ASIAN JOURNAL OF PHARMACEUTICAL AND CLINICAL RESEARCH

INNOVARE ACADEMIC SCIENCES
Knowledge to Innovation

Special Issue (April)

Online - 2455-3891
Print - 0974-2441

Full Proceeding Paper

LAYERED ARCHITECTURE TOWARD MINING AYURVEDIC FACTS

GAYATHRI M*, JAGADEESH KANNAN R

Department of , SCSVMV University, Kanchipuram, Tamil Nadu, India. Email: mgayathri@kanchiuniv.ac.in

Received: 23 January 2017, Revised and Accepted: 03 March 2017

ABSTRACT

The main objective of this paper is to analyze the difficulties in finding the relevant information in ayurvedic medical system and the changes and opportunities that the information technology brings for different aspect of traditional Indian medicine. We describe the ontology's and semantic tools to obtain the deep knowledge among the data. Moreover, we adapted the ontology-based model which helps in finding the semantic information with limited time complexity. This model proves that it suits to find the useful data without degrading the efficiency and performance of the system.

Keywords: Text mining, Traditional medicine, Ontology, Distributed data storage.

INTRODUCTION TO AYURVEDA

Ayurveda [1,2] is the system of traditional medicine prevalent in India since 2000 B.C. Ayurveda means the "science of life." Ayurveda derives medicine from nature. After thorough study, experimentation and documentation of hundreds of plants over a period of more than a thousand years, India's ancient sages have come to accurate conclusions about the efficacy of different plants and herbs. Although efficacy of Ayurveda for a variety of human ailments is well known in and around India, most of the world is not aware of the benefits that could be derived from this unique Indian system of medicine. Most of the ayurvedic preparations are free from side effects or reactions. Ayurveda [3] provides rational means for the treatment of many internal diseases which are considered to be obstinate and incurable in other systems of medicine. Life according to Ayurveda is a combination of senses, mind, body, and soul. Hence, it is clear that Ayurveda is not only limited to body or physical symptoms but also gives a comprehensive knowledge about spiritual, mental, and social health. Ayurveda describes three fundamental mind/body types or doshas called Vata, Pitta, and Kapha, which embody different combinations of the five elements: Air, ether, fire, water, and earth. If vata dosha is the main life force, they are more likely to develop: Anxiety, asthma, heart disease, nervous system disorders, rheumatoid, and skin problems. If pitta dosha is the main life force, they are more likely to develop: Anger and negative emotions, Crohn's disease, heartburn a few hours after eating, high blood pressure, and infections. If kapha dosha is the main life force, they are more likely to develop: Asthma and other breathing disorders, cancer, diabetes, nausea after eating, and obesity. When these elements are balanced, one is healthy. Illness is defined as an imbalance of these elements; all disorders are excesses of one or more element. Ayurveda is a form of treatment by natural remedies, which makes use of the power of nature to restore human beings to a state of balance. Diagnosis is a very vital aspect of ayurvedic treatment. Diagnosis according to Ayurveda is to find out the root cause of a disease. Unless the proper diagnosis is done, it is difficult to provide medicine and cure the disease. The diagnosis and the treatment of disease are always individual to each patient. As Ayurveda treats according to the constitution of an individual, it is known as a highly accurate and personalized method of analyzing diseases. The treatment mainly comprises powders, tablets, decoctions, medicated oils, etc., prepared from natural herbs, plants, and minerals. Plant-based treatments in Ayurveda may be derived from roots, leaves, fruits, bark, or seeds such as cardamom and cinnamon and the animal products used in Ayurveda include milk, bones, and gallstones.

