Elicitation of Relation Schema with Primary Key Attribute from Natural Language Text

S. Geetha
Research Scholar, Dept. of CSE
JNT University, Hyderabad, India

G.S. Anandha Mala
Professor, Department of CSE
St.Joseph's College of Engineering, Chennai, India

ABSTRACT

In natural language, a word or phrase represents the concept. This paper presents an idea of extracting the key attributes of the object. The construction of the logical structure of a relation emerges from the unstructured text. It overcomes the limitation of extracting the structured information. This approach is based on extracting the key information from the scattered unstructured text. This process starts with splitting the sentence, tagging the individual words in the input document by using Parts of Speech (PoS). PoS categorize the input data into Object Oriented Elements (OOE) which includes entities, attributes, actions and builds the relations among these entities and actions. The proposed approach identifies the structure of the relation which is extracted from the Software Requirement Specification (SRS) to the user and constructs a schema of the relation by identifying primary key attributes based on adjectives and by applying mapping rules.

Keywords

Structured Data, Entity, Mapping rules, Key Attribute, Relation Schema

1. INTRODUCTION

The unstructured data is a classical source of information for obtaining the extraction model. The unstructured data refers to any data that has no identifiable structure. The text document hides the valuable structured data. SRS specifies the complete description of a system and lists all necessary requirements which gives thorough understanding of user needs-input data in terms of requirements. The input data is the unstructured information and output is the structured data. The automation of the process is achieved by creating the schema of the relation. A relation schema describes the basic information about a table or relation. It is the logical information of the relation. This includes a set of attributes, the data types associated with each attribute. Tables might also have indexes on them to looking up values on assured columns. Attributes can be classified as identifiers. Identifiers are commonly called as keys or key attributes that uniquely identifies an instance of an entity. Identifier describes a non unique characteristic of an entity instance. An entity has an attribute whose values are distinct for each individual entity. Such an attribute is called as key attribute. Key attributes in a table are of strong significance in relation schema design process.

Relation schema defines the meta data elements which represent the particular domain which are used to describe structures and constraints of data representation of an entity. When identifying attributes of entities, identifying key attribute is very essential. By using key attribute, each record in a relation can be easily identified. There are number of issues to be addressed in designing the schema of a relation specifically to identify the key attribute identifier as a key attribute. This is another approach that brings automatic identification of the primary key attribute depending on the list of attributes of an entity identified from SRS. A relation key can have only one primary key. Primary keys enforce uniqueness on the column(s) of the relation. Primary key emphasizes entity integrity by uniquely identifying entity instances. Once the key attributes have been identified, it is easy to design the schema of the relation. The key attribute makes the relation row unique and ensures that no duplicate records exist.

The paper starts with the review of the requirement specification followed by the natural language processing techniques which are applied and discussed in section 2. The proposed system is explained in section3, implementation and results in section 4. The conclusion and future work in section 5.

2. RELATED WORK

Natural Language Processing (NLP) deals with determining, accepting and generating the languages that humans use naturally to interact with computers. By applying the database queries for extracting the information on text corpus [7], Natural language system uses the knowledge about the language structure which comprises words, grouping the words to form sentences, word meaning to match with sentence meaning, etc. [3]. There are several user services based on keyword search and structured querying [10]. Information Extraction (IE) identifies and extracts the information about particular Class of events or relationships in a natural language text and converts into a structured representation. IE system needs either hand written rules or training data to extract information [1, 2]. It is quite successful in accumulating databases from unstructured text, but they have been applied in situations where the schema is known. The KnowItAll system [4] needs manually written extraction rules for each relation and many systems have tried hand chosen training samples [5].

TGen, an algorithm is another assuring start for a schema discovery system [6] which constructs the relation schema for extracted data. The essential features of the object oriented system namely the classes as entities, attributes, methods and relationships between the methods among the classes are analyzed by mapping the PoS tagged words of the natural language text into object oriented modeling elements using mapping rules [8]. The previous works have concentrated less on identifying schema for the relation with primary key attribute. The proposed work considers the attributes of the classes to identify the primary key and assign data type to all the attributes to construct the schema for the relation by analyzing SRS.

3. THE PROPOSED SYSTEM

A relation is a set of data values is organized into rows and columns. Row values are called as records and column values are called as attributes. Each row is identified uniquely with the key attribute. The primary key is the identity of the entity. A primary key is an attribute that allows information for an entity to be uniquely identified.

