Semantic Graph Knowledge Repository Based Knowledge Discovery System for Customer Value

Semantic Graph Knowledge Repository Based Knowledge Discovery System for Customer Value Assessment

Jiandong Li
Graduate School of Computer and Information Sciences
Hosei University
Tokyo 184-0003, Japan
jiandong.li.94@stu.hosei.ac.jp

***Abstract*—Retail data available for consumer-oriented company is a precious asset which can deliver useful insights in decision making and marketing strategy. KID (Data-Information-Knowledge) model is a generic from data to knowledge cognitive model. It bases on how human process outside information. It can be applied to retail business for supporting retail data analytics. Knowledge repository is a key element of KID model. In this paper, a retail semantic graph knowledge repository based knowledge discovery system is proposed and developed for KID model to apply in limited retail data. The proposed knowledge repository integrates Neo4j graph database, retail ontology designed by Maryam Fazel Zarandi and Jess rule engine. It interprets streaming data into meaningful information and assimilate meaningful information into graph knowledge repository to update knowledge by pre-embedded prior objective-oriented algorithms knowledge in algorithm pool. The deductive reasoning capability is provided by Jess rule engine, so it can deduce answers to retail queries from graph knowledge repository. A customer value assessment case study by using Recency-Frenquecy-Monetary (RFM) analytic model and K-means algorithms is given to demonstrate the proposed semantic graph knowledge repository based knowledge discovery system.**

Keywords—KID Model; Graph Knowledge Repository; Retail Ontology; Knowledge Discovery Rule Engine; Business Intelligence

I. INTRODUCTION

Data analytics can capture value and delivery useful insights from data. Now it is a time that companies must take advantage of all available data to do advanced data analytics to guide their decision making to occupy a chunk of market share against competitors. However, retail companies still have difficulties in processing large amount of data to make better predictions and marketing strategies for their survival. As marketing becomes more customer-centric, more analytics and research should be done on customer purchase behavior.

KID model [1] based on a cognitive approach is proposed to support retail big data analytics. KID model is human-like cognitive data-information-knowledge cyclic process. Streaming data is interpreted into meaningful information by prior or expert knowledge stored in the knowledge repository of KID model. Then, this meaningful information is absorbed or assimilated into existing knowledge repository. Knowledge will be added, modified or updated according to this meaningful information. Knowledge management and knowledge store design are always a hot topic. The core part of KID model is K-Store. It is a knowledge repository which should represent items and relationships of knowledge by corresponding structures and support data interpretation and knowledge assimilation.

Retail ontology is a formal description of retail concepts, entities and their relationships. The retail concepts and their ties exist in business processes from supply chaining to sales information to customer transaction data [2]. Retail ontology can be used to define retail business logic, capture business semantics and draw further knowledge as deductive reasoning capability is provided by an inference engine. Maryam Fazel Zarandi designed and developed retail ontology which combined description logic and first-order logic [3].

A rule-based knowledge discovery engine embeded retail semantic graph knowledge repository [4] is proposed and developed as K-store of KID model. It bases on Grüninger and Fox method [5], Neo4j graph database [6], retail ontology designed by Maryam Fazel Zarandi and Jess rule inference engine [7]. It interpret retail data into information, assimilate information into graph knowledge repository to update knowledge and deduce answers to retail queries based on the general knowledge of retail business. Learned from Grüninger and Fox method, the semantic graph of our limited retail data can be constructed from existing retail ontology developed by Maryam and be modified. Neo4j graph database are used as a knowledge repository and store knowledge. It models concepts as nodes, relationships as edges and stores properties as hash-objects according to retail semantics captured from retail ontology. A rule-based knowledge discovery engine includes working memory, an inference engine and a rule base which has forward chaining rules and backward chaining rules. It is embedded in graph knowledge repository and uses hybrid reasoning method to deduce answers to retail queries.

