Detection Approaches for Table Semantics in Text

Saleh Alrashed and W.A. Gray

Department of Computer Science, Cardiff University,
Cardiff, Wales, UK.
{scmsa1,wag}@cs.cf.ac.uk

Abstract. When linking information presented in documents as tables with data held in databases, it is important to determine as much information about the table and its content. Such an integrated use of Web-based data requires information about its organization and meaning i.e. the semantics of the table. This paper describes approaches that can be used to detect and extract the semantics for a table held in the text. Our objective is to detect and extract table semantics that are buried in the text. For this goal to be achieved, a domain ontology that covers the semantics of the term used in this table must be available to the information system. The overall aim is to link this tabular information in an interoperable environment containing database and other structures information. ...

1 Introduction

The amount of online structured, semi-structured and unstructured data is growing rapidly. The reasons for this are the growth of e-commerce, e-services, e-government and e-library, areas which publish a very large amount of data on the internet in tabular form. Web pages can be designed as static html pages or as dynamic web page. A dynamic web page gives its user the ability to query an associated database, with the result being represented as either an html table or an XML document. Joining these tables from different sources will leads to semantic conflicts. The information needed to detect semantic conflicts when combining tables is often buried deep within the text of the documents associated with the table or in the web site itself . In order to resolve problems caused by semantic conflict, an information system must be able to ensure semantic interoperability by discovering and utilizing this contextual information. The goal of this research is to enable the discovery and effective use of context information. The context of a piece of data in a table is the metadata relating to its meaning.

D. Lopresti, J. Hu, and R. Kashi (Eds.): DAS 2002, LNCS 2423, pp. 287–290, 2002.

2 Semantics and Representational Conflicts

When trying to integrate data from different sources drawn from the same domain, semantic conflicts and representational conflicts can occur. Semantic conflicts are related to differences in the metadata of the table (attributes, names etc). Such a conflict occurs when different data names represent the same data (synonyms), the same name represents a different data domain (homonyms), or a hidden semantic relationship exists between two or more terminologies. For example the relationship between cost and price or between profit and net-profit, and the similarity between car, vehicle and truck cannot be understood unless we use a domain knowledge base to relate these terms. To overcome conflicts we use semantic metadata, which provides information about the meaning of the available data and any semantic relationships . The sources of semantic information can include ontologies.

Representational conflicts occur due to the way data is represented and in particular the measurement unit being used. These conflicts are concerned with the values of an attribute. If we have two attributes, which are semantically the same, they are not in conflict when their values are represented in the same way, i.e. in the same units. This is important, if we are bringing attributes together. Thus representational metadata provides information about the meaning of the values of an attribute, its representational relationships and units of representation.

3 Semantic and Representation Detection Framework

SRD (Semantic and Representation Detection framework) is a proposed system for discovering and interpreting the context information about tables present in a document containing tabular data. (Fig. 1) shows the proposed system architecture of the SRD system, which will extract and structure the context data about a table held within a textual document.

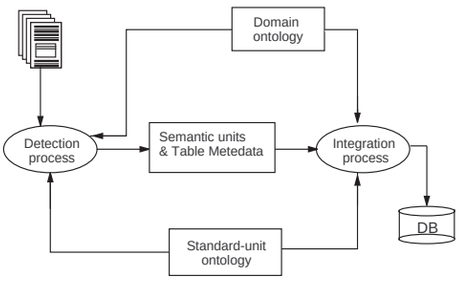

Fig. 1. Semantic and representation detection framework

It consist of five units; namely: detection process – analyses the content of the document and detects the beneficial context information about tables present

in a document containing tabular data; domain ontology - provides information about the representation and description of the data on the basis of the model and details about how to convert between different representations; standard-unit ontology – used in discovering some of the value representations in the text; semantic units and table metadata – represent extracted semantics and representation information which is stored for further processing; integration unit – integrates semantic units and table metadata with the corresponding data from a database. In this paper we concentrate on the detection process subunit.

3.1 Detection Process

This process analyses the content of a document and detects the useful context information about tables present in a document containing tabular data. We divide this process into three parts; context detection, context extraction, and context representation.

Context detection. There are a number of approaches that can be used to analyze a document which contains tabular data to detect contextual information about the table.
A. Text searching:Use the table headings as keywords to search for related semantics in the adjacent paragraphs to the table. We know that the table header or table metadata indicates the main concepts that the table represents, therefore concentrating on the table metadata insures the correctness and accuracy of the detected contexts.

B. Augmenting header using an ontology: In many situations the table header alone is not enough to describe the semantics of that table. Another approach is then used, namely: the table header is used to extract the corresponding concept from the domain ontology and search for that concept and all related synonyms in the text.
C. Table data: In some cases, the author explains some of the table data immediately after or before the table itself. Therefore searching the surrounding paragraphs might lead us to some of the table semantics.
D. Table title: Another approach is to take the table title and search for the phrase "figure x" in the text. We can also use the table title which informs that this table is about car prices which leads us to an understanding that the field name "value" is equivalent to "price". We can then search for both value and price in the adjacent document context.

Context extraction. After detecting the corresponding context in a document, these contexts will be extracted and kept for farther processing.These contexts will have no meaning until they have been mapped to a corresponding concepts in the domain ontology.

Extraction algorithm. After selecting the approach to take, the first thing is to decide which part of the document to search. Logically the author is going to describe the table either before or after. We can also search for the table

figure number in other paragraphs. After identifying the section to be searched in detail, a search will start looking for the column headers and their synonyms until it finds matching words followed by a mathematical operator (+, *, = ...) or (is, are ...). If no mathematical operators found than this sentence will be stored for further assessment by the user.

Representing the extracted contexts. After detecting and extracting the corresponding contexts, our internal model represents them in two ways; as semantic units or as table metadata. This can then be used to link corresponding information in different tables together.

3.2 Semantic Units

A semantic unit represents a data item together with its underlying semantic context. This consists of a flexible set of meta-attributes that describe the meaning of the data item. However, because we cannot describe all modeling assumptions the semantic context always has to be recognized as being a partial representation. In addition, each semantic unit has a concept label associated with it that specifies the relationship between the unit and the real world aspects it describes. These labels have to be taken from the ontology. Our Semantic Unit represents a value v as: "(C , V , ST)" where C represents the knowledge concept derived from the domain ontology which represents the corresponding value, V represents the value, and ST represents the semantic contexts that have been discovered in the text.

3.3 Table Metadata

The table metadata describes the full table with its corresponding contexts. This metadata is enhanced with any semantic context found in the text.

4 Related Work

Christof Bornhovd proposed a representation model for explicit description of implicitly described semi-structured data, and used this model for the integration of heterogeneous data sources from the web. In this model they don't distinguish between semantic and representational contexts and don't support the conversion of data between different contexts. Thus our proposal extends their work by supporting the conversion between different representational values.

5 Conclusion

In this paper we have described the architecture of a system for the discovery and resolution of semantic and representational conflicts and have shown the differences between them and when these occur. Also we have described the different approaches that can be used to detect the table semantics by using the domain ontology. After detection and extraction of the corresponding contexts, our internal model represents them in two ways; as semantic units or as table metadata. This can then be used to link corresponding information in different tables together.