A Framework for an Ontology-Based Data-Gleaning and Agent Based Intelligent Decision Support PPDM System Employing Generalization Technique for Health Care

K. Murugesan
Department of Computer Science and Engineering
Anna University
Chennai,Tamilnadu
India

J.Indumathi
Department of Computer Science and Engineering
Anna University
Chennai,Tamilnadu
India

D.Manjula
Department of Computer Science and Engineering
Anna University
Chennai,Tamilnadu
India

Abstract----The obligatory to anticipate the privacy benefits of heavy downpour of monsoon rain from the firmament clouds of Privacy Preserving Data Mining (PPDM) Techniques have recently grown leaps and bounds. The desiccated users & miners look for the petite clemency from these little heavens in the form of a framework with PPDM Techniques. In this paper we have developed two things namely, the ontology based data gleaning system, the gleaned data is sent to a PPDM system which has an in-built generalization privacy technique and the agent-based intelligent decision support system. The primary report is on the implementation of existing generalized framework with alternate technology (i.e. implementation using Natural language processing instead of heuristic based method). Our Data Gleaning system will also allow new algorithms and ideas to be incorporated into a data extraction system. Extraction of information from semi-structured or unstructured documents is a useful yet complex task. Ontologies can achieve a high degree of accuracy in data extraction system while maintaining resiliency in the face of document changes. Ontologies do not, however, diminish the complexity of a data-extraction system. As research in the field progress, the need for a modular data-extraction system that decouples the associated processes continues to grow.

We also propose a generalization conceptual framework in this paper, where we guide the extracted data from the data Gleaning system to the generalization framework. The QI Generalization technique in the generalized framework is used to visor sensitive information and then publishes the privacy preserved data for knowledge discovery.

Keywords—Ontology; Data extraction; generalization; privacy preservation.

I. INTRODUCTION

Making sense of vast amount of information available on the World Wide Web has become an increasingly important and lucrative endeavor. The success of web search companies such as Google demonstrates the vitality of this industry. Yet the web search has much still to deliver. While traditional search engines can locate and retrieve documents of interest, they lack the capacity to make sense of the information those documents contain [1].

II. RELATED WORK

Data extraction addresses many of the problems associated with typical web searches based on standard information retrieval techniques. Data extraction is the activity of locating values of interest within electronic textual document, and mapping those values to a target conceptual schema (Laender .et al., 2002)[10] .The conceptual schema may be as simple slots in a template (a wrapper) used to locate relevant data within a web page, or it may be as complex as a large domain ontology that defines hierarchies of concepts and intricate relationships between those concepts. The conceptual schema is usually linked to a storage structure such as an XML file or a physical database model to permit users to query the extracted data. In this way the meaning of a document is detected, captured, and made available to user queries or independent software programs. Much of the research in data extraction has aimed at developing more accurate wrappers

Much of the research in data extraction has aimed at developing more accurate wrappers while requiring less human intervention in the process [7,8,4,2]. The primary drawback of wrappers, whether they are generated manually or semi automatically, is that they depend on the particular syntax of the mark up to detect boundaries between relevant and irrelevant data. The main implication is that when a site's mark up changes, the corresponding wrappers often break.

The generalized framework [1] has focused on the use of richer and more formal conceptual schemas (ontologies) to improve accuracy in data extraction. This system modifies the generalized framework with use of NLP in order to improve the performance. Because ontology describes a subject domain rather than a document, ontology-based data-extraction systems are resilient to changes in how source documents are formatted, and they can handle documents from various sources without impairing the accuracy of the data extraction.

This system will accepts multi-record HTML documents, determines record boundaries within those documents, and extracts the data from each record. It generates SQL DDL statements for the model structure and stores the extracted information as DML statements. This facilitates the querying of the results but also removes certain metadata attached to the data during the extraction process.