We also develop semantic graph knowledge repository based knowledge discovery system to make the proposed retail semantic graph knowledge repository applicable in real retail

Supervisor: Professor Runhe Huang

business world. RFM analytic model is a method to differentiate important customer from large data. This paper illustrates a customer value assessment case study by using RFM analytic model and K-means algorithm to demonstrate the proposed knowledge discovery system..

II. Semantic Graph Knowledge Repository Based Knowledge Discovery System

The system architecture of retail semantic graph knowledge repository based knowledge discovery system is shown in Fig. 1. Expert knowledge in retail domain is pre-embedded into the algorithm pool. When streaming retail data comes, function from algorithm pool is called to load piece retail data into D-store. The design of D-store combines relational database and a secondary storage named Enhanced KStore [8]. The D-store continuously stores small sets of data that has passed through. As time goes by, it becomes retail business big data which is enough to delivery useful insights. Data interpretation is an object-oriented process. Given a specific objective, corresponding algorithms pre-embedded in algorithm pool are called to interpret data into meaningful information. The information will experience a series of information transformation and new information generation. Relational database (I-store) is also used to store information. Knowledge is contained in information. What we expect for the objective is knowledge. Then knowledge is assimilated into our graph knowledge repository (K-store) based on Neo4j database. As time goes, knowledge can accumulate and aggregate. With more and more knowledge about customers contributed by our system, each customer model may approximate its real customer referring to their purchase behavior aspect. The forward chaining and backward chaining rules are applied to deduce answers to retail queries to help retailers to make decision and identify new market opportunities.

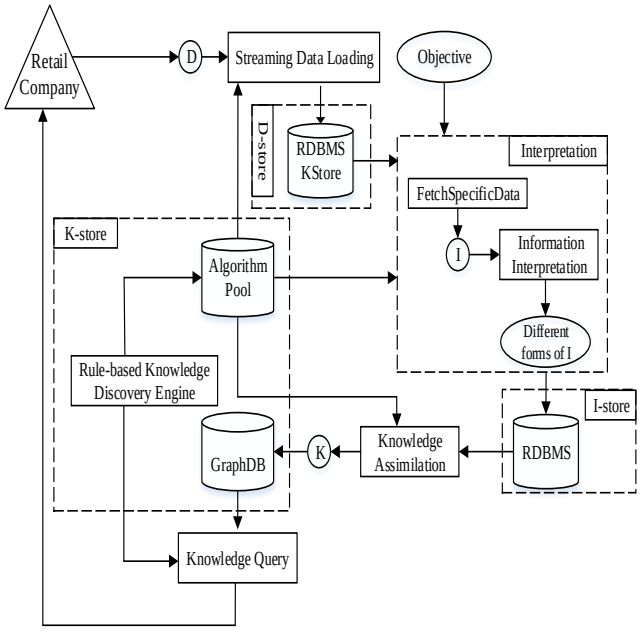

Fig. 1 Architecture of semantic graph knowledge repository based system

III. Overview of KID Model

How humans process continuously incoming small sets of data and response it is considered into the design of KID model.. Instead of processing big data as usual, KID model processes incoming data piece by piece. This is a natural form of human cognition. KID model based system can accumulate experience, knowledge and gain useful insights in each transformation process of iterative and incremental data-information-knowledge cycle.

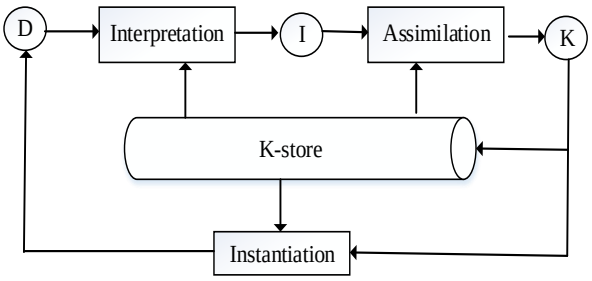

Fig. 2 The overview of KID conceptual model

Fig. 2 illustrates the KID conceptual model. Streaming data, *D*, is regarded as stimulant to activate items of knowledge in the knowledge repository. The data is interpreted into meaningful information, *I*, which is understandable to a specific domain. The generated meaningful information, *I*, is assimilated into knowledge repository, *K-store*, by identifying elements in *K-store* which is related to the generated information by traversing *K-store*. The knowledge in knowledge repository can be instantiated to other practical application.