The Key attribute of the relation allows database user to locate the specific information which is extracted from unstructured text. The proposed methodology shows automatic identification of primary key attribute of the relations by using handcrafted rules and machine learning techniques. The processes are shown in figure 1.

3.1 Domain Knowledge Extractor

Domain Knowledge Extractor takes as input, a natural language text describing the user needs. It identifies the OOE namely classes, attributes, methods and relationships and it performs this activity using the following steps.

3.1.1 Sentence Splitting

The problem statement SRS is split into sentences to reduce the processing overheads.

3.1.2 Tagging and Phrase Chunking

The PoS tag designates each word of all sentences and classifies the words as nouns, verbs, adjectives, etc. QTAG is a probabilistic PoS tagger which is used for this purpose[9]. The OOE namely classes, attributes, methods and relationships have to be chosen depending upon their tags. The noun and verb phrases are recognized based on simple grammar.

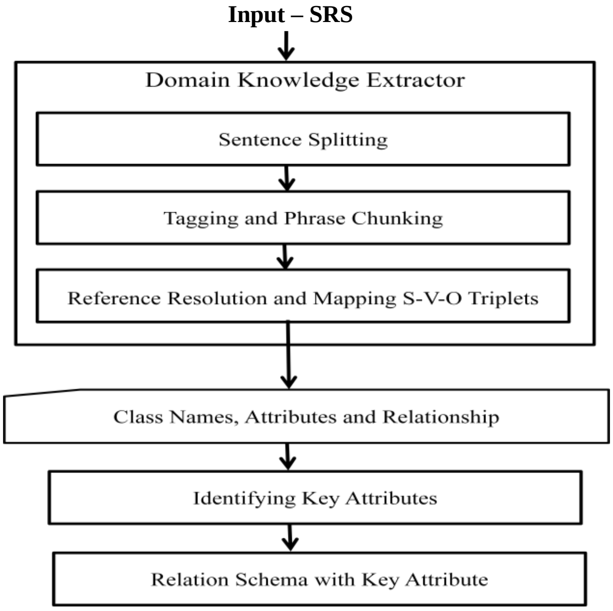

Input – SRS
Domain Knowledge Extractor
Sentence Splitting
Tagging and Phrase Chunking
Reference Resolution and Mapping S-V-O Triplets
Class Names, Attributes and Relationship
Identifying Key Attributes
Relation Schema with Key Attribute

Figure 1. Architectural Design of Relation Schema Extractor

3.1.3 Reference Resolution and Mapping S-V-O (Subject-Verb-Object) Triplets

In a sentence, the subjects and objects can sometimes be pronouns. In that case, it has to be resolved to their respective noun phrase. The text has to be interpreted into the S-V-O to map the words into OOE based on rules.

3.2 Class names, Attributes, Methods and Relationship Identification

To identify the OOE, simple rule based approaches are followed to perform the transforming and matching process. Nouns are considered to be classes and verbs are considered to be methods of the classes. First noun is considered as class and second noun is considered as attribute based on verb phrase. Transforming S–V–O structure to subject and object are classes sharing verb as candidate method. The relationship is an association between two OOEs. After creating a standard Noun Verb Noun structure and mapping this information onto the classes, it is used to identify the relationship with each other.

3.3 Identification of Primary Key Attribute

The primary key of the table uniquely identifies each record in the table. The table is defined with set of tuples (d1, d2… dn) where each element “d” is a member of a data domain “D”. Each Domain will have a set of attributes (a1, a2…an) which is called a relation schema. To identify as a primary key for an entity, an attribute must have

- It must have a non null value for each instance of the entity.
- The value must be unique for each instance.
- The values must not change.

Each column represents attributes of the object modeled by a relation. Primary key attribute has to be selected out of n attributes to maintain the uniqueness of the relation. Each row represents an individual occurrence of the data. A training data set is used to identify the key attribute for specific domain with the predefined set of letters and words. The primary key attribute is identified based on the rules listed below.

- If the sentence is in the form of “Subject+ Possessive verb + Adjective + Object”, then the object is a key attribute.
- When the sentence is in the form of “Subject+ Possessive verb + Object”, then the object is prefixed or suffixed with set of predefined letters.
- When the sentence is in the form of “Subject+ Possessive verb + Object”, then the object is from the set of predefined words.