According to the restraints forced on generalization, privacy algorithms can be alienated into two groups. The first category developed by Samarati in 2001[13] employs "full-domain generalization", which presumes a chain of command on each QI attribute, and entails that all the divisions in a general domain should be at the same level of the hierarchy. For example, the value *sinusitis* is generalized to *respiratory-infection*. Such a constraint is espoused by the binary search algorithm 13], the exhaustive search method [15] and the Apriori-like dynamic programming approach [11] all of which minimize information loss based on various metrics. The second category (i.e., "full-subtree recoding" as termed in [11]) drops the same-level obligation mentioned earlier, since it often leads to pointless information loss. Following this idea, Iyengar 2002[9] develops a genetic algorithm, whereas greedy algorithms are proposed in [6,16], based on top-down and bottom-up generalization, respectively. These approaches, however, do not minimize information loss. Bayardo et al., 2005[3] remedy the problem with the power-set search strategy. Our work also belongs to this category, but significantly extends it to incorporate customized privacy needs. Say for example, in practice, the recipients of the published data are often specialized users (e.g. scientists), who may explicitly stipulate the analytical tasks (such as association rule mining [14] required. This information may be utilized to release a table that is highly effective for those tasks, without breaching the privacy constraints formulated by data owners. We are also proposing a new hybrid system which combines the PPDM system and the agent based intelligent decision support systems. These systems coalesce the compensations of both systems namely, the privacy of the PPDM and the intelligent support of the agent based decision support systems.

III. OSMX ONTOLOGY CONSTRUCTIONS

For ontology construction our system relies on OSMX ontology editor [12]. Ontology editor is a predominantly WYSIWYG tool. OSMX is the ontology language which is derived from the existing OSML (Object oriented System Model Language) [12]plus XML schema.

We use the Java Architecture for XML binding (JAXB) to generate java classes and interfaces that represent OSMX constructs from the OSMX specification. Modifying the OSMX definition is generally a straightforward process: we adjust the definition in the XML schema document, and then execute a JAXB program that rebuilds the classes and interfaces automatically.

OSMX's ORM (Object relationship Model) with its data frame can serve as an Ontology language. In ORM concepts are represented by object sets, which group values (objects) that have similar characteristics; and connections between concepts are expressed via relationship sets, which group object tuples (relationships) that share common structure. Generalization-specialization is a special type of relation that expresses "is-a" relationships between object sets.

Data frame describe the characteristics of objects, similar to an abstract data type[5]. Data frames are attached to object sets, and provide a means to recognize lexical values that correspond to objects in the ontology. The following diagram explains the structure of data frames.

ObjectSet

|

*--DataFrame (0 or 1)

|--defaultCanonicalizationMethod (attribute)

|--InternalRepresentation <TypeSpecification> (exactly 1)

|--ValuePhrase (0 or more)

| |--hint (attribute)

| |--confidenceTag (attribute)

| |--caseSensitive (boolean attribute)

| |--canonicalizationMethod (attribute)

| |--ValueExpression <DataFrameExpression> (0 or 1)

| |--ExceptionExpression <DataFrameExpression> (0 or 1)

| |--LeftContextExpression <DataFrameExpression> (0 or 1)

| |--RightContextExpression <DataFrameExpression> (0 or 1)

| |--RequiredContextExpression <DataFrameExpression> (0 or 1)

| |--SubFromExpression <DataFrameExpression> (0 or 1)

| |--SubToExpression <DataFrameExpression> (0 or 1)

| *--<Keyword Phrase> (0 or more)

|

|--<KeywordPhrase> (0 or more)

*--Method (zero or more)

|--hint (attribute)

|--name (attribute; required)

|--language (attribute; default is 'java')

|--ReturnType <TypeSpecification> (0 or 1)

|--Parameter (0 or more, ordered)

| |--name (attribute; required)

| |--decorator (attribute) <--- e.g., 'OUT', 'BYVAL', 'const', etc.

| |--Type <TypeSpecification> (exactly 1)

| *--<KeywordPhrase> (0 or more)

*--<KeywordPhrase> (0 or more)

TypeSpecification (NOTE: children are mutually exclusive)

|--ObjectSetReference

| |

| *--name (attribute; required) <--- ref to an object set

*--DataType

|

|--typeName (attribute; required)

*--unitOfMeasure (attribute)

KeywordPhrase

|--hint (attribute)

|--confidenceTag (attribute)

|--caseSensitive (attribute)

*--KeywordExpression <DataFrameExpression> (exactly 1)

DataFrameExpression

|--color (attribute) <--- only for graphical tool use

|--ExpressionText (exactly 1)

| |

| *--<element content> <--- contains the extended regular expression

*--MatchedText (0 or more)

|--document (attribute)

|--location (attribute)

|--startPos (attribute)

|--endPos (attribute)

|--status (attribute)

