Journal of Computer Science 9 (11): 1451-1455, 2013
ISSN: 1549-3636

doi:10.3844/jcssp.2013.1451.1455 Published Online 9 (11) 2013 (http://www.thescipub.com/jcs.toc)

ANALYSIS OF INTELLIGENT DATA MINING FOR INFORMATION EXTRACTION USING JAVA AGENT DEVELOPMENT ENVIRONMENT PLATFORM

[1]M. Vinoth Kumar, [2]G. Tholkappia Arasu and [3]V. Palanisamy

[1]Department of CSE, K.S.R. College of Engineering, Tiruchengode, India
[2]A.V.S. Engineering College, Salem, Tamil Nadu, India
[3]Info Institute of Engineering, Coimbatore, Tamil Nadu, India

Received 2013-07-12, Revised 2013-09-19; Accepted 2013-09-20

ABSTRACT

In this study, the problem of unstructured information extraction has been analyzed and the need for a new Information Extraction algorithm is justified. We propose an Intelligent Information Extraction using Java Agent Development Environment (JADE) to be an optimal solution for intelligent information extraction. This proposed algorithm first assigns intelligent agents to gathering data, which they then classify, match and organize to construct sequential information for a user query. It improves the efficiency and performance for retrieving a proper information results as a sequence that satisfy user's needs. It gives the user needed documents based on similarity between query matching and relevant document mechanism. The results obtained from the Intelligent Information Extraction are optimal.

Keywords: Intelligent Agent, Intelligent Information Extraction, Intelligent Agent, JADE, Data Mining

1. INTRODUCTION

Information Extraction is the process of finding the exact Information for the given data from the Internet. Many search engines are used for data extraction but these would not give a structured data. The unstructured characteristic of the information sources on the World Wide Web makes automated discovery of Web-based information difficult. Because of the explosive development of information source available on the Internet and on the business, government and scientific databases, it has become quite necessary for the users to utilize automated and intelligent tools to extract knowledge from them (Singh *et al*., 2013).

The amazing growth of the WWW, Mobile Devices (Saravanan and Sumathi, 2013), Cloud Computing (Saravanan *et al*., 2013) and Wireless Communication (Saravanan and Sumathi, 2013) are the indications of its strength. However, as more and more information becomes available, the task of finding the right information becomes more and more difficult. Since Information extraction has to produce structured data for the further processing, it is very difficult to transform the input data into a structured one. So that content mining and additional intelligent tools are needed to get the desired data (Jiang *et al*., 2012). There is a growing consensus in the Internet community that one of the most promising solutions to the problem of Internet information retrieval is the use of software agent technology. The main feature of an agent based WWW is that information agents perform the role of managing, manipulating or collating information from many distributed sources (El-Bathy *et al*., 2011).

A search engine provides results related to the given key word but it could not provide the structural information and categorization, filtering or interpretation of the data. It makes researchers to design an intelligent framework to provide a structured data from the search engine.

Agents are considered one of the most important paradigms that on the one hand may improve on current methods for conceptualizing, designing and implementing software systems and on the other hand may be the solution to the legacy software integration problem.

Corresponding Author: M. Vinoth Kumar, Department of CSE, K.S.R. College of Engineering, Tiruchengode, India

Agent technology is originated from the branch of artificial intelligence known as distributed artificial intelligence. Intelligent agent is a collection of components that is characterized by, among other things, autonomy, pro-activity and an ability to communicate. Being autonomous it can independently carry out complex and often long-term, tasks. Being proactive it can take the initiative to perform a given task even without an explicit stimulus from a user. Being communicative it can interact with other entities to assist with achieving their own and others' goals (Chan *et al*., 2009).

Here intelligent agent is developed in which to construct a sequence set of information for a given key word. For that a set of agent framework is proposed and it extract a raw information from search engine and filter the closed information for the input and order the data into a sequential data with the multi agent environment. Here agents are communicate and coordinate each other to achieve a common goal. **Figure 1** shows the one of the web page for the given key word 'dengue fever' to know the status of affected people in a particular area.

This is taken from the famous news paper which shows the number of people affected in the area of Tirunelveli and the **Table 1** shows the sample extracted information of the given web page.

So getting the information about the total number of people affected for a particular disease in a particular area as a single data is very difficult from the internet. The agent framework defines the key word and finding the closed web pages for that disease and filters the exact count and the place from the web page and order the count and place as a single data. Likewise a set of disease spreading details can be analyzed and reported.