The core part of KID model is *K-store* which is a cognitive model of the memory system and contains two components, short-term memory for current information processing and long-term memory for storing useful information and knowledge. The design and implementation of *K-store* is related to such factors like knowledge representation, an inference mechanism. Different *K-store* representation structure can lead to different implementation of interpretation and assimilation function.

IV. Streaming Data Loading

Instead of processing big data as usual, small sets of incoming data are loaded into D-store. The new coming data stimulates the graph knowledge repository based knowledge discovery system. It call functions such as *loadDataToDB* pre-embedded in algorithm pool to load data. By continuously importing incoming small sets of data, it becomes retail business big data which is sufficient to form useful insights.

We need have a good data warehouse or data structure to store and access business data to support efficient data access and analytics. Traditional relational database is still a mainstream business data storage warehouse. With the advent of graph database, the use of it to store retail business data is becoming popular. However, to preprocess raw retail data and model them as a graph is a very complex and time-consuming

task. The design of D-store combines relational database and KStore which is a tree-based datastore comprising two or more levels of forests of interconnected trees.

KStore [9] is a data structure proposed by Jane Campbell Mazzagatti based on the Phaneron of C. S. Peirce. It is designed and developed as a storage engine to support business intelligence data storage, queries and analysis. Fig. 3 shows an instance of KStore data structure. The generation and data access of KStore use two kinds of linked lists of which one represents one hierarchical relationship of nodes in the tree, and the other records all the other relationships between the dataset elements encountered in the input. When the input business dataset is large, the length of linked lists become longer and KStore query becomes time-consuming. We improved the query performance of KStore by using Trie data structure and Dictionary data structure. Table 1 shows the average query performance comparison between KStore and enhanced KStore.

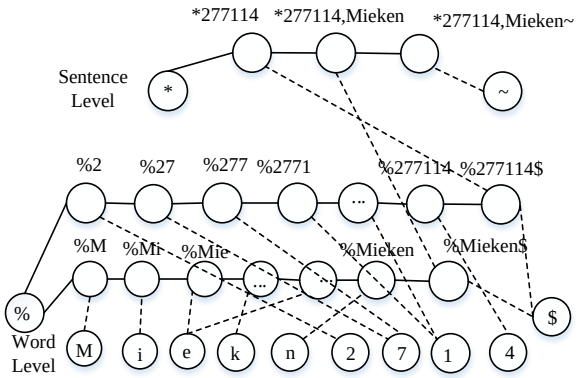

Fig. 3 An instance of KStore data structure

TABLE I. AVERAGE DATA ACCESS TIME COMPARISON

Query Request	Mazzagatti's KStore	Enhanced KStore
Words including letter '2'	865.5254s	0.0083s
Records including word 'Efiacoce'	0.0094s	0.0014s
Words starting with letter 'E'	0.0040s	0.0006s
Records starting with word '277121'	865.5271s	0.0017s

V. INTERPRET DATA INTO INFORMATION

It is worthwhile to point out that our semantic graph knowledge repository based knowledge discovery system is a task-driven process. Given a specific objective, *fetchData* function from algorithm pool is called to extract required retail data. The fetched data is endowed with meaning. It is meaningful to the existing knowledge and can be interpreted into information by activated knowledge stored in algorithm pool. In the interpretation process, information experienced a series of information transformation and new information generation. The generated information can be as valuable as one's collected experience and is also stored in the relational database (I-store).