*--<element content> <--- contains a matched text string

Figure 3.1 shows the graphical view of Digital Camera ontology

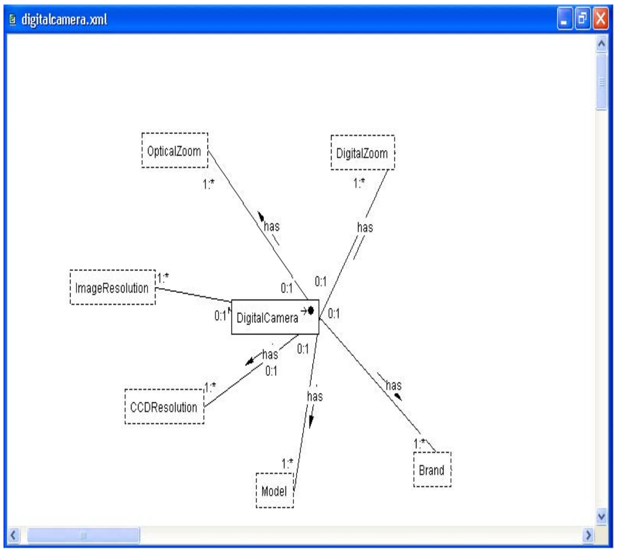

Figure.3.1. Digital Camera application ontology

The most basic element of an extraction rule for a data frame is a matching expression. This an augmented forms of a Perl-5 compatible regular expression. The level of regular expression support is defined by the java regular expression package java.util.regex. We augment regular expression by allowing the rule designer to embed macro and lexicon references within the expression itself.

IV. THE FRAMEWORK ARCHITECTURE

A. Problem Statement

The goals of An Ontology-Based Data-Gleaning And Agent based decision support PPDM System employing Generalization Technique were to design, develop and implement functionalities like privacy preservation, User friendly framework, Reusability, Portability secure protocol for preserving private data's and knowledge. Figure 4.1 provides the overall architecture of the system.

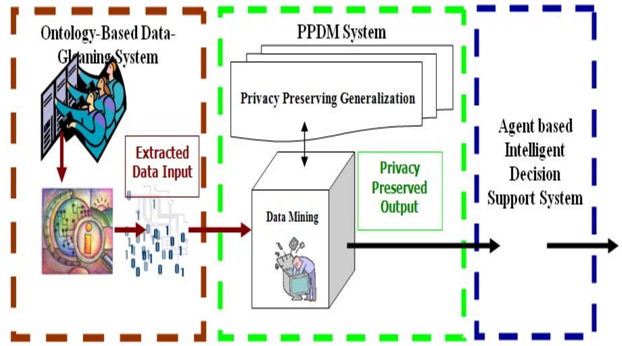

Figure 4.1. An Ontology-Based Data-Gleaning and Agent based Intelligent Decision support PPDM System: High-Level

a. Sub System Architecture of the Data Gleaning System

This section explains the overall architecture and functionality of the system. A graphical overview of the system is described in figure 4.2. The user is expected to initiate the system by giving the input files names.

The first four stages, document retriever, document recognizer, document parser and content filter together called as document pre-processing. The document retriever is responsible for supplying input document to the system. The current system will retrieve the document from the local file system. This accepts the URI as the input and produces a set of documents.

The document recognizer analyzes the input document to determine which available document parser is best suited to decompose the document. This module is needed when we giving the different types of file as inputs. The current system accepts the HTML and plaintext sources.

The document parser breaks up the documents into sub-documents in order to make extraction easier. It divides the multi-record document into sub-documents, each constituting an individual record, allows us to process one record at a time without having to worry about missing a record boundary and extracting values from an adjacent record.