MINING MEDICAL INFORMATION

Researchers, Life Scientist working in the domain of Ayurveda tries to publish their findings in the form of electronic information. Such information exists abundantly on the web on daily basis. Whenever we try to find the information in any of the popular search engine, regarding any kind of diseases, we may result with thousands of result of which some are relevant and some are irrelevant pertaining to our search. Our result is narrowed by the irrelevant hits produced and it is difficult for the user to retrieve the most appropriate information. To solve this, we use many data mining techniques to discover the knowledge. Data mining [4] seems helpful when we try to find the relevant information from the structured database. Since the medical information across the internet is available in unstructured way we need to preprocess the content using Some Natural Language processing technique. Hence, the process of finding the useful information from unstructured content with the help of NLP is what we call as text mining. In relate to this context, there is various text mining tool exist both commercially and free of cost. Open source software is of interest. Some of the tools which are suitable for medical data include GATE, rapid miner, and prolog.

ISSUES IN FINDING THE RELEVANT INFORMATION

There are various issues in finding the relevant information from heterogeneous source. Here, we listed as,

1. The medical information lacks structure means that the content available in the internet is of unstructured.
2. Most of the content holds biomedical (medical subject headings) terms.
3. Term synonyms - the same word holds multiple meaning.
4. Term polysemy - the same word can be expressed differently.
5. Massive volumes of information.
6. Term ambiguity - this exist due to the variation in the names of the Indian herbs from region to region.

ONTOLOGY IN TIM

With the improvement in technology and growth of internet large number of medical databases have emerged. Semantic web based on ontologies are introduced to address the problem of data heterogeneity in different databases. The term ontology refers to the understanding of domain of interest. Ontology [5,6] is an explicit specification of a conceptualization, a conceptualization means an abstract model of some aspect of the world, taking the form of a definition of the properties of important concepts and relationships. An explicit specification means that the model should be specified in some unambiguous language, making it amenable to processing by machines as well as by humans. Ontologys are becoming of increasing importance in fields such as knowledge management, information integration, cooperative information systems, information retrieval, and electronic commerce the effective use of ontology's requires not only a well-designed and well-defined ontology language but also support from reasoning tools. Fig. 1 shows the graphical representation of concepts and its relation using ontology.

OVERVIEW OF KNOWLEDGE DISCOVERY IN TRADITIONAL MEDICINE

The following Fig. 2 illustrate the overview of knowledge discovery in traditional medicine. By taking the traditional medicine information sources, such as the TIM bibliographic literature, TIM clinical data and the ancient TIM literature as the main data sources, the text mining tasks include the recognition of TIM named entities and the relationships of those entities, the extraction of the constituent herb information (e.g. herb name and herb dosage) in formula, and the discovery of novel clinical facts and events.

One of the main objectives of text mining [7] in Ayurveda medicine is to help generate scientific hypotheses and clinical guidelines for practical diagnoses and treatments. To achieve this aim, it is essential to extract the clinical facts and events from the data. There are two important kinds of ayurvedic medicine knowledge which should be extracted by text mining methods: The relationships of the ayurvedic named entities (e.g., syndrome-symptom relationship, disease-syndrome relationship, and herb-symptom relationship), and the constituent herb information of formula in Ayurveda.

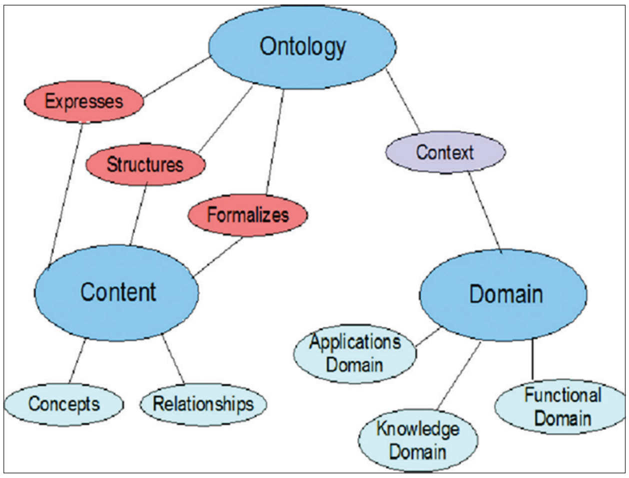


Fig. 1: Ontology representation

In addition, to improve the quality and efficiency of the text mining process, information retrieval and text classification would be indispensable to facilitate the data searching and filtering of the ayurvedic literature. As the insights and hypotheses are most likely to be found by integrating multiple traditional medicine data sources, the development of integrative data mining methods would be a promising step for TIM text mining.