VI. ASSIMILATE KNOWLEDGE INTO GRAPH KNOWLEDGE REPOSITORY

The newly generated information in the interpretation process is linked to relevant knowledge in retail semantic graph knowledge repository (K-store). The existing knowledge is updated or enriched when new knowledge is derived due to the newly assimilated information. It can accumulate and aggregate knowledge by continuously absorbing knowledge into graph knowledge repository. With more and more knowledge about customers contributed by the graph knowledge repository, the knowledge referring to customer's purchase behavior aspect can accumulate and aggregate. The retail semantic graph knowledge repository is a digital description of a retail business which grows with continuous contribution of knowledge about the business.

A. *Semantic graph of limited retail data*

Customer purchase record data are provided by a retail business corporation for retail data analytic. Customer profile dataset has 22 attributes, such as customer ID, last purchasing data, prefecture, etc. Purchase records dataset contains 38 attributes including item name, number of purchased, order code, customer ID, shipment, tax, discount price, etc. These two kinds of dataset have over 20000 customers and 400,000 purchase records over a year. Fig. 4 shows the detail of the two datasets.

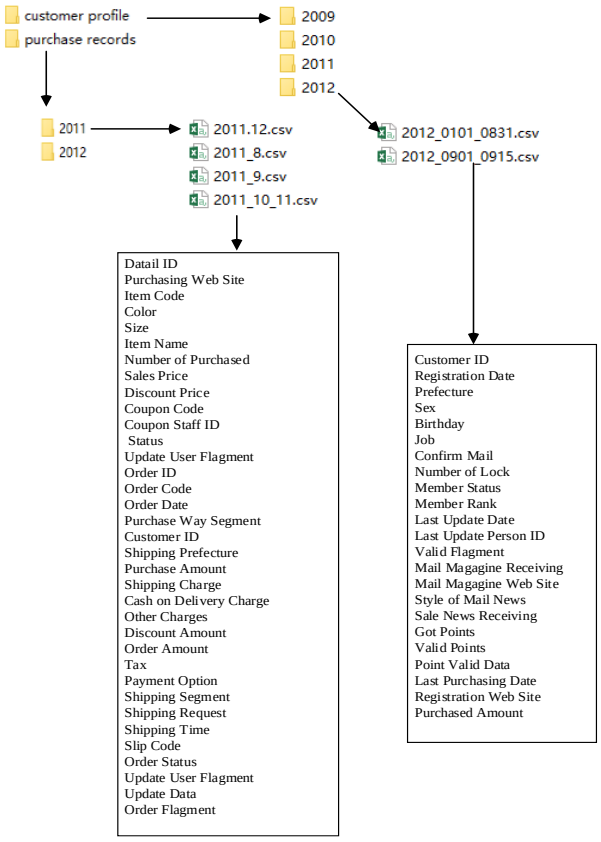

Fig. 4 A glance at limited retail data

The method that transforms retail ontology to semantic graph has are 5 important steps. It includes reusing the existing retail ontology developed by Maryam Fazel Zarandi, enumerates concepts, capture relationships between concepts, define the properties of concepts and define the facets of properties. Fig. 5 illustrates the semantic graph of our limited retail data.

Limited to owned dataset, not all concepts and their relationships extracted and enumerated from retail ontology are used to build the semantic graph of our data. As shown in Fig. 6, the concepts including *Address, Person, Customer, Mail, CustomerOrder, OrderLineItem, Shipment, Tax, Zone, SKU (stock keeping unit), Coupon, Promotion*, etc., are modeled as nodes in the Neo4j graph database. The relationships including *hasDiscountPrice, hasMail, hasPerson, hasCoupon, belongsToRFMSegment, CustomerCustomerRelationship, hasTax, hasCustomerOrder, hasOrderLineItem, hasZone, customerOrderHasSKU, hasSKU, hasInventorItem, hasProduct, hasAddress,* are represented as edges in the graph database. The properties of each concept and their values are stored as hash objects.

