A Survey for Prevention of Chronic Diseases by Information Extraction

Ms. Anie Pearlin M[1], *Mrs. Thamarai Selvi R*[2]

[1]*M.Phil Scholar, Department of Computer Science, Bishop Heber College (Autonomous), Trichirappalli, Tamil Nadu, India*

[2]*Asst.Professor, Department of Computer Applications, Bishop Heber College (Autonomous), Trichirappalli, Tamil Nadu, India*

Abstract

Chronic diseases are a long-lasting problem, which can be controlled but not cured. Many search engines are dumped with large number of scientific papers. The medical practitioner cannot maintain the patient records and risk factors. For that, a computing system used is called Chronic Illness Surveillance System (CISS). CISS is an assemblage of scientific papers which relates to chronic diseases and it constructed to retrieve relevant scientific research papers that link with chronic diseases. CISS helps the medical practitioners to identify the relationship between patient risk factors and chronic diseases. To evaluate CISS, it is compared with PubMed search engines using Question Answering (QA) System. QA system is usually in Natural Language (NL) formulation. The survey papers discuss about QA system and classification technique in information retrieval.

Keywords: *Biomedical Informatics, Information Retrieval, Question Answering system and Text Classification Technique.*

1. Introduction

The Chronic diseases are the major health problem and cause death. The World Health Organization reported that there is 1.7 million people are affected by chronic diseases. It cannot prevented by immunizing agent and not curable, but can control it. Chronic diseases are categorized like heart diseases, type 1 and 2 diabetes, obesity, etc. The main causes for chronic diseases are smoking; absences of physical activity, extreme use of intoxicant and short temper are the key reasons.

An information retrieval (IR) is the tracing and extracting relevant document from the collection of information resource. The main aim of IR is to afford the exact answer for the questions asked by the questioner [2]. Question answering is an advanced class of information retrieval which is qualified by information needs and conveyed as natural language or questions, and is one of the natural forms of human computer interaction.

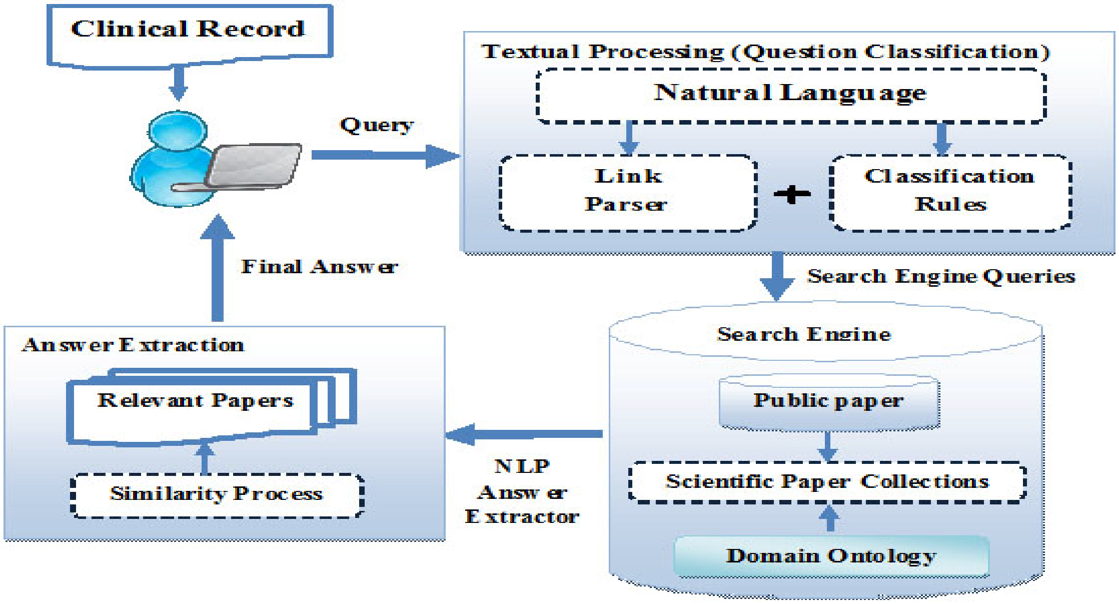
QA system is commonly used Natural Language that means human language and is a specific concept of information retrieval. QA system has classified as three parts, namely, question classification, information retrieval and answer extraction. In Question classification the questions are classified automatically according to the question taxonomy and the template. [4].

2. Methodology

The CISS is combined with information retrieval systems. It is used to direct the medical practitioners about the dangerous factors on chronic diseases like heart diseases, type 1 and 2 diabetes, obesity.

In design and implementation part the CISS are first it prepares and updates the collection of scientific papers which relate to the risk factors. In this it also finds the relevance between the documents and it affords ranking process. Then in the second step it processes the textual information from the papers and stores the metadata. Finally the pre-processor extract the relevant scientific papers which are related to the medical records [1].

Figure 1: Question Classification and Answer Extraction Process



Natural language is a rising field of bioinformatics. The process of CISS is pictured in Figure 1. The CISS system uses the Vector Space Model to find out the similarity between the scientific papers and weight matrix is used to find the similarity of term in document.