LAYERED APPROACH TOWARD MINING

To meet the objective in mining, the appropriate and relevant textual data and improving the quality and efficiency of the text mining process. We proposed the layered architecture for retrieving information about Ayurveda. Fig. 3 describes the architecture of this layered approach.

It includes four layers:

1. Application interface layer
2. Knowledge discovery layer
3. Ontology layer
4. Data content layer.

Application interface layer

Considering researches, users, life scientist, or medical practitioners their at most goal is to find the information which they actually want. They used to submit their problem in the form of query. This is responsible for getting the query from the user and submitting the response as a query result to the user. To yield knowledge about ayurvedic facts, herb name used in Ayurveda treatment can be given as an input. Ranking is presentation of information fetched by text retrieval engine to user based on efficiency parameters and performance measure, i.e., recalls precision, F-measure, and ultimately information gain. Ranking algorithm are used to find the most and the relevant and important web pages. We used page rank algorithm which allocates a arithmetical worth to each building block of a set of hyperlinked corpus (i.e., web pages) within the World Wide Web with the purpose of measuring the relative importance of the page. The key

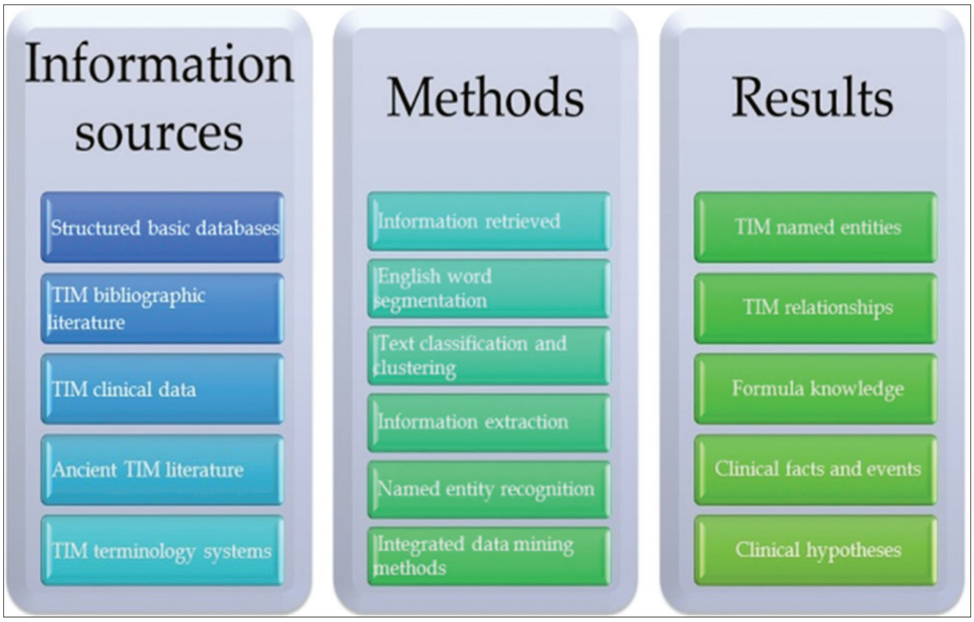


Fig. 2: Overview of knowledge discovery in traditional medicine

idea in the algorithm is to give a higher page rank value to web pages which are visited often by web surfers.

Knowledge discovery layer

This layer is responsible for processing the query which actually involves the task of understanding the semantic of herbs given as input. As far as the Indian herbs are concerned it varies from one region to another region. Therefore, it results in term ambiguity and it also holds with the drawback of term polysemy and hyponyms. To handle word synonyms and polysemic words of every Indian herbs of database is maintained. Moreover, it understands the relation between the input queries along with the related terms. Here, the Indian herbal database is constructed using the ontology model which best understands the relationship between the herbals and the diseases that the particular herb cures. It includes text mining discovery algorithm for retrieving the relevant textual data pertaining to the user request.

