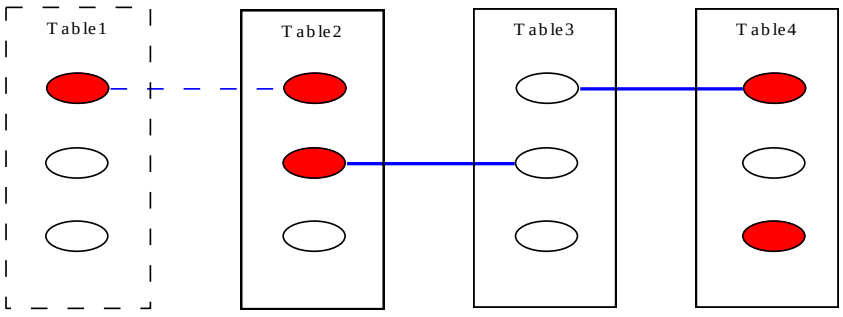

Figure 5. Pruning Process

The figure shows the pruning process. Table1 provide only a requested attribute which is also provided by Table 2. No tables are related through Table1. So Table 1 is deleted. Table 3 provide no requested attribute. However, it linked Table 2 and Table 4 together. So, it is not deleted.

4.2. Algorithm of Meta Structure Recommendation

E-R graph of an information system is normally a connected graph. If it is not a connected graph, it can be regarded as different systems.

At the semantic searching process, E-R graph can be represented as a graph with attributes as notes and all relations and inclusion as edges. Let A be the required attribute set for integration. Let Gg be the E-R graph in meta layer, N be nodes in Gg and R be edges of Gg. The semantic searching algorithm can be described as the algorithm in Figure 6.

```
Input Attribute Set A={A1,....,An}, Metadata Graph Gg(N,R)
Begin
  A=A∩Gg.N
  for every node Ni in Gg
   if Ni∈A then
     Gk=(Ni)
     GS=GS+ Gk
     k=k+1
   end if
  next i
  do while not ∃(Gi=Gg)
     for every Gi in GS
         Gi=Gi+Neighbor(Gi)
         if ∃(Gi∩Gj≠ϕ and Gj⊆GS) then
            Gi=Gi+Gj
            GS=GS-Gj
         end if
         if A⊆Gi then return Gi
     next i
  loop
  return Gg
End
```

Figure 6. Generating Candidate E-R Graph

After generating the semantic E-R graph, pruning algorithm should be implemented to simplify the interface E-R graph. In the pruning process, tables are regarded as nodes. Input of the pruning algorithm become E-R graph G(T,R) generated from the semantic searching algorithm and Attribute set A needed for data integration. The algorithm is described in Figure 7.


```
Input:Attribute Set A={A1,....,An}, E-R Graph G(T,R)
Begin
  do while G changes
     for every table Ti in G
        if Ti.Neighbornumber<2 then
           KeyAttNum=0
           for all Attributes a in Ti.Attributes
              if a∈A and a∉R.Attributes then
                KeyAttNum=KeyAttNum+1
              end if
           next
           if KeyAttNum=0 then G=G-Ti
     next i
  loop
  return G
End
```

Figure 7. Pruning Algorithm

Pruning of initial E-R graph is shown in this figure. This algorithm generates simple E-R graph for data extraction and integration.

5. Implement and Results

5.1. Implement Algorithms to Recommend Data Extraction

The realizing of recommendation algorithms in software takes five steps which are described below.

- Input XML scheme which describe the attributes needed for integration.
- Find attributes in the E-R graph in meta layer.
- Run the searching and pruning algorithms in systems to generate E-R graphs as candidate graphs.
- Divide candidate E-R to loop-free E-graphs and create candidate views.
- Combine candidate views to one view for ETL.