1.1. Related Work

Many Researchers have developed several approaches for intelligent agent based information retrieval. Among them, a handful of significant researches are presented in this section 1.1. Chan *et al*. (2009) have introduced an agent-based cognitive semantics retrieval system for location-aware information. They implemented intelligent agents for information grasping, classification, matching and organizing. They evaluated the performance of their system. In their method, intelligent agents first gather location-aware data and then, using semantic graphs in the Word Net English dictionary, they classified, match and organize the information to find a best match for a user query. Their experimental result showed that the effectiveness of our system leads to have more accurate and faster search result.

El-Bathy *et al*. (2011) have presented an Intelligent Extended Clustering Genetic Algorithm (IECGA) using Business Process Execution Language (BPEL), to be an optimal solution for data clustering. It improved the speed and accuracy for retrieving an exact data that satisfied user's needs. The proposed IECGA used different level of iteration to move next generation. This series of different iteration process depends on chromosome best fitness in the population and rely on high relevancy as well. The iteration operation was guaranteeing the result of IECGA for data clustering since it expands the search. So the continuous iteration provides the greater effects on the genetic process. Finally, IECGA for data clustering produce exact data based on similarity between query matching and relevant document mechanism. The results obtained from the web intelligent search engine are optimal.

Kang and Sim (2011) have presented a four-stage, agent-based Cloud service discovery protocol. Utilizing an ontology description, in which each resource was described semantically and relatively to other resources, they developed a multi-agent system that cooperates successfully by implementing a ontology-based matching. To increase the performance and the rate of finding match of the given information, they used a database to store and keep track of historical data for making intelligent recommendation based on attribute value prediction. Empirical results showed that when broker agents in their system used Cloud ontology and a connection procedure with a recommendation stage achieved better performance in finding the appropriate Cloud services.

Ma *et al*. (2011) have proposed an agent-based distributed environment where, weather data can be queried, loaded using agent-grid services according to the requirements. They presented a multi-agent-based framework to manage, share and query weather data in a geographical distributed environment, named Weather Data Sharing System (WDSS). In each node, some services were designed for querying and accessing data sets based on agent environment. The agents local and remote search was evaluated and the transfer speeds for different file types were also evaluated. From the presented platform, the system extensibility was analyzed.

1.2. System Model

The sample block diagram of our recommended technique is shown in the **Fig. 2** Here; we are giving the keyword related to the disease status to the search engine. The search engine finds the related documents u_i and sends it to the Information Gathering System. The Information Gathering System then performs the tasks such as named entity recognition and extracts the required data which we need to process and collects the required information in a database.