Ontology layer

As we mentioned earlier, ontology [8] defines the domain of interest. It identifies classes, objects, and its properties. Data are collected from heterogeneous sources such as www, traditional knowledge digital libraries, and other online resources using crawling and fetching technique. Since the collected data are in unstructured form it is to be preprocessed using NLP Techniques. Various text mining tools are available for preprocessing. We used GATE [9]. It is open source free software which has infrastructure for developing and deploying software components that process human language. GATE developer, an integrated development environment for language processing components bundled with a very widely used information extraction system and a comprehensive set of other plugins. GATE is distributed with an IE system called ANNIE which includes Tokenizer, Gazetteer, sentence splitter, POS tagger, semantic tagger, orthomatcher, etc. Fig. 4 show the GATE shot used for text analyzing.

Stop word removal

The words which occurs more frequently and commonly in a document but does not add meaning are referred to as stop words. Examples of stop word includes the, is, at, which, and on.

Stemming

It is the process of reducing inflected (or sometimes derived) words to their word stem, base, or root form – generally a written word form. Stemming algorithm reduces the words “fishing,” “fished,” and “fisher” to the root word, “fish.”

Tokenization

It is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens.

Automatic summarization

Produce a readable summary of a chunk of text. Often used to provide summaries of text of a known type, such as abstract from the literature content.

Coreference resolution

Given a sentence or larger chunk of text, determine which words (“mentions”) refer to the same objects (“entities”). Anaphora resolution is a specific example of this task, and is specifically concerned with matching up pronouns with the nouns or names that they refer to. The more general task of coreference resolution also includes identifying so-called “bridging relationships” involving referring expressions. For example, in a sentence such as “He entered John’s house through the front door,” “the front door” is a referring expression and the bridging relationship to be identified is the fact that the door being referred to is the front door of John’s house (rather than of some other structure that might also be referred to).

Disclosure analysis

This rubric includes a number of related tasks. One task is identifying the discourse structure of connected text, i.e., the nature of the discourse relationships between sentences (e.g., elaboration, explanation, contrast). Another possible task is recognizing and classifying the speech acts in a chunk of text (e.g., yes-no question, content question, statement, assertion, etc.).

Named entity recognition

Given a stream of text, determine which items in the text map to proper names, such as people or places, and what the type of each such name is