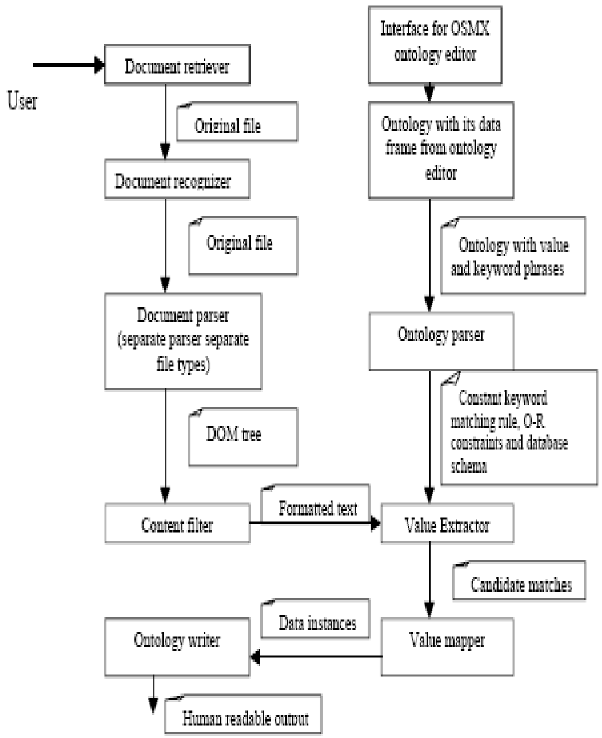

Figure.4.2. Sub System Architecture of the data Gleaning system

Content filter eliminates all tag information from the HTML documents since these tags do lend additional meanings. It is convenient for extraction the elimination of tags.

Ontology parser produces the set of constant-keyword matching rules, object relationship constraints and a set of database schema by parsing the input ontology. The above mentioned processes are only the preliminary process for extraction. The following steps initiate and extract the information from input documents.

The responsibility of value extractor is to apply the value recognition rules produced by the ontology parser to the input document and producing the set of candidate matches. The conflicts among the candidate matches do not resolved in this stage. The one more functionality of value extractor is to maintain the location information for each candidate values. This provides a traceable back to the document content and also can supply the useful data for algorithms that resolve match conflicts and create mappings from candidate values to elements to the ontology.

The most important and difficult part of the extraction system is the process that takes the candidate matches and uses them to build a data instance consistent with the constraints specified by the ontology. The value mapper module does this job. The generalized framework implements the value mapper by using the heuristics approach. This system implements the value module by using the Natural Language Processing.

The important tasks of value mapper module to transform the candidate matches into data instances are (1) resolve conflicting claims that different elements of the ontology make upon the same matched value, (2) transform lexical values into objects (instances of concepts defined in the ontology), (3) infer the existence of objects that have no direct lexical representation in the text, and (4) infer relationships between objects. Finally a collection of objects and relationships between those objects are obtained.

The ontology writer will converts the data instance into suitable for storage. Also it displays the result as a human readable HTML format.

If desired we can guide the output to a Generalization Framework for Privacy Preserving Data Mining.

b. Sub System Architecture of the Privacy Preserving Data Mining System

A overview of the sub system is described in figure 4.3.

Data Sets-These are the datasets extracted using our bespoken framework of data extraction sub system.

User Specification-User personal preference can be easily solicited from an individual through a node in the taxonomy when s/he is supplying his/ her data.

QI Generalization-The system perform QI generalization on QI attribute (the attributes which may reveal a personal identity with the aid of external information)& provide the generalized QIgroup. Initially these generalizations should choose a unique generalization function with single partition for every QI attribute .A generalization function for every QI attribute should be decided by a general domain of that attribute.

Sensitive Generalization- The system perform the sensitive generalization on sensitive attribute of QI group generalized values with the help of user specification &

provide the optimal generalized information. The SA Generalization should choose a different generalization function for each QI group.

Checking Information Losses-Check the total information losses for the current publishable table with the previously discovered table .The output of this process decide the further process, i.e., if the information lossses are mininmum, it will invoke rounding algorithm process, else the current table is best table.

Rounding Process - The software should carry out different level of generalization in order to release various tables for fulfilling wanted data analysis requirements. The software should preserve the information for data researcher & limit the usefulness of unwanted information that may be derived from the release of data. Data analysis requirements are different, so the software should choose different generalization function to release different tables during each round. This software process is the “Rounding Process” which is responsible for the subsequent execution of different level of QI generalization intended to satisfy the analyst requirements and next the Sensitive Generalization process will be invoked. In each round different QI generalization function is chosen without loss of privacy, which functions performing the QI generalization on original QI attribute.