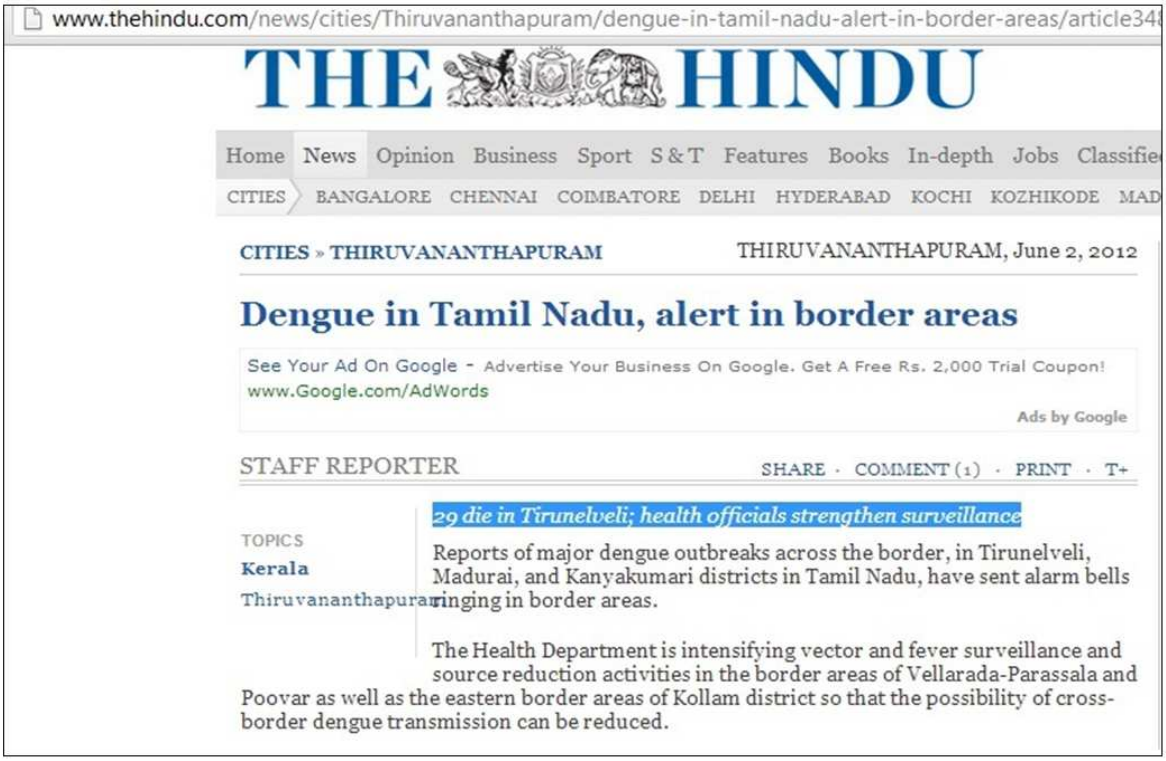

Fig. 1. Web page which shows the dengue fever details of Tirunelveli District

Table 1. Sample extracted data

Disease	Location	No. of persons affected	Time period
Dengue Fever	Tirunelveli	29	June 2012

The Data Transfer Agent transfers the required data The Information Gathering System then performs the tasks such as named entity recognition and extracts the required data which we need to process and collects the required information in a database.The text documents dj are the documents which contains the information in the form of text. These text documents are used to extract the required data for our proposed technique. The text documents are obtained from the search engine by giving a keyword. The search engine would give a list of links based on the keyword. The text documents di is collected from those links separately to obtain our required data for processing our proposed technique. The sample text document with its contents as follows Equation (1):

$$di = \{c1, c2, c3 \ldots cn\} \quad (1)$$

Where:

di = Number of document.

i = 1,2,3,...., n

cj = Contents in the document.

Where:

j = 1,2,3,..., n

The number of document di represent as a collection of content related to that document.the contents are represented as c1, c2 upto cn.

1.3. Information Gathering System

This system performs the operations such as named entity recognition, gathers the names of the locations, gathers the number of persons affected and the time period which the disease affected a particular location.

1.4. Named Entity Recognition

This is a subtask of data extraction which is used to categorize the names of persons, organizations, locations. The other names of Named Entity Recognition are Entity Identification and Entity Extraction. We have considered three attributes to process our proposed technique. The attributes we considered are location, number of person affected and the prevalence of the disease.

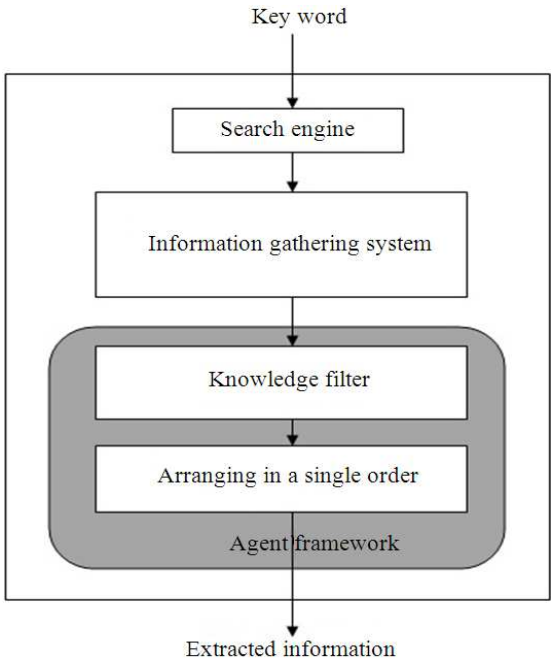

Fig. 2. Model of intelligent extraction system

1.5. Extraction of Location

The location is extracted from the document d based on the following procedure: first we have to create a database ld which contains the name of places. The location database ld is then compared with the contents cj of the document d to identify the name of the place in that document d. Similarly, the location database is compared with the entire documents di separately to find the place mentioned in those documents. A sample dataset ld that contains the name of the places and the process of identifying a location L from a document is as follows Equation (2 and 3):

$$ld = \{c_1, c_2, ..., c_n\} \quad (2)$$

$$L = \begin{cases} d_i c_j, & \text{for } d_i c_j = ldc_m \\ nil, & else \end{cases} \quad (3)$$

Where:
Ld = Database with the names of the location
L = Extraction of location from a document
Djcj = j^{th} content in ith document
ldcm = m^{th} content in the database ld

1.6. Extraction of Number of Persons Affected

The extraction of number of persons affected from the documents di is as follows: we need to create a database pd which contains the word 'affected' and the related words which gives the meaning of the word 'affected' or 'diagnosed'. Thereafter, we have to check this database pd with the document d and if the document d contains the word which is in the database pd, we need to check the preceding word pdck-1 and the following word pdck+1 of that word which is in the database and in the document and if the preceding pdck-1 or following word pdck+1 is in numerical, we have to take it as the number of the affected persons. A sample dataset that contains the word 'affected' and the related words which gives the meaning of the word 'affected' or 'diagnosed' is as follows Equation (4):

$$pd = \{c_1, c_2, ..., c_n\} \quad (4)$$

where, pd is database contains the word 'affected' and the related words.