5.2. Input XML Scheme

An input scheme corresponds to a view in data integration. In order to find attributes accurately, XML scheme should provide enough information. The syntax of XML scheme can be described in Figure 8.

```
<View Name=[Name]>
<Description>[Description of view]</Description>
<Connection>[Connection String]</Connection>
<Attribute Name=[Name] Datatype=[Data Type]>
<Description>[Description]</Description>
<Alias>[Alias]</Alias>...<Alias>[Alias]</Alias>
    </Attribute>
    ...
<Attribute Name=[Name] Datatype=[Data Type]>
<Description>[Description]</Description>
<Alias>[Alias]</Alias>...<Alias>[Alias]</Alias>
    </Attribute>
</View>
```

Figure 8. XML Scheme for Meta Data

5.3. Attribute Matching

Attributes in schemes and system E-R graphs can be matched according to the name, description or alias. All attributes matched are regarded as initial nodes for semantic searching program.

5.4. Recommendation as Data View

After the matching of attributes, algorithms can be run for the recommendation of views. The system also provides the matching score which is computed based on the similarity of the generated views and the target view. After the recommending process, SQL statements can be optimized manually. Then, candidate views can be merged to generate ETL schedule.

5.5. Dividing and Combination of Views

Sometimes, an attribute can be matched with two or more attributes in different tables. When this situation happens, generated view will be divided into two or more views according repeated attributes and their distribution in tables. One candidate view can consist of one set of attributes.

5.6. Result of Implementing

The system is implemented in the integration of systems in a university. 1836 tables and about 16000 fields were concerned in this integration. The result can be shown in Table2.

Table 2. Depth and Effect Of Integration Manually and Automatically with Recommendation System

Metrics	Manually	With Recommendation
Data views integrated	12	186
Fields integrated	126	1831
Relations integrated	6	372
Time used (Days)	90	4

It is resulted that the implementing of recommending system can greatly enhance the efficiency and depth of system integration.

6. Conclusion

This paper advanced a semantic recommendation method based spreading activation theory for data integration in complicated data environment. The method was implemented in the integration of digital campus project. The result of practice indicated that the implementing of the recommending system can greatly enhance the depth and efficiency of data integration.

7. Discussion and Future Works

Spreading activation based algorithms are efficiency aggressive algorithms [5]. It is often too complicated to search the full combinations of all nodes in complicated data environments. It is also possible to regard the full E-R graph of all systems as the result of semantic searching and run pruning algorithm with it. However, it takes much more to erase redundant tables while get almost the same result in practice.

In the pruning process, table is regarded as the basic element. All attributes of remained tables are shown in the E-R graph. If some attributes can not be matched is found in E-R graph, it is available to be added manually. It is more of a problem of adaptability than efficiency to use table as the basic element in pruning process.

Acknowledgements

This research was financially supported by the Open Fund of Shanghai Dianji University for Computer Application Technology (No. 13XKJ01). Special thanks should also be given to Dr. Yuan Ren as the communication author.