Incremental Data Preservation-The proposed software technique also considers incremental data dissemination, where a dataset is continuously incremented with new data. This technique anonymizing the incremental dataset with considering previously released information, due to this consideration, releasing table with incremental data keep away from inferences like difference attack. QI group that is affected by difference attack is refined with the intention of preserving privacy. This technique gives better solution to difference attack.

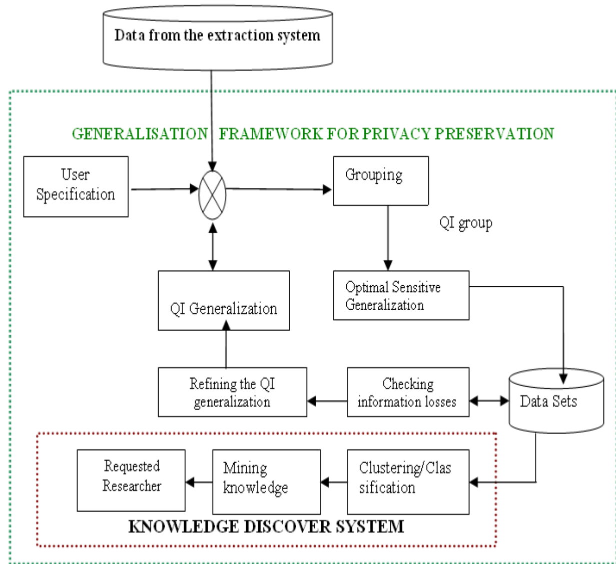

Figure 4.3. Block Diagram of the proposed architecture

c. Sub System Architecture of the Agent-Based Intelligent Decision Support Systems

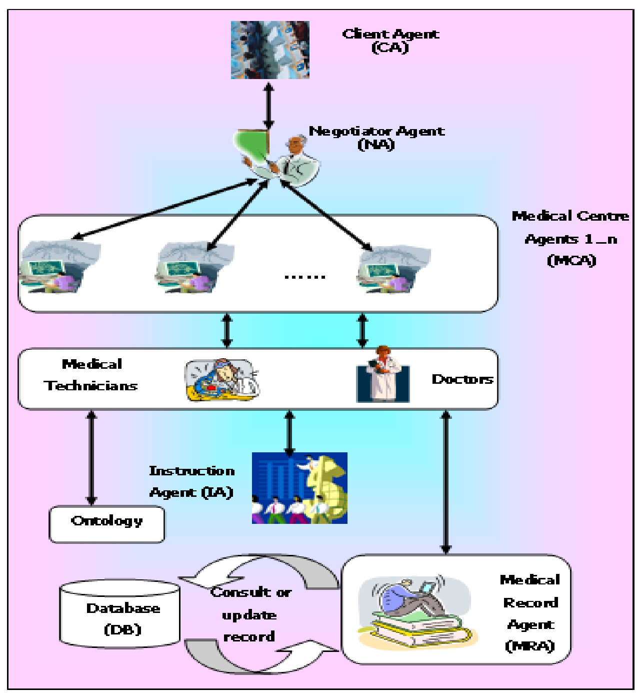

Figure.4.4. Sub System Architecture of the data Gleaning system

At the crest of the design seen in figure 4.4 is positioned the client, who works together with the system through the Client Agent (CA). CA stores static data related to the client (e.g. national healthcare number, name, address, access information, login, password, and keys) and dynamic data (the timetable and the preferences of the user. The Negotiator Agent (NA) is an agent that is on familiar terms with about all the medical centers located in a certain area. A Medical Centre Agent (MCA) centralizes and monitors the outsider's accesses to the agents that administer the information of a medical centre. A MCA monitors all of its departments, and a set of general services linked to human or physical resources (e.g. a blood test service). Each department has a staff of several doctors, and offers more specific services.

In addition, each department contains an Instruction Agent (IA) that performs all actions involved with instructions (e.g. it can retrieve the CPG associated to a specific illness). This IA contains only CPGs related to the department where it is located. At the bottom of the architecture, a Medical Record Agent (MRA) controls the access to a database that stores all medical records of the patients of the medical centre.