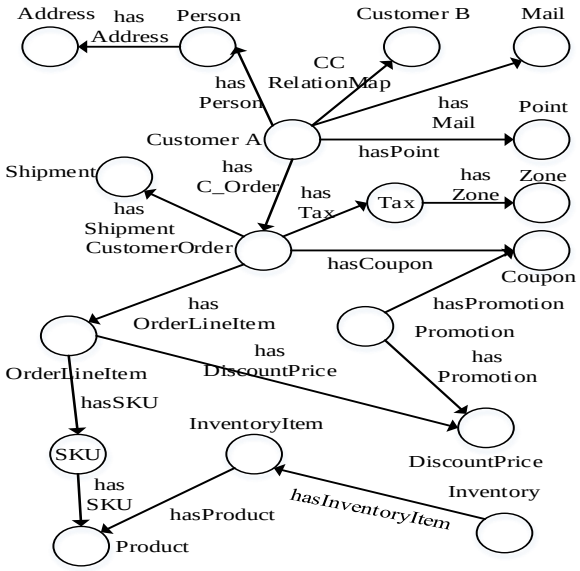

Fig. 5 Semantic graph created on limited retail data

B. Retail semantic graph knowledge repository

As is shown in Fig. 6, the novel graph knowledge repository of KID model is designed and based on retail business semantic relationships in retail ontology using Neo4j graph database technology [6]. Entities and their properties, relationships, constraints and behaviors in retail business are all represented in this knowledge repository. A rule-based knowledge discovery engine is also embedded in graph knowledge repository to deduce answers to retail queries based on the general knowledge of retail business. The inference engine controls the whole process of applying the rules to the working memory to obtain the outputs. It includes three parts: pattern matcher, an agenda and an execution engine. All the rules is a kind of instruction like if-then statement and are stored in the rule base. The working memory stores the facts and the instances of facts that rule engine operates on.

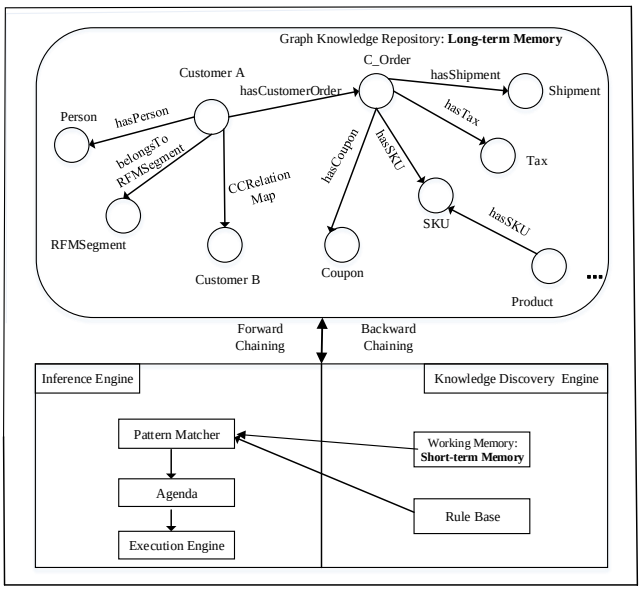

Fig. 6 The graph knowledge repository based inference mechanism

VII. DEDUCE ANSWER TO RETAIL QUERIES

A Jess rule engine [7] is used in the retail semantic graph knowledge repository to provide the reasoning ability to deduce answers to retail queries. Learned from the hybrid reasoning method of semantic digital library [10] [11], we adopt both forward and backward reasoning. The backward reasoning method is used to gather knowledge from graph knowledge repository based on Neo4j graph database and assert them as facts into working memory of rule engine. The forward reasoning method is responsible for answer a given query.

VIII. CASE STUDY: CUSTOMER VALUE ASSESSMENT

Recency-Frequency-Monetary (RFM) analytic model is a method to identify important customer from large data. It represents customer behavior characteristics by the following three variables: Recency which refers to the interval between the last purchase data and present, Frequency which refers to the transactions number in a particular period and monetary which refers to consumption money amount in a particular period.. This case study constructs a customer value assessment by uses RFM analytic model [12] and K-means [13] algorithms. It interprets customer data and purchase record data into customer value and customer segmentation, assimilates customer value and customer segmentation into retail semantic graph knowledge repository and deduces answers to retail queries from graph knowledge repository including

(1) What is the R-value, F-value or M-value of customer X?