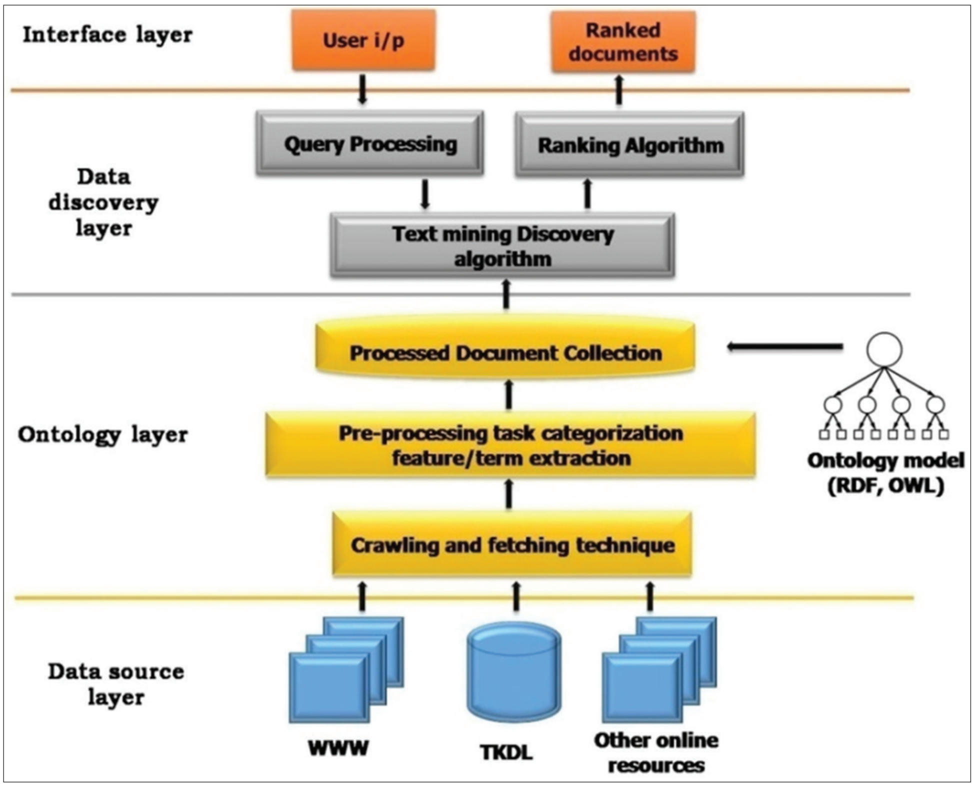
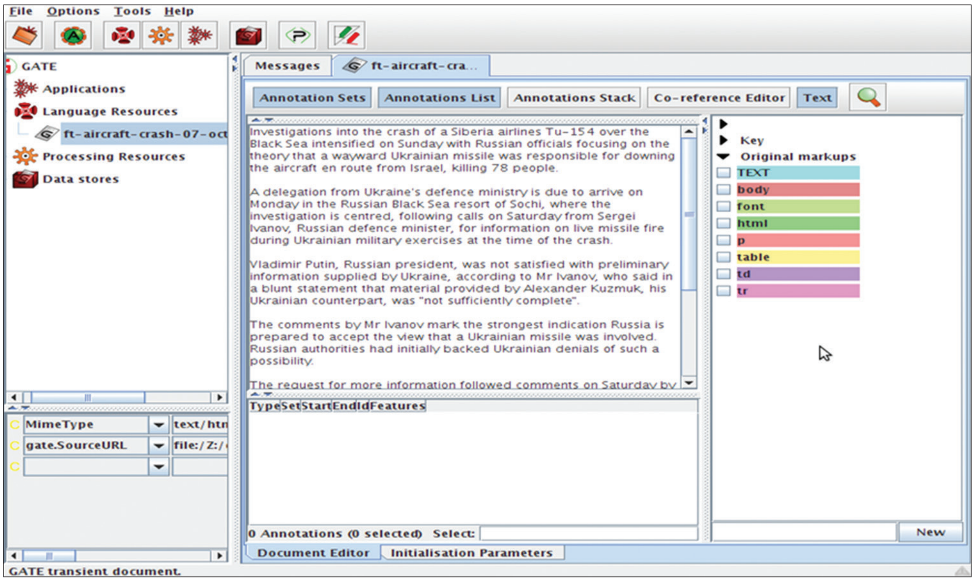

Interface layer
User i/p
Ranked documents
Query Processing
Ranking Algorithm
Data discovery layer
Text mining Discovery algorithm
Processed Document Collection
Ontology layer
Pre-processing task categorization feature/term extraction
Ontology model (RDF, OWL)
Crawling and fetching technique
Data source layer
WWW
TKDL
Other online resources

Fig. 3: Layered approach in mining



Investigations into the crash of a Siberia airlines Tu-154 over the Black Sea intensified on Sunday with Russian officials focusing on the theory that a wayward Ukrainian missile was responsible for downing the aircraft en route from Israel, killing 78 people.

A delegation from Ukraine's defence ministry is due to arrive on Monday in the Russian Black Sea resort of Sochi, where the investigation is centred, following calls on Saturday from Sergei Ivanov, Russian defence minister, for information on live missile fire during Ukrainian military exercises at the time of the crash.

Vladimir Putin, Russian president, was not satisfied with preliminary information supplied by Ukraine, according to Mr Ivanov, who said in a blunt statement that material provided by Alexander Kuzmuk, his Ukrainian counterpart, was "not sufficiently complete".

