HETEROGENOUS KNOWLEDGEABLE APPROACH FOR RETRIEVING MEDICAL DATA

M. Muralidharan
Research Scholar, Periyar Maniammai Institute of Science and Technology, Vallam, Thanjavur
J. Jeyachidra
Associate Professor, Department of Computer Science and Applications, Periyar Maniammai Institute of Science and Technology Vallam, Thanjavur,

Abstract: Now a days, hospitals and healthcare centers maintain the record of clinical and medical data of patients in electronic formats. Medical practitioners find easy to access these records for improving the quality of diagnosis and treatment. The medical dataset are growing at a fast rate and therefore, day by day it becomes difficult to search and get the right data from large data set. Moreover, medical terminology has vagueness and increases the complexity while seeking or retrieving. To settle this issue, we propose a syntactic based heterogeneous approach to retrieve the unique result from Electronic Medical Record (EMR). It has been done through two steps. The first step is query transfiguration, where Natural Language Processing and Medical Ontology transform the user query to key phrase sets. The second step is re-ranking of the initial key phrase query list, by considering both the relevance and originality of user query. This can be utilized as an important reference for the development of similar application in medical environment.

Keywords: EMRs, Information Retrieval, Query Transfiguration, Sub Query, Key Phrase, Medical Ontology, NLP.

Introduction: The interest towards automating the health record is excessive in both grown and growing countries. The motive for trying to change to an electronic system is important. People are concerned in moving from paper to paperless environments in healthcare. It is a foremost step and few healthcare centers have only been effectively implemented. They should no longer focus on simply going paperless. Healthcare Institutions must give attention on encouraging every department's medical practitioners to use towards an automated system that will enhance the correctness of the data logged in a record. Medical practitioners can get the details of patient's information in electronic format enabling it to be used by all for the current and preserving with care for the future and to enhance the satisfaction of care. Since health information is readily handy at any times for patient care, it is highly advisable to adapt to such transformation.

The Electronic Medical Record (EMR), is an automated arrangement based on record imaging or structures and has been documented inside a medical practice or public health center. It has been used appreciably by means of accepted practitioners in many developed international locations and consists of affected person identification details, drugs, prescription generation, laboratory results and in some case an over healthcare record is also recorded by the physician at each visit by the patient. In some nations, EMR is defined as an electronic record system inside a health care center that consists of medical information updated by the healthcare professional. People use internet for medical related searching. However, such search may not sure about the correction or accuracy of the result for the query. Search engine understands the Query in different ways. A single query from the user may have wide range of results. For example, a person is running temperature and has rashes throughout his/her body. He is not clear about the problem; he feeds the search engine input as "fever" and "rash". In this case, 'fever' may be symptoms for many diseases. It makes the user to worry or put himself in a confusing state. As a result, he/she is loaded with many questions for doctors. In another scenario, user may have a little medical learning, and feeds the search engine as "pain in abdomen" and "pregnant" as input query. Here pain is common term which means "stabbing pain", "distending pain" and "labor pain" etc,. All the results displayed would be entirely different. Therefore, the search engine would produce

answer for a set of possible results, which include all available chances of a query. Earlier information retrieval technique is classified in to two ways. They are text-based and syntactic based approaches. In text-based approach, query keyword is searched inside the document. In syntactic based approach, keyword in query and its meaning has been searched in the documents. So, the syntactic based approach provides better outcomes than content based approach. It produces the output of document with ranking independently. Even the best-ranked document contains more repetitive information. To satisfy the user needs, originality of the message is taken in to account for additional measure of document ranking. Traditional approaches have been classified as implicit and explicit approach and the latter is more viable than implicit approach. Explicit methods are mainly implemented in web search scenario, but this is not suitable for medical search setting. Web search engine uses query log and domain knowledge for its operation. Query log is unavailable, because it produces inappropriate output. Large range of domains are covered by website search engine for medical related search. It should have sound knowledge and medical domain, in order to show an appropriate result. There are some lack of approaches in existing domain knowledge. Consider a case, whose aspects are identified for user query. The first two core query matches with similar topics and the third one has new idea. Therefore, ranking comes in place, where similar topics are ranked and other irrelevant topics are removed from the top-ranked list. Our proposed approach is a combination of syntactic based and web search result technique for medical search.

This paper discuss about the following information: the proposed concept is medical information retrieval from the client input query. It is performed in two stages. Query understanding is the first step and that is done by Natural Language Processing and Medical Ontology and generating the multiple sub queries. In the following stage, numerous sub queries are checked for originality and record relevance, then shows the information based on re-ranking of derived sub queries. The proposed strategy efficiency is demonstrated in the real medical dataset reports. This study is formulated with the following sections, the previous similar works, the proposed methodology, the experimental evaluation results for the medical dataset and the conclusion.