(2) What is the customer value of customer X?

(3) Which customers belong to customer segmentation Y?

Customer value and its segmentation are clustered based on RFM attributes and K- means algorithm. The RFM model is regarded as input attributes to yield quantitative value for K-means clustering. Fig. 7 illustrates the workflow in this case study.

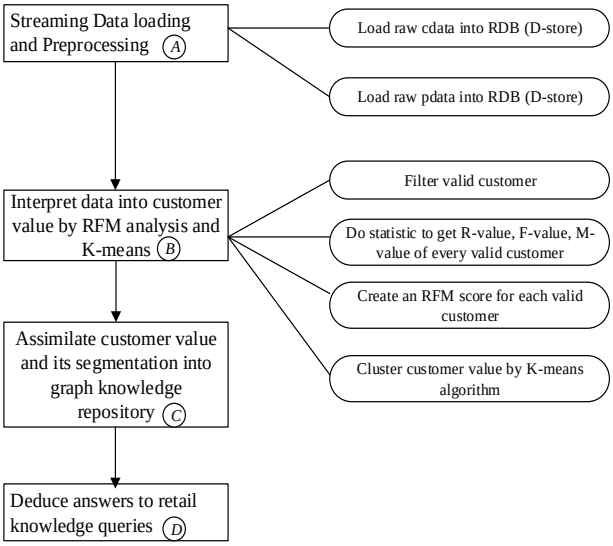

Fig. 7 Flow diagram of customer value assessment

The algorithm knowledge include *LoadCdataToDB, CreateRFMDaysTable, LoadPdataToDB, FilterValidCustomer, CreateRFMTable, CreateRFMScoreTable, ExportDB2CSV, CSVToArff, CustomerValueKmeansClustering, WriteToGraphKR*. They are implemented as implemented by Java class and are pre-embedded in the algorithm pool. *ReasoningQueryFromGKR* rules can deduce answers to above three retail queries.

A practical dataset which includes customer data of 2011 year and purchase record data from 2012/01/01 to 2012/09/12 is used in this case study to demonstrate the proposed procedure. The proposed procedure as given in Fig. 7 can be divided into the following four processes:

A. Streaming data loading and preprocessing

At first, *LoadCdataToDB* and *LoadPdataToDB* function are called to load streaming customer data and purchase record data into mysql relational database (D-store). The raw dataset is provided as CSV file. The two kinds of raw dataset have 101703 customers and 287964 purchase records. There are 60 attributes in the raw dataset. However, only 4 attributes including *CustomerId, LastPurchaseDataTime, Frequency, Monetary*, are used to cluster customer value. It took about 4 hours to import raw dataset into mysql relational database.

B. Interpret retail data into customer value

To process raw data to generate RFMScoreTable is a complex work. At first, *FilterValidCustomer* function is called to filter valid customer because not all customer of 2011 had purchase record in 2012. We get 15211 valid customer after filtering. Then, *CreateRFMTable* function is called to do statistics to generate RFMTable. R-value can be obtained directly from *LPDT* attribute value of customer data. F-value could be got from counting the number of transactions from purchase record data. M-value is acquired by the sum of *PurchaseAmount* value of every purchase record belonging to customer from purchase record data. The partial detail of RFMTable is shown in Table 2. Next is to call *CreateRFMDaysTable* function to calculate the interval between the time that the latest consuming behavior happens and 2012/9/13. At last, defining the scaling of R-F-M attributes is processed by *CreateRFMScoreTable* function which partitions the three R-F-M attributes respectively into 5 equal parts. It took about 25 hours to generate RFMScoreTable. The partial description of RFMScoreTable is depicted in Table 3.

According to the quantitative value of R-F-M attributes for each valid customer in RFMScoreTable, partition customer into 5 clusters using K-means for clustering customer value. The raw retail data is interpreted into customer value and customer segmentation by *CustomerValueKmeansClustering* function and K-means algorithm in algorithm pool. Table 4 shows the clustering results. It took averagely 1697ms to cluster customer value and segmentation.