V. IMPLEMENTATION

The labeling technique described above has been implemented, and experiments have been conducted with web-based document articles selected from several domains. All documents used in these experiments have HTML-based file format. Samples consisting of 20 documents from four domains (car-ads, genealogy, computer job ads. obituary) downloaded from various online news papers. The current framework has been implemented by using java 1.5. To convert ontology as a java class file the system used JAXB (Java Architecture for XML Binding). The accuracy of result is same as that of generalized framework. This project is implemented and tested for the following application ontology: real estate, restaurant, cell phone plan, digital camera, camp ground and person.

a. Data Gleaning Processes

Data entities-The Data entities in the flow diagram are Input file (formatted text),application ontology with its data frame, human readable output.

Data frame definition-The data frame defines the extraction rules through its value phrase and keyword phrase constructs. The data frame's regular expression is compatible with Perl-5 regular expression

Ontology parser-This Process combines application ontology and its data frames and produce constant keyword matches rule, objects, relationships and constraints and database schema.

Value extractor-This process uses OSMX with its data frames as the ontology language. OSMX will be formed by OSML with XML specification. Locates the recognition rules specified by each data frame and applies them to the input text. The extractor identifies all substrings in the text that match the recognition rules and for each substring construct a matched text object.

Value Mapper-This process is most important and difficult part of the extraction system and it takes candidate value matches and uses them to build a data instance consistent with the constraints specified by the ontology.

Ontology writer-This process provides a human readable hierarchical list of objects and relationships in each data instance stored with the input ontology. The output format is HTML, which suits our present purpose.

b. PPDM Process of Generalization

The processes involved are:

QI generalization- QI Generalization process is responsible for

(i). Getting the QI attribute: The QI attributes are invoked from input database.

(ii). Performing Generalization functions

(iii). Output**:** This process yield the generalized QIgroup that have set of tuples with identical QI attribute values.

Optimal Generalization

This process is responsible for

(i). Getting the guarding node value: The guarding node values are invoked from the input database.

(ii). Performing SA generalization

(iii). Output**:** This process yield the optimal generalized information table without loss of privacy.The quality of output depends upon the effectiveness of the generalization.

Calculating Information Loss& Privacy Factor:

The calculating information losses & privacy factor will facilitate the software to check out the total information losses & privacy factor for every publishable table

The screen shots of the generalized QI datasets (QI group) are shown below. Figure 5.1 shows the QI generalized medical data sets that database contain the following field such as tupleno, sex, age, and zip code, disease where disease is the sensitive information Age, sex, Zip code these are Quasi Identifier (QI) attributes. QI group is, the sets of tuples which have same QI value except sensitive information

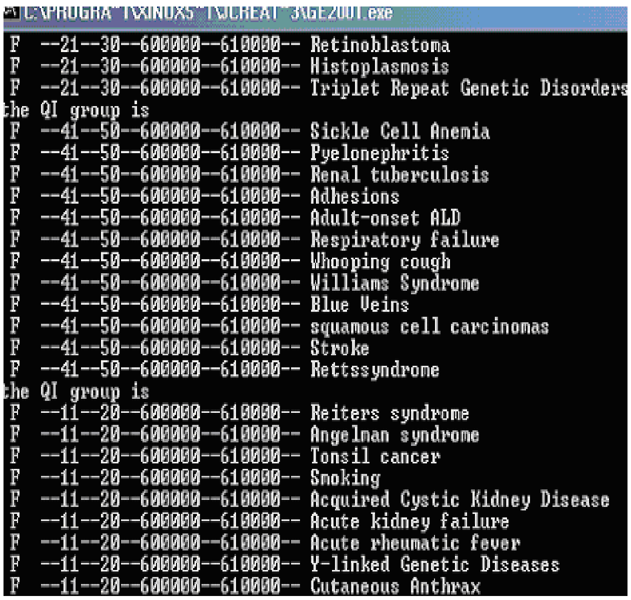

Figure 5.1 QI Generalization Output

Rounding of algorithm- Data analysis requirements are different, so the software should choose different generalization function to release different tables during each round. This software process the "Rounding of Algorithm" which is responsible for the subsequent execution of different level of QI generalization intended to satisfy the analyst requirements and next the Sensitive Generalization process will be invoked. In each round different QI generalization function is chosen without loss of privacy, that function performing the QI generalization on original QI attribute.

Displayer- This process invokes the preserved information from the database, and displays the preserved information

Topology- The software process follows a linear topology.