References

[1] Anderson John R. and Pirolli Peter L., "Spread of activation", Journal of Experimental Psychology: Learning, Memory, and Cognition, vol. 10, no. 3, **(1984)**, pp. 791-799.
[2] Collins, A. M., Loftus, E. F., "SPREADING ACTIVATION THEORY OF SEMANTIC PROCESSING", Psychological Review, vol. 82, no. 6, **(1976)**, pp. 407-428.
[3] Crestani, F., Loftus, E. F., "Application of spreading activation techniques in information retrieval", Artificial Intelligence Review, vol. 11 no. 6, **(1997)**, pp. 453-482.
[4] Yolanda B. et al, "Exploring synergies between content-based filtering and Spreading Activation techniques in knowledge-based recommender systems", Information Sciences, vol. 181, **(2011)** , pp. 4823-4846.
[5] Xixu Fu and Hui Wei, "Problem Solving by soaking the concept network", Computer Science and Information Systems, vol. 8 no. 3, **(2011)**, pp. 761-778.
[6] Juan W. et al, "MHS: A distributed metadata management strategy", Journal of Systems and Software, vol. 82, **(2009),** pp. 2004-2011.
[7] Arun S., "Metadata management: past, present and future", Decision Support Systems, vol. 37, **(2004)** ,pp. 151-173.
[8] H. Stuckenschmidt and F. van Harmelen, "Generating and managing metadata for Web-based information systems", Knowledge-Based Systems, vol. 17, **(2004)**, pp. 201–206
[9] Wu Z. H. "Knowledge Base Grid: A Generic Grid Architecture for Semantic Web", J. Comput. Sci & Technol, vol. 18 no. 4, **(2004)**, pp. 462–473.
[10] H. Kevser Sunercan. "A systematic approach to the integration of overlapping partitions in service-oriented data grids", Future Generation Computer Systems, vol. 27, **(2011)**, pp. 667–680.
[11] Melike S. and Vincent W., "Automatic metadata mining from multilingual enterprise content", Web Semantics: Science, Services and Agents on the World Wide Web, vol. 11, **(2012)**, pp.41-62.
[12] A. Thomo and S. Venkatesh, Rewriting of visibly pushdown languages for XML data integration. Theoretical Computer Science, 412, 5285–5297 **(2011)**
[13] Yi P. "An incident information management framework based on data integration, data mining, and multi-criteria decision making", Decision Support Systems, vol. 51, **(2011)**, pp. 316-327
[14] H. Kevser Sunercan. "A systematic approach to the integration of overlapping partitions in service-oriented data grids", Future Generation Computer Systems, vol. 27, **(2011)**, pp. 667–680.

[15] Park W. “A Personalized Multimedia Contents Recommendation Using a Psychological Model”, Computer Science and Information Systems, vol. 9, no. 1, **(2012),** pp. 1-21
[16] Davidson S. “Propagating XML constraints to relations”, JOURNAL OF COMPUTER AND SYSTEM SCIENCES, vol. 73, no.3, **(2007)**, pp. 316-361
[17] Vincent M. W. “On the equivalence between FDs in XML and FDs in relations”, ACTA INFORMATICA, vol. 44, no. 3-4, **(2007)**, pp. 207-247
[18] Peiyun Z. and Rongjian X., "Heterogeneous Information Integration Based on Services Composite Process", INFORMATION- An International Interdisciplinary Journal, Vol.14, No.12, **(2011)**, pp. 3941-3948.
[19] Fan, X. H. and Sun, M. S., "Knowledge representation and reasoning based on entity and relation propagation diagram/tree", Intelligent Data Analysis, vol. 10, no. 1, **(2006)**, pp.81-102.
[20] Fan X. H. and Sun M. S., "A method of recognizing entity and relation", Lecture Notes in Artificial Intelligence, vol. 3561, (2005), pp. 245-256.
[21] Cheng X. Y. "The Overview of Entity Relation Extraction Methods", INTELLIGENT COMPUTING AND INFORMATION SCIENCE, vol. 134, **(2011)**, pp. 749-754.
[22] Chakrabarti S. "Index design and query processing for graph conductance search", VLDB JOURNAL, vol.20, no. 3, **(2011)**, pp. 445-470.
[23] John L. Knapp, "ER isomorphisms and uniqueness conditions", Data & Knowledge Engineering, vol. 26, no. 3, **(1998)**, pp. 271-290.
[24] De Lucia. "An experimental comparison of ER and UML class diagrams for data modeling", EMPIRICAL SOFTWARE ENGINEERING, vol. 15, no. 5, **(2010)**, pp. 455-492.

Authors

Xixu Fu, got his master degree in 2007 at Fudan University. Now he is an engineer in Shanghai Ocean University and a ph. D candidate in Fudan University. His research interests include artificial intelligence, psychology, cloud computing and software engineering.

Yuan Ren, was born in Changchun city in 1984. He earned his doctorate of science from the School of Computer Science of Fudan University in 2013. He is now an instructor in Shanghai Dianji University. His research interests include computer vision and computer teaching.