C. Asssimilate customer value into knowledge repository

WriteToGraphKR function is called to absorb the customer value and customer segmentation into graph knowledge repository based on Neo4j graph database. In this case study, the conceptual storage model of customer and customer segmentation in graph knowledge repository is shown in Fig. 8. It took averagely 14679ms to absorb R-F-M attribute value, customer value and customer segmentation into graph knowledge repository.

D. Deduce answer to retail quries

The forward chaining and backward chaining *ReasoningQueryFromGKR* Rules for deducing answer to customer value knowledge queries can address the mentioned three retail queries. Fig. 9 shows the backward chaining rule to deduce answers to retail question "What is the R-value of customer X from graph knowledge repository". Table 5 illustrates the average reasoning time to deduce the above three retail queries from graph knowledge repository based on Neo4j graph database with the comparison from relational database [10] [11].

TABLE II. PART INSTANCES IN RFMTABLE

CID	R-value	F-value	M-value
201381	2012-08-30 18:30:00	28	108684
209370	2012-01-01 00:00:00	1	9524
201450	2012-04-06 00:42:00	21	151376
201641	2012-01-01 00:00:00	1	9524
106721	2012-09-13 00:47:00	629	9895961
100292	2012-09-08 00:08:00	239	3686797
…	…	…	…

TABLE III. PART INSTANCES IN RFMSCORETABLE

CID	R-score	F-score	M-score
108268	5	5	5
108298	5	3	3
101944	5	2	4
99373	5	1	1
120893	4	4	4
37985	3	4	5
114511	3	4	5
192452	1	5	5
…	…	…	…

TABLE IV. CLUSTER RESULTS BY K-MEANS WITH 5 CLASSES

ID	R center	F center	M center	Customer Value	Customer Level	Count
0	4.3186	4.8039	4.7588	8.0233	Very High	3346
3	3.8026	3.5376	3.4194	6.2183	High	2725
2	1.4474	3.6541	3.8488	5.5009	Medium	2414
1	3.8273	1.7528	1.6985	4.5373	Low	3058
5	1.5350	1.5664	1.6142	2.7221	Very Low	3669

TABLE V. QUERY TIME COMPARISON

Query Question	RDB+Rule	GraphDB+Rule	Number of Results
(1)	1251ms	47ms	1
(2)	1217ms	24ms	1
(3)	1775ms	181ms	3346

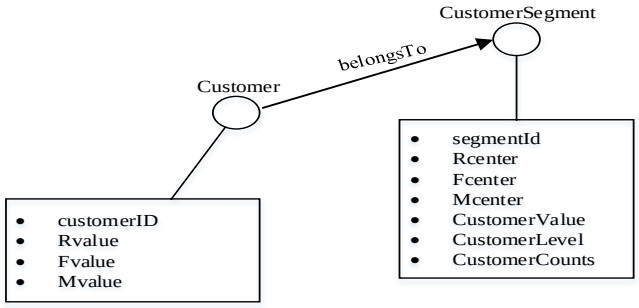

Fig. 8 Conceptual storage model of customer and customer segmentation

```
...
(do-backward-chaining RFMSegment)
(do-backward-chaining Customer)
(defrule customerValueCheck
   (do-customerValue-check (cId ?cid))
   (customerValue (CID ?cid) (CV ?cv))
   =>
   (printout t "CustomerValue of " ?cid " is " ?cv))
(deftemplate customerSegment (slot customerId) (slot segmentId))
(defrule find-customerValue
   (do-customerValue-check (cId ?cid))
   (customerSegment (customerId ?cid) (segmentId ?sid))
   =>
   (bind ?r (call BackwardAssistor getCustomerValue ?sid) )
   (assert (customerValue (CID ?cid) (CV ?r))))
(defrule find-segmentId
   (do-customerValue-check (cId ?cid))
   =>
   (bind ?segmentIdResult (call BackwardAssistor getSID ?cid))
   (assert (customerSegment (customerId ?cid) (segmentId ?segmentIdResult))))
(assert (do-customerValue-check (cId 108268)))
```