The process of identifying the number of persons affected is as follows Equation (5):

$$P = \begin{cases} pdc_{k-1}, & \text{if } pdc_{k-1} = \text{numerical and } d_i c_j = pdc_k \\ pdc_{k+1}, & \text{if } pdc_{k+1} = \text{numerical and } d_i c_j = pdc_k \\ nil, & else \end{cases} \quad (5)$$

Where:
P = Extraction of number of persons affected
pdck-1 = k-1th content in the database pd

2. MATERIALS AND METHODS

Information extraction is done by two computer systems. One system will act as search engine system.It collects important information from different web pages.The extraction agent is implemented to search the keyword from different web pages. The results of all links are stored in this system. The second system is the extraction system which extract the exact information from the first system. The transmission agent is implemented to transfer data from search engine system to the extraction system.The extraction agent and the data transfer agent is implemented by Java Agent Development Environment framework.

3. RESULTS AND DISCUSSION

After identifying the location, the number of people affected and the time period of the prevalence of disease from every document, we have to store those data in a database. This database is then used for our technique. The **Table 2** shows the sample database which contains the gathered information.

Table 2. Database with the gathered information

Disease	Location	No. of persons affected	Time period
D1	Madurai	80	Mar. 2012
D2	Tirunelveli	71	Jul. 2012
.	.	.	.
.	.	.	.
.	.	.	.
Dn	Salem	92	Nov. 2012

4. CONCLUSION

In this study, we formalized an unconventional and promising approach towards structured information extraction from the Web. The approach uses a model of the visual representation of web pages as rendered by a web browser and, therefore, shifts the problem of information extraction from the lower level of code interpretation (HTML tag structure, CSS, JavaScript code) to the higher level of visual features (2-D topology WWW 2007/Track: Data Mining Session: Identifying Structure in Web Pages 79 and typography). We have also presented an idea for representing web table structures along technology that we have used derive instances of the model given some arbitrary web pages. Our approach strives to perform well even without tuning for specific application domains such as the interpretation of product catalogues. Although our results are preliminary at the current state, we believe that applying a visual paradigm towards automatic information extraction from web tables is promising, especially given the rising complexity in the encoding of web pages on the source code level. Specially, highly dynamic pages which tend to get more popular with the rise of Web 2.0 cannot be processed without complex interpretation of the source code.

5. REFERENCES

Chan, E.C.L., G. Baciu and S.C. Mak, 2009. Cognitive location-aware information retrieval by agent-based semantic matching. Proceedings of the 8th IEEE International Conference on Cognitive Informatics, Jun. 15-17, IEEE Xplore Press, Kowloon, Hong Kong, pp: 435-440. DOI: 10.1109/COGINF.2009.5250701

El-Bathy, N., C. Gloster, I. Kateeb and G. Stein, 2011. Intelligent extended clustering genetic algorithm for information retrieval using BPEL. Am. J. Intell. Syst., 1: 10-15. DOI: 10.5923/j.ajis.20110101.02

Jiang, C., B. Zhang, Y. Yu and X, Zhang, 2012. An intelligent agent-based framework for information security management. Instrument. Manage. Circ. Syst., 127: 807-814. DOI: 10.1007/978-3-642-27334-6_95

Kang, J. and K.M. Sim, 2011. Towards agents and ontology for cloud service discovery. Proceedings of the International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Dec. 10-12, IEEE Xplore Press, Beijing, pp: 483-490. DOI: 10.1109/CyberC.2011.84

Ma, T.H., W. Tian, B. Wang, D.H. Guan and S.Y. Lee, 2011. Weather data sharing system: An agent-based distributed data management. IET Soft., 5: 21-31. DOI: 10.1049/iet-sen.2009.0027

Saravanan, V. and A. Sumathi, 2013 Biologically-inspired vertical mobile handoff with seamless mobility. Int. J Innovat. Res. Sci. Eng. Technol., 2: 4211-4220.

Saravanan, V., S. Thirukumaran, M. Anitha and S. Shanthana, 2013. Enabling self auditing for mobile clients in cloud computing. Int. J. Adv. Comput. Technol., 2: 53-60.

Singh, A.K., D. Sharma and A. Pathak, 2013. Web usage Mining: A concise survey on tools and applications. Int. J. Comput. Applic., 74: 0975-8887. DOI: 10.5120/12846-9076