The comments by Mr Ivanov mark the strongest indication Russia is prepared to accept the view that a Ukrainian missile was involved. Russian authorities had initially backed Ukrainian denials of such a possibility.

The request for more information followed comments on Saturday by

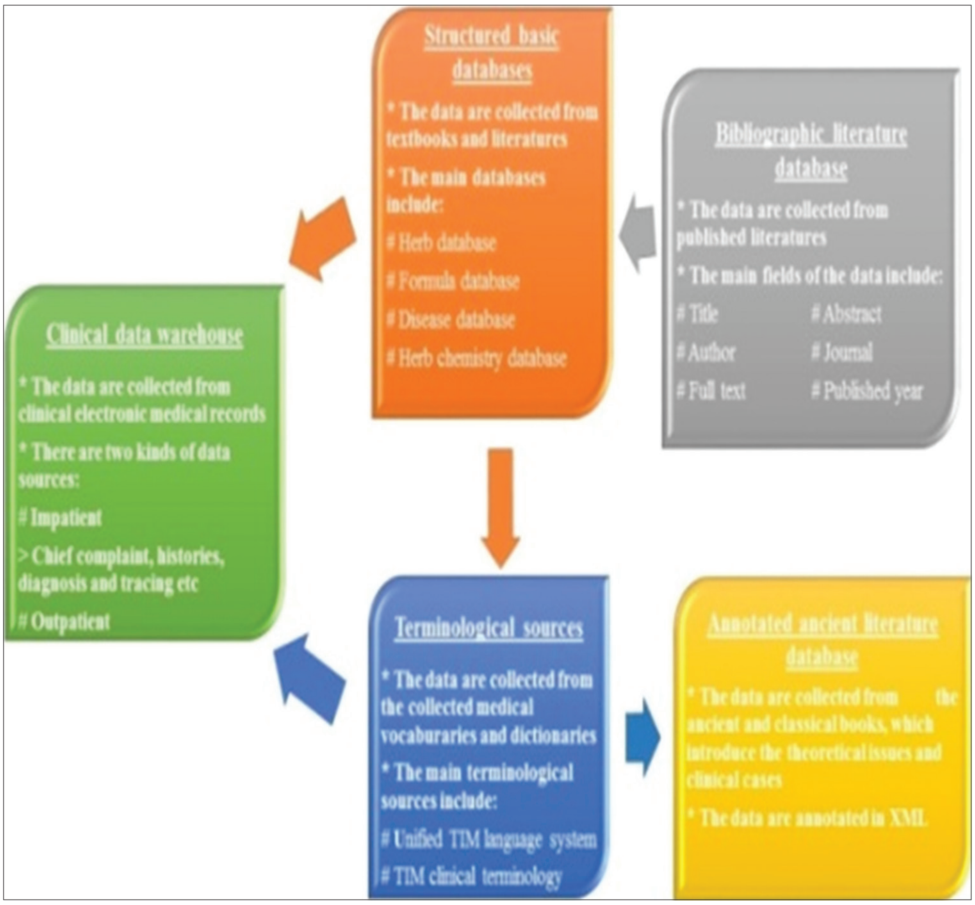
Type	Set	Start	End	Id	Features

0 Annotations (0 selected) Select:

MimeType	text/htm
gate.SourceURL	file:/Z:/

GATE transient document.