The CISS is evaluated with PubMed search engines like Google using natural language. Then the extracted papers are classified into four categories:

- All papers are closely related (+)
- Most papers are fairly related (+/-)
- Most papers are not related (-/+)
- All the papers are irrelevant (-).

2.1. QA System

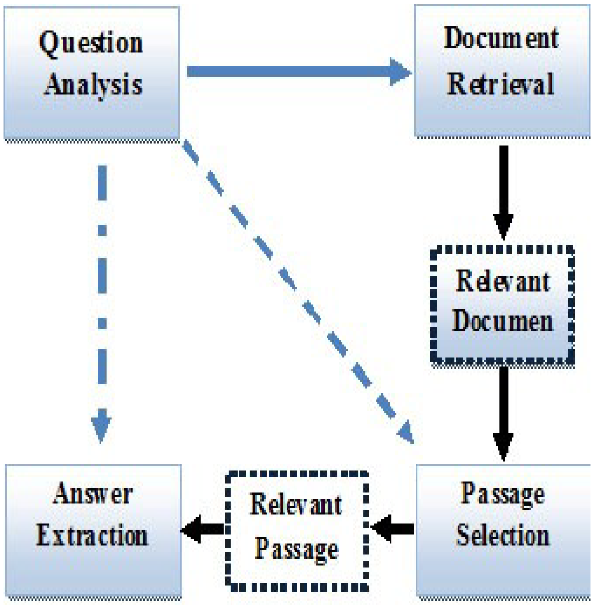
The QA System plays a main role in CISS system. QA system has three main modules. Question classification is most important part of the Query Processing Module, like wise information retrieval is the essential concept for the Document Processing Module and Answer Extraction is necessary notion for the Answer Processing Module.

Kolomiyets et al. [3] discuss that question is defined as a natural language clause, which usually starts with a query. Question types are characterized by some common properties. They are, Factoid Questions usually start with a Wh-interrogated word, List Questions requires as an answer a list of entities or facts, Definition Questions finding the definition of the term starts with “What is”, Hypothetical Questions requires information about a hypothetical event “What would happen if”, Causal Questions requires explanation of an event like “Why”, Relationship Questions asks about a relation between two entities, Procedural Questions requires list of instructions for mentioned, Confirmation Questions requires a Yes or No answer to an event. The natural language questions and contents are translated into structured form using low level bag-of-words. The ranking process is afforded to rank the information which is most relevant to the query.

The questions are evaluated with Precision and Recall which are based on the complete list of known distinct cases of the answers. Question answering is a composite task requires effective betterments including information retrieval, natural language processing, database technologies, Semantic Web technologies and computer interaction. Most of the medical question classifications were designed for literature based question and answering systems. Question answering technology for clinical needs relies on relevant matches answer, according to the semantic interpretation of the question.

In question answering approach the user questions are casting into three categories. Those are semantic based, inference based and logic based. The QA processing stages rely on four main modules those are Question Analysis, Document retrieval with relevant document, Passage Selection with relevant passages and Answer Extraction, as shown in Figure 2.

Figure 2: QA System Architecture



Nigam H Shah et al. [6] built a prototype system for ontology based notation and classification of biomedical terms. The system procedures the text of various resource elements like medical terms, description, PubMed abstracts to notation and index them with ideas from reserving ontology. The system constructed with all concept names and synonyms of medical terms like UMLS. The question classification uses the Unified Medical Language System (UMLS), which is kept by the US National Library of Medicine. The UMLS is useful to utilize the semantic information to perform the question classification work [2].

2.2. Text Classification Technique

In general, text classification plays a vital role in information extraction and question answering system. The task of Text classification technique is to classify the scientific papers under a predefined category. The classification process uses machine learning techniques to find the best papers to use to rank succeeding search outcome. Bader Aljaber et al. [5] proposed a new method of selecting and utilizing citation context used in academic publication. The citation context helps to improve information retrieval applications and MeSH (Medical Subject Heading) term classification. MeSH is a list of terms built by the National Library of Medicine for indexing and searching medical content. Machine learning technique is also working on document similarity that used to find the best answer for the medical practitioner query.

3. A Review of Literature

This section discusses about the techniques used for information extraction.

[1] *Surveillance for the prevention of chronic diseases through information association - Juliana Tarossi Pollettini, et. al.,*

This paper discusses about the extraction of scientific research papers which are related to the chronic diseases like cardiovascular disease, obesity and diabetes. They built a computational system named as a Chronic Illness Surveillance System, which is used to retrieve the relevant scientific papers according to the diseases. They used machine learning technique and vector space model to read the scientific papers which are stored in database and find the similarity between the papers. The surveillance system uses question answering system and tf–idf weighting scheme, which is a classification technique frequently used in search engines for scoring and ranking a document's relevance for user query with the help of UMLS and extract relevant research papers. Then the system has evaluated with the other search engines like Google, Google Scholar and PubMed.