The following screen 5.2 shows the Output of Digital Camera domain

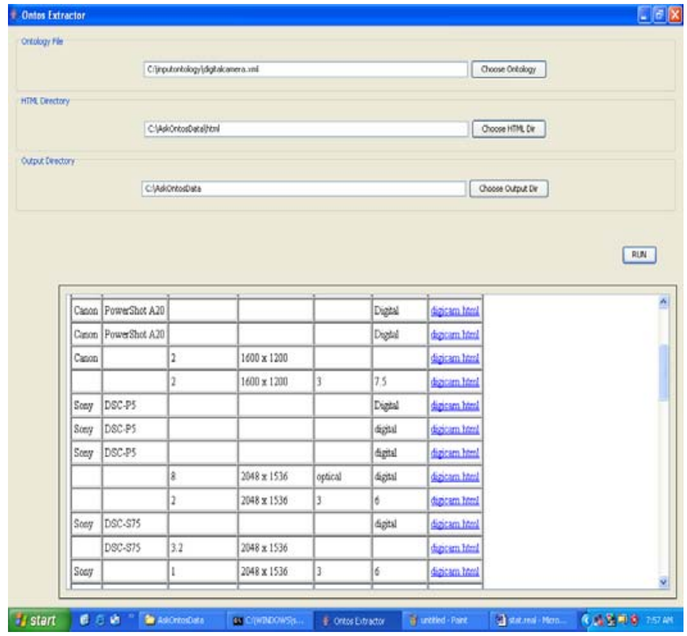

Figure 5.2. Digital Camera domain output

The following screen 5.3 shows the result statistics of Digital Camera domain

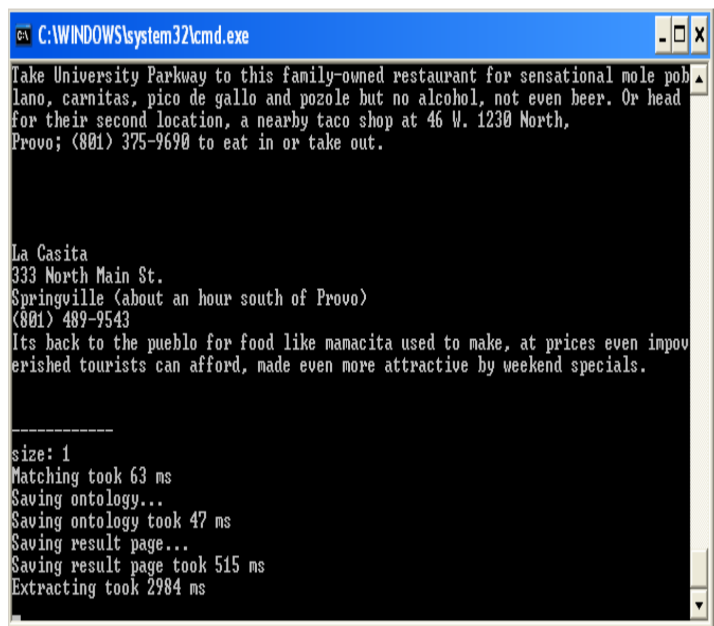

Figure 5.3. Result statistics of Digital Camera domain

VI.CONCLUSION AND FUTURE ENHANCEMENT

We have designed and implemented a modified framework for an ontology-based Data-extraction. The replacement of existing technique by our technique has toiled our efforts. The generalized framework is less flexible for module replacement. The modified framework result is also same as that of generalized framework. This work provides a solid basis for continued research on ontology-based Data-extraction. In this project, the solution was determined for extraction of text information from HTML document and plaintext document. The input documents are retrieved from the local system. The proposed solution will not consider the image, audio and video files. For small scale input the developed system is more efficient than the search engine. For large scale input the accuracy will be high but it will take more time than search engine.

The existing generalization methods are inadequate because they cannot guarantee privacy protection in all cases, and often incur unnecessary information loss by performing excessive generalization on all the attributes of the table. This proposed software proposes the technique to achieve privacy preservation is personalized anonymity and this new generalization framework that takes into account, customized personal privacy requirements. The releasing generalized information utilized for highly effective tasks, without breaching the privacy constraints formulated by data owners. Our technique has successfully prevented privacy intrusion.