Background Study: Medical information extraction is the process of finding the user query related to medical information from large set of data in EMR. Here the input is Layman query and the output is list of related documents or information based on original query. In Indian medical systems, Electronics Health Record is used for storing the patient full details from history, diagnosis, drugs, treatments, vaccination, allergy, scan image, and lab test report. This helps the medical practitioner for decision-making and diagnosis (second opinion). In [12], they discussed that EHR is implemented in cloud infrastructure for easy accessing and it is integrated with data from health care providers all over India. The issue facing here are data security, integrity and handling different type of data from different religion of people.

In [17] the author came up the semantic medical information retrieval. The major challenge here is effective use of the domain area knowledge in therapeutic knowledge database. It has two stages. Query extension technique is the primary stage where expansion is to improve the matching score between user query and documents. Second stage is to enhance the execution of data recovery with therapeutic information base. Some authors [7] implemented the visual and literary data recovery for the medical query via mobile application. Since mobile is handy, anybody can get to it any time and can permit the user to comprehend the difficult scenario in less complex and speedier way than the text retrieval.

Some authors [14], describe the concept mapping in medical field which means mapping the user query to its synonyms and extracts the natural language query for user. Synonyms are obtained from WordNet and UMLs Meta thesaurus. Generally, EMR does not contain the nursing reports. It contains the reports from medical practitioners alone. However, the nursing report are also important. In paper [21], they use the facultative analysis result of inpatient nursery and passage records. This discover the vocabularies related to treatment methods and summarized their results from the above-mentioned records.

Proposed Methodology: The proposed method for retrieving the medical or clinical information to done in two steps broadly. The first step is query transfiguration where the original user query

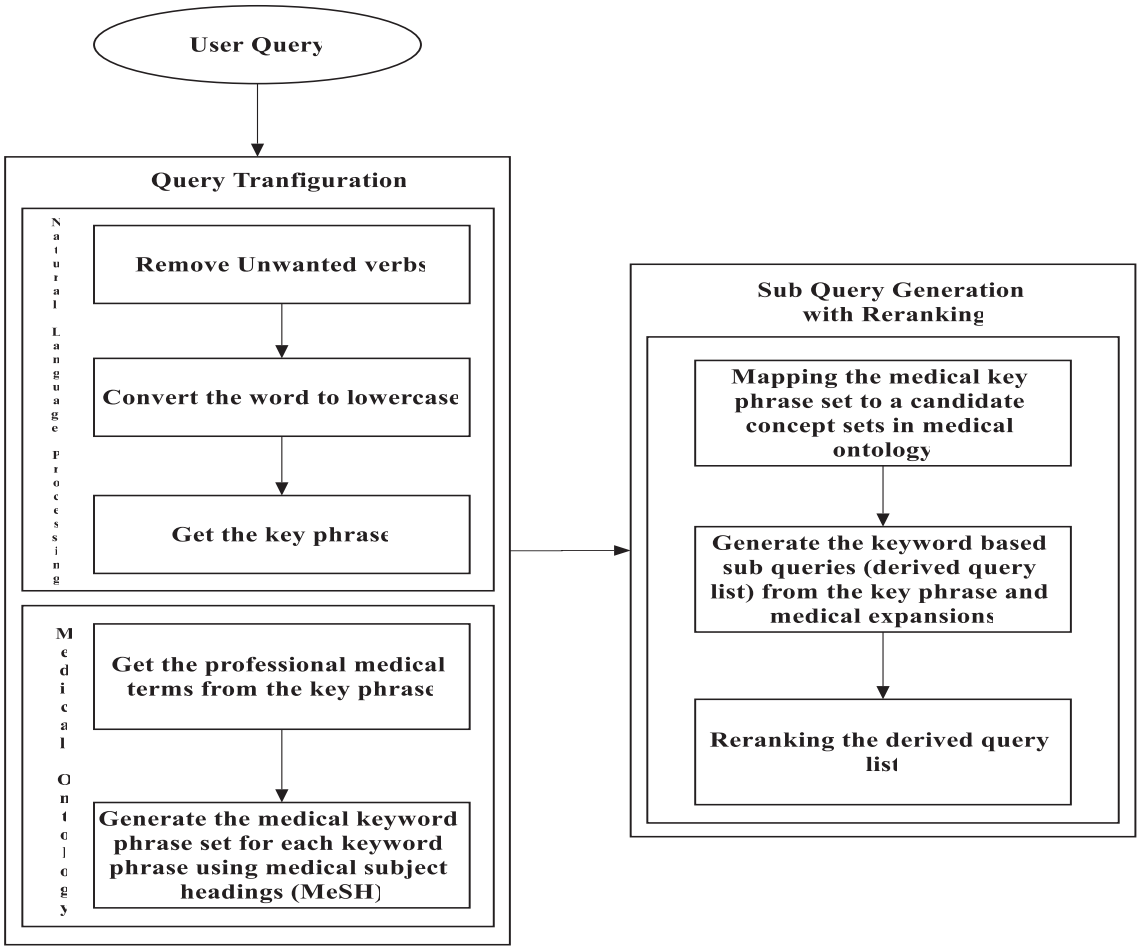
transformed to multiple sub query. It is achieved by Natural language processing and using medical ontology. The followed step re-ranks the derived query list and their document list. It has candidate mapping and derives the query list.