3.4 Relation schema with Key Attribute

A relation schema describes the basic information about a table or relation. This includes a set of column names and each column is associated with numeric or alphanumeric data type. The key attribute of the relation identifies unique record in the relation. A data type is applied to every attribute in a logical model. The data type is determined by the domain from which the attributes derives its properties. Every attribute of the relation is assigned with data type that contains a specific type or range of values. In the initial stage, the attributes of the relation schema are considered as either numeric or non numeric. After identifying the key attribute, data type of each column of the relation has to be identified with set of rules.

- Assign the data type of the key attribute to number.

- Assign the data type of the non key attributes to character.

4. IMPLEMENTATION AND RESULTS

In this Section the results obtained from experiments performed based on the following sample.

Sample Input:

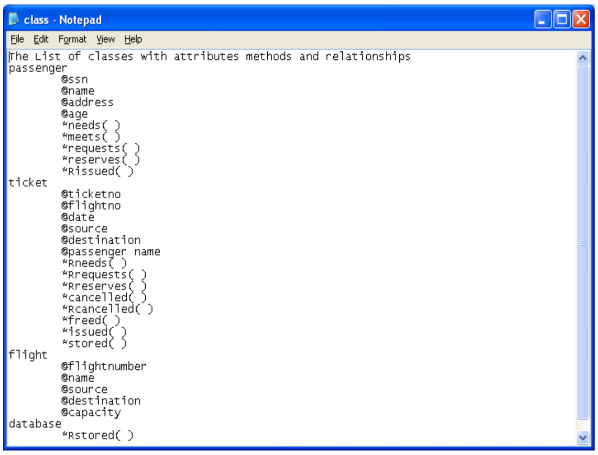
The passenger has the unique ssn, name, address and age. When the passenger needs a ticket and meets the receptionist. The receptionist gets passenger trip details. The receptionist checks for the availability of flight and ticket which is requested by the passenger. The flight has the flightnumber, name, source, destination and capacity. If the receptionist finds a ticket available on the flight then the receptionist issues the ticket to the passenger. The ticket has ticketno, flightno, date, time, source, destination and passenger name. The receptionist blocks the ticket. If the passenger requests a ticket and if there is no availability of ticket, the passenger reserves the ticket. The ticket can also be cancelled. If it is cancelled, the ticket is freed up and is available for booking. If the ticket is issued to the passenger, the ticket is stored in the database.

Here the various stages of processing. The process starts with

Step 1. Domain knowledge extractor : It takes as input, a natural text describing requirements for a system. The processing starts with sentence splitting, tagging, phrase chunking, pronoun resolving and finally interpreting the text into the S-V-O pattern.

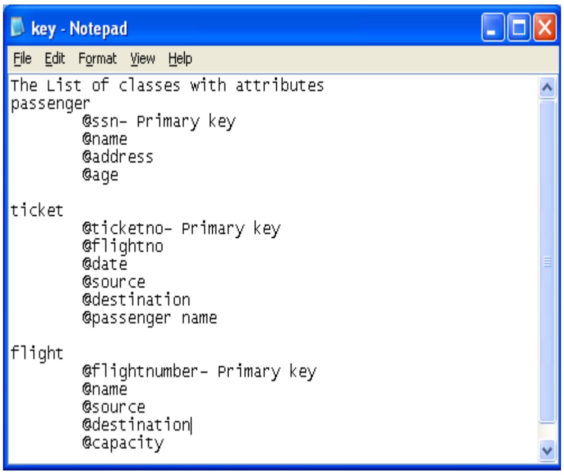
Step 2: Class Names, Attributes, Methods and Relationships: Rules are used to identify OOE as attributes and methods and relationships .The List of classes with attributes, methods and relationships are shown in figure 2.

Step 3. Identifying Primary Key Attributes: The primary key is an attribute that uniquely identifies a specific instance of an entity. The List of classes with Specifying Primary key attributes are shown in figure 3.



```
The List of classes with attributes methods and relationships
passenger
        @ssn
        @name
        @address
        @age
        *needs( )
        *meets( )
        *requests( )
        *reserves( )
        *Rissued( )
ticket
        @ticketno
        @flightno
        @date
        @source
        @destination
        @passenger name
        *Rneeds( )
        *Rrequests( )
        *Rreserves( )
        *cancelled( )
        *Rcancelled( )
        *freed( )
        *issued( )
        *stored( )
flight
        @flightnumber
        @name
        @source
        @destination
        @capacity
database
        *Rstored( )
```

Figure 2. List of classes with their attributes, methods and relationships



```
The List of classes with attributes
passenger
        @ssn- Primary key
        @name
        @address
        @age

ticket
        @ticketno- Primary key
        @flightno
        @date
        @source
        @destination
        @passenger name

flight
        @flightnumber- Primary key
        @name
        @source
        @destination
        @capacity
```

Figure 3. List of classes with key attributes

Step 4. Relation schema with key attribute: The list of classes with column names and their data types are shown in table 1.