[2] *Biomedical question answering: A survey - Sofia J. Athenikos, et. al.,*

This paper discusses about the QA system, question classification and answer extraction of the biomedical information. QA system uses the semantic knowledge-informed technique, which takes the benefit of the lexico-semantic information encoded in WordNet. The inference based approach describes the structure of natural language and analysis of textual information related for their biomedical QA system. The logic based approach converts the WordNet glossaries into axioms through Logic Form Transformation.

[3] *A Survey on Question Answering Technology from an Information Retrieval Perspective - Oleksandr Kolomiyets, et. al.,*

This article provides an overview of question answer technology. The QA task from an information retrieval view and expresses the importance of the retrieving models. This paper discusses about data sharing between information extraction, natural language processing and information retrieval. Information extraction techniques are necessary for examining natural language questions. These techniques often depend on natural language processing tools. Information retrieval affords the necessary functions for computing and ranking answers. Affording ranking is depending upon the relevance between the query and the information. In another extreme keyword-based searches existed from the natural language questions, and answers existed through reasoning.

[4] *An ontology for clinical questions about the contents of patient notes - Jon Patrick, et. al.,*

This paper focuses on the question process and classification methods for answering the medical report. The question taxonomy, which is built for inquiring and responding biomedical questions, such as knowledge gap reorganization, question formulation, information retrieval and answer generation. The taxonomy is a simple hierarchy concept and its main objective is to impart the question processing engine in the clinical QA system by classifying questions according to the required answer which represents a unique answering system. The automatic Knowledge Discovery and Knowledge Reuse (KD–KR) process is used to improve the function of the template classifier by detecting the subject of the training questions and re-using the questions for the learning and prediction functions.

[5] *Improving MeSH classification of biomedical articles using citation contexts - Bader Aljaber, et. al.,*

This paper discusses about the importance of citation context in scientific research papers. The author uses intrinsic and extrinsic evaluation to improve the citation terms on UMLS conceptual database to find the synonyms of the citation terms and automatic classification of MeSH terms for MedLine documents over the citation terms. Finally they diagnose the result of document enrichment using citation terms and ontological terms which taken from the UMLS. This paper also establishes the statistically significant improvements in citation feature quality, which is a better document feature representation in biomedical text processing.

[6] *Ontology-driven indexing of public datasets for translational bioinformatics - Nigam H Shah, et. al.,*

This paper describes the prototype execution of an ontology-based annotation and indexing system. They evaluate the quality of ontology-based indexing scheme exactly identify experiments related to particular diseases and identify gene and expression datasets according to diseases. This method helps the researchers to determine relevant biological data sets for combination analyses.

4. Conclusion

This survey paper investigates the different automatic classification techniques like tf-idf, latent semantic indexing and MeSH classification for information extraction. The Chronic Illness Surveillance System is useful to prevent the chronic diseases by alarming the health professional about their risk factors, by retrieving relevant scientific research papers from the resource. The automatic classification techniques are used in a Question Answering system and Text classification which helps classify the document as text and to extract the research papers. The system enables the medical practitioners to find relevant biological information for the risk factors.

References

[1] Juliana Tarossi Pollettini, José Augusto Baranauskas, Evandro Seron Ruiz, Maria da Graça Pimentel and Alessandra Alaniz Macedo, “Surveillance for the Prevention of Chronic Diseases through Information Association”, BMC Medical Genomics, Vol. 7, No. 7, 2014, PMID: 24479447. http://www.biomedcentral.com/1755-8794/7/7.

[2] Athenikos S, Han H, “Biomedical question answering: A survey”, Comput Methods Programs Biomed, Vol. 99, No. 1, pp. 1-14, 2010, PMID: 19913938.

[3] Kolomiyets, Oleksandr Moens, Marie-Francine, “A survey on question answering technology from an information retrieval perspective” Information Sciences, Vol. 181, No. 24, pp. 5412-5434, ISSN: 0020-0255.

[4] Jon Patrick , Min Li, “An ontology for clinical questions about the contents of patient notes”, Journal of Biomedical Informatics, Vol. 45, No. 2, pp. 292–306, April 2012, PMID: 22142949.

[5] Bader Aljaber, David Martinez, Nicola Stokes, James Bailey, “Improving MeSH classification of biomedical articles using citation contexts”, Journal of Biomedical Informatics, Vol. 44, No. 5, pp. 881–896, October 2011, PMID: 21683802.

[6] Nigam H Shah, Clement Jonquet, Annie P Chiang, Atul J Butte, Rong Chen and Mark A Musen, “Ontology-driven indexing of public datasets for translational bioinformatics”, BMC Bioinformatics, Vol. 10, No. 2, February 2009, PMID: 19208184.

Authors

First Author – Ms. Anie Pearlin M, Pursuing M.Phil (CS), Phil Scholar, Department of Computer Science, Bishop Heber College (Autonomous), Trichirappalli, Tamil Nadu, India.

Second Author – Mrs. Thamarai Selvi R, Assistant Professor, Head of Computer Application, Bishop Heber College (Autonomous), Trichirappalli, Tamil Nadu, India.