Fig. 4: GATE for preprocessing



Structured basic databases

* The data are collected from textbooks and literatures
* The main databases include:
Herb database
Formula database
Disease database
Herb chemistry database

Bibliographic literature database

* The data are collected from published literatures
* The main fields of the data include:
Title # Abstract
Author # Journal
Full text # Published year

Clinical data warehouse

* The data are collected from clinical electronic medical records
* There are two kinds of data sources:
Impatient
> Chief complaint, histories, diagnosis and tracing etc
Outpatient

Terminological sources

* The data are collected from the collected medical vocaburaries and dictionaries
* The main terminological sources include:
Unified TIM language system
TIM clinical terminology

Annotated ancient literature database

* The data are collected from the ancient and classical books, which introduce the theoretical issues and clinical cases
* The data are annotated in XML

Fig. 5: Data warehouse (conversion from unstructured to structured data)

(e.g., person, location, organization). Note that, although capitalization can aid in recognizing named entities in languages such as English, this information cannot aid in determining the type of named entity, and in any case is often inaccurate or insufficient. For example, the first word of a sentence is also capitalized, and named entities often span several words, only some of which are capitalized.

Part of speech tagging

Given a sentence, determine the part of speech for each word. Many words, especially common ones, can serve as multiple parts of speech. For example, "book" can be a noun ("the book on the table") or verb ("to book a flight"); "set" can be a noun, verb or adjective; and "out" can be any of at least five different parts of speech.

Word sense disambiguation

Many words have more than one meaning; we have to select the meaning which makes the most sense in context.

To have formal representation of the knowledge by a set of concepts within a domain and the relationships between those concepts ontology model is used to construct the database. To implement such a representation, several languages have been developed. The one that currently gets the most attention is probably the web ontology language. Ontologies do provide the means to store such information, which allows for a much richer way to store information. This also means that we can construct fairly advanced and intelligent queries. Query languages such as SPARQL have been developed specifically for this purpose. To design the plant ontology for our work, we used Protégé. It is a free, open source ontology editor and a knowledge management system. Protégé provides a graphic user interface to define ontologies. It also includes deductive classifiers to validate that models are consistent and to infer new information based on the analysis of ontology.

Data source layer

This layer actually contains the web resources where they can find the information. The amount of information available in the internet is increased exponentially. There is been an explosion in the amount of textual data in the biomedical domain. Researchers, life scientist tries to publish their findings in the form of literature thereby the growth of digital data in the internet and other online web resources are expanding dramatically. Simply considering the online digital repository such as PubMed, a database containing 12,000,000 references of biomedical publications. The information are said to be scattered and most of the information are weakly structured. The clinical data warehouse is typically maintained to gather the information from the web sources. Fig. 5 illustrates data warehouse which actually holds the structured data.

CONCLUSION AND FUTURE WORK

Here, we presented a layered approach toward effective mining of textual data regarding Ayurvedha medicine. The system clearly identifies the semantically related data thereby improves the rate of retrieving the relevant document pertinent to the user request. Moreover, the ontology-based model which helps in finding the semantic information with limited time complexity without degrading the efficiency and performance of the system. Future work include: (i) A deep analysis of expressiveness and complexity issues in mining the biomedical facts, (ii) an extension of the approach that enables to exploit semantic web standard ontology language, (iii) the adoption of other ontology representation approaches and parsing strategy.

REFERENCES

1. Ravishankar B, Shukla VJ. Indian systems of medicine: A brief profile. Afr J Tradit Coplement Altern Med 2007;4:319-37.
2. Web Resource. Available from: http://www.en.wikipedia.org/wiki/AYUSH.
3. Patwardhan B,Mashelkar RA.Traditional medicine-inspired approaches to drug discovery: Can Ayurveda show the way forward? Drug Discov Today 2009;.
4. Han J, Kamber M. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann; 2001.
5. Oro E, Ruffolo M. XONTO: An Ontology-based System for Semantic Information Extraction from PDF Documents 20th IEEE International Conference on Tools with Artificial Intelligence; 2008.
6. Horridge M. A Practical Guide to Building OWL Ontologies Using Protégé 4 and CO-ODE Tools; 2008.
7. Kayed M, Shaalan KF. A survey of web information extraction systems. IEEE Trans Knowl Data Eng 2006;18(10):1411-28.
8. Chen SW, Tseng YT, Lai TY. The design of an ontology-based service-oriented architecture framework for traditional Chinese. Med Healthcare 2012;:353-6.
9. Web Resource. Available from: https://www.gate.ac.uk.