Fig. 9 A part of rules used in backward chaining inference

IX. CONCLUSION AND REMARK

Our work is a pragmatic application of conceptual KID model to retail data and verifies the feasibility of KID model. We also proposed a semantic graph based knowledge representation structure for retail business. We also use rule and inference engine to deduce answers to retail queries automatically. In our customer value case study, the experiments show that the use of rule-based engine to deduce answers to retail queries from graph knowledge repository has a better time performance than deducing answers to retail queries from relational database because of its time-consuming joint operation.. Our work also help retailer to have a better understanding on customer purchase behavior.

However, we did not validate the accuracy of customer value assessment result owing to limited retail data. Since limited study and research time, we don't build a complete graph knowledge repository for our retail data. It also should be pointed out that the proposed enhanced KStore data structure is not used to store retail data in the customer value assessment case study. But KStore data structure can record all possible contextual relationships in business data, has a good data compression and can grow as time going. It will be used in some resource-limited device to store data and do data analytics. We also extend our work to other domain not only retail domain.

ACKNOWLEDGMENT

I would like to express sincere gratitude to my research advisor, Prof. Huang, for providing inspiration, guidance, and support during this research and preparation of this thesis, and for teaching me how to be a good researcher. I also give thanks to Prof. Koike and Prof. Hidaka for their careful paper review and kind guidance in graduation thesis writing. I also extend my thanks to lab room members for their constructive suggestion.

REFERENCES

[1] A. Sato and R. Huang, "From Data to Knowledge: a Cognitive Approach to Retail Business Intelligence," IEEE International Conference on Data Science and Data Intensive Systems, pp. 201-217, December, 2015.

[2] E. Siegel, "Predictive Analytics Delivers Value Across Business Applications," B-Eye Networks Business Intelligence and Data Warehousing Resources, 2009.

[3] M. F. Zarandi, "A retail ontology: formal semantics and efficient implementation," University of Toronto, master thesis, 2007.

[4] J. Li and R. Huang, "A Rule-based Knowledge Discovery Engine Embedded Semantic Graph Knowledge Repository for Retail Business," International Conference on Advanced Cloud and Big Data(CDB 2016), Chengdu, China, 13-16 August, 2016.

[5] L. M. Fernández, "Overview of methodologies for building ontologies," Proceedings of the IJCAI Workshop, Madrid, Spain, 1999.

[6] L. Robinson, J. Webber and E. Eifrem, "Graph databases," O'Reilly Media, Inc., 2013.

[7] E. Friedman, "Jess in action: rule-based systems in Java," Manning Publications Co., 2003.

[8] J. Li and R. Huang, "Enhanced KStore with the Use of Dictionary and Trie for Retail Business data," IEEE International Conference on Big Data Analysis(ICBDA 2016), Hangzhou, China, 12-14 March, 2016.

[9] J. C. Mazzagatti, "KStore: A Dynamic Meta-Knowledge Repository for Intelligent BI," IJIIT, vol. 5, pp. 68-80, 2009.

[10] J. Bak, C. Jedrzejek and M. Falkowski, "Usage of the Jess engine, rules and ontology to query a relational database," Rule Interchange and Applications, Springer Berlin Heidelberg, pp. 216-230, 2009.

[11] J. Bak, M. Falkowski and C. Jedrzejek, "The SDL Library: Querying a Relational Database with an Ontolgoy, Rules and the Jess Engine," RuleML201@BRF Challenge, 2011.

[12] M. Baier, K. Ruf and G. Chakraborty, "Contemporary database marketing: concepts and applications," Evanston: Racom Communications, 2002.

[13] J. Han and M. Kamber, "Data mining: Concepts and techniques," San Francisco: Morgan Kaufmann Publishers, 2001.