Excessive generalization increases the privacy of an individual, but to increase the dataset utility, minimal generalization is required. So the proposed system releases different set of generalized tables to facilitate data researchers. The same data is anonymized and published multiple times to assist the data researchers, each of time in different form.

The current system will process only the text information and accepts the HTML/plaintext document as its input. In future the current system may be extended to process the image, audio and video information. The current system may be extended to accepts the other types of documents, such as PDF as its input in our further research we plan to extend our works and then implement them. Future work is related to investigate various inferences attack.

VII. ACKNOWLEDGEMENT

Primarily, my gratitude goes to Dr. D.Manjula, A.P, DCSE, Anna University Chennai, who guided me to carry out this research with interest and involvement. With profound reverence and high regards, I thank Dr. P. Narayanasamy, Professor and HOD, DISE, Anna University Chennai for the motivation provided by him during this research work which enabled me to complete this work successfully.

REFERENCES

[1] Alan wessman, Stephen W.Liddle, David W.Embley. 2005.A generalized framework for an ontology-based data-extraction.In the proceedings of the International Conference On Information Systems Technology And Its Application.Volume 63,pp.239-253.

[2] Ashish, N. and C. Knoblock. 1997. Wrapper generation for semi-structured Internet sources. *SIGMOD Record*, Volume 26, Number 4, pp. 8-15.

[3] Bayardo R. and R. Agrawal. Data privacy through optimal k-anonymization. In *ICDE*, pages 217–228, 2005.

[4] Crescenzi, V., G. Mecca, P. Merialdo. 2001 RoadRunner: Towards automatic data extraction from large Web sites. In *Proceedings of the 27th International Conference on Very Large Data Bases*, Rome, Italy, pp. 109-118.

[5] Embley, D.W.1980.Programming with data frames for everyday data items. *AFIPS '80 Proceedings*, Anaheim, California, pp. 301- 305.

[6] Fung, B., C. M. K. Wang, and P. S. Yu. Top-down specialization for information and privacy preservation. In *ICDE*, pages 205–216, 2005.

[7] Hammer, J., H. Garcia-Molina, S. Nestorov, R. Yerneni, M.Breunig, V.Vassalos.1997.Template-based wrappers in the Tsimmis system," In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Tucson, Arizona, pp. 532-535.

[8] Kushmerick, N., D. Weld, R. Doorenbos. 1997.Wrapper induction for information Extraction.In *Proceedings of the International Joint Conference on Artificial Intelligence*, Nagoya, Japan, pp. 729-737.

[9] Iyengar. V. Transforming data to satisfy privacy constraints. In *SIGKDD*, pages 279–288, 2002.

[10] Laender, A.H.F., B.A. Ribeiro-Neto, A.S. da Silva, J.S.Teixeira. 2002. A brief survey of Web data extraction tools," *SIGMOD Record*, Volume 31, Number 2, pp. 84-93.

[11] LeFevre K., D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain *k*-anonymity. In *SIGMOD*, pages 49–60, 2005.

[12] Liddle, S.W., D.W. Embley, S.N. Woodfield.2000.An active, object-oriented, model- equivalent programming language.*Advances in Object-Oriented Data Modeling*, MIT Press, pp. 333-361.

[13] Samarati. P. Protecting respondents' identities in microdata release. *TKDE*, 13(6):1010–1027, 2001.

[14] Srikant R. and R. Agrawal. Mining quantitative association rules in large relational tables. In *SIGMOD*, pages 1–12, 1996.

[15] Sweeney. L. , Achieving *k*-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):571–588, 2002.

[16] Wang K., P. S. Yu, and S. Chakraborty. Bottom-up generalization: A data mining solution to privacy protection. In *ICDM*, pages 249–256, 2004.

AUTHORS PROFILE

K. Murugesan**,** is pursuing his Ph.D in Computer Science from Anna University and also working for Anna University as a Superintendent. His research interests include Data Mining, Multi agent based systems etc.,

Indumathi.J is working for Anna University as Asst.Professor. Her field of interest spans and is not limited to Computer Science and Financial Management.

D.Manjula is working for Anna University as an Assistant Professor. Her research interests include Natural Language Processing, Text Mining, Artificial Intelligence, Databases and Parallel .