Table 1. Relation schema

Passenger			Ticket			Flight		
Attribute list	**Primary key**	**Data type**	**Attribute list**	**Primary key**	**Data type**	**Attribute list**	**Primary key**	**Data type**
Ssn	✓	number	ticketno	✓	number	flightnumber	✓	Number
Name	X	string	flightno	X	string	name	X	String
Address	X	string	date	X	string	source	X	String
Age	X	string	source	X	string	destination	X	String
			destination	X	string	capacity	X	String
			passenger name	X	string			

5. CONCLUSION AND FUTURE WORK

This paper presents an idea to extract the structured data from the unstructured text. Natural processing techniques and set of rules are used to extract domain knowledge from the requirement specification. By applying a set of rules, the system can easily identify entities, attributes and the primary key attribute from the list of attributes for different entities. It gives information about tables, columns, domains and constraints. By using the concept of machine learning techniques and handcrafted rules, the schema of the relations are identified. A relation schema is described with a set of attributes and data types associated with it.

Further work can be extended for identifying the relationship among different entities. By using foreign key the database schema can be constructed by associating a column in one table to a column or set of columns in another table. The system can be extended by automating the construction of database schema and it gives information about all the relations and association among the attributes of all the relations.

6. REFERENCES

1. E.Agichtein and L.Gravano. 2000. “Snowball : Extracting relations from large plain-text collections”. In Proc. of the 5th ACM International Conference on Digital Libraries, ACM Digital Library.

2. I.Mansuri and S.Sarawagi. 2006. “A system for integrating unstructured data into relational databases”. In Proc. of the 22nd IEEE International Conference on Data Engineering (ICDE).

3. Mirzanur Rahman, Sufal Das and Utpal Sharma. 2009. “Parsing of part-of-speech tagged Assamese Texts”, IJCSI International Journal of Computer Science Issues, Vol. 6, No. 1.

4. O.Etzioni, M.Cafarella, D.Downey, S.Kok, A.Popescu, T.Shaked, S.Soderland, D.Weld and A.Yates. 2004. “Web-Scale Information Extraction in KnowItAll”. In Proc. of the 13th International Conference on WWW, ACM Digital Library, pages 100–110.

5. S. Brin. 1998. “ Extracting Patterns and Relations from the World Wide Web”. In WebDB Workshop at 6th International Conference on Extending Database Technology, pages 172–183.

6. Michael J. Cafarella,Dan Suciu, Oren Etzioni. 2007.

“ Navigating Extracted Data with Schema Discovery”. In Proc. of the 10th International Workshop on Web and Databases.

7. Luis Tari, Phan Huy Tu, Jorg Hakenberg, Yi Chen, Tran Cao Son, Graciela Gonzalez and Chitta Baral**.** 2012. “Incremental Information Extraction Using Relational Databases”, IEEE Transactions on Data & Knowledge Engineering, pages 86-99, Vol. 24, No. 1.

8. G.S. Anandha Mala, J. Jayaradika and G.V. Uma. 2005. “Restructuring Natural Language Text to Elicit Software Requirements”. In proc. of the International Conference on Cognition and Recognition.

9. Lee, S. E., & Han, S. S. K. 2011. “Qtag : Introducing the qualitative tagging system. In Proc. of the 18th conference on Hypertext and hypermedia”, pages 35–36. ACM Retrieved.

10. A. Doan, J. F. Naughton, R. Ramakrishnan, A. Baid, X. Chai, F. Chen, T. Chen, E. Chu, P. DeRose, B. Gao, C. Gokhale, J. Huang,W. Shen, and B.Q. Vuong. 2008. “Information extraction challenges in managing unstructured data”, *SIGMOD Rec.*, pages 14 – 20,Vol. 37, No. 4.