Query Transfiguration: Query transfiguration is achieved by NLP and medical ontology. The user or nonprofessional searches for the query. The queried input has to check for the superfluous verbs and punctuations. If the input query has the less important phrases, they can be removed the user input query can be changed into lower case letter. Scan the user query for getting the key phrase. It should be like longest maximum matching method to find the key phrase. The key phrase should be mapped to professional medical terms using (MeSH) Medical Subject Heading. It provides better search result when compared to the previous works.

Consider if a nonprofessional raises a query like, "I am having Frequent Urination and Thirsty", we have to remove the unwanted verbs and punctuation marks from user query. The query is modified to "Frequent Urination Thirsty" by removing the words 'I', 'am', 'having' and 'and'. Convert the changed user query to lower case letters as "frequent urination thirsty". Get the lengthiest key phrase from the query. Here "frequent urination" is a lengthiest key phrase and "thirsty" is another key phrase. Using medical subject heading, finding medical professional for the key phrases. For frequent urination, the professional term is "Polyuria" and thirsty is rephrased to "Polydipsia". With this query, transfiguration step is completed. To summarize, the key phrase should be expanded as follows.

$Key_i \in query = \{key_1, key_2, ...key_n\}$

$Keyset_i = \{keyphrase_{i1}, keyphrase_{i2}, ...key\ phrase_{in}\}$



Re-Ranking The Derived Sub Queries: The first sub step is the query concept mapping: Map the user key phrase from the key set to a set of query concept with ranking scores using medical ontology.

$Key_i \epsilon$ query= $\{key_1, key_2,.....key_n\}$
$Concepts_i$= $\{con_{i1}, con_{i2},.....con_{in}\}$
With ranking scores
$Rank_i$= $\{r_{i1}, r_{i2},.....r_{in}\}$
The condition for selecting the candidate with ranking score is greater than 0.75., where this value is opted for balancing the search accuracy and finding the complexity of the derived sub queries.
To model the various aspects of query, construct a derived query list. With the help of Medical Ontology, sub-graph list is extracted here and every sub-graph covers minimum one concept for every key phrase.
$Keyphrase_i \epsilon$ query= $\{keyphrase_{i1}$, key $phrase_{i2}$, ... $keyphrase_{in}\}$
Sub queries are generated from each sub graph representing the interpretation of the user search information.
Consider each key phrase represents a concept by the user, then theoretically
Concept combination $X = \prod_{m}^{i-1} n_i$ for query Qn_i represents the number of concepts and m the number of key phrase in Q.

A sub graph set is extracted from the medical ontology for each concept combination. Each sub-graph (SG_i) denotes the possible definition of query Q and it has graph weight score for the user into need.
Sub-graph (SG_i) ranking is performed based on the weight score for all X concept combinations. Sub-graph ranking is based on the weight value. The sub-graph whose weighted score is larger than threshold value then, it is added to top ranked sub-graphs. An investigation has been directed to test the diverse threshold values, which will result in our proposed approach to achieve the best on all the retrieval when the threshold esteem is 0.75. The key-based sub query can be generated from the sub-graph whose threshold esteem is 0.7 which contains the key phrase in the user input query Q, and key phrase in concept nodes covered by sub-graphs. Score of the corresponding sub query derive the derived sub-graphs weight scores.

To discover the re-ranking (RR) that has extreme coverage and minimum redundancy got from the original query initial ranking list R, the local best document has extreme coverage of user input query and least redundancy to aspects covered by the queries initial ranking R and re-ranking RR. Our proposed search depends on this approximation idea.

The problem definition is as follows:m Initial user input query is termed as Q and derived query list is termed as DQ. A document list for each query and each derived query has been produced. A ranking model used here may be a vector space model, probabilistic rank model or legency based model, rank model, suitable for relevance computing. So, the base ranking is a biopic list for the initial Query and derived ranking is generated for derived queries. The search results could be re-ranking RR of the based ranking that depends on derived query list Q and their document list.
The top first position of re-ranking list is taken from top first of base ranking list, since this is more relevant to the query. The next best document that is local-best is consecutively selected.

Algorithm: Sub Query Generation:
Input: a combination of concept
$Concept_i$= $\{con_{i1}. Con_{i2},.....con_{in}\}$,
Medical Ontology MO.
Output: SGQ= {(Query, Weight),.....} with weight ϵ |0,1|
Start
1. Initialize a graph vector V and Edge E
V=ϕ, E=ϕ
2. Concept mapping initialization
S_{con}= $\{con_1, con_2,.....con_n\}$
$S_{initial}$=ϕ
3. Assign value to $S_{initial}$ from choosing a concept from S_{con} randomly
4. While the $S_{initial}$ is not equal to concepts

5. Assign length as 1 and S_{length} as null
6. For each concept in $S_{initial}$

Do
Calculate the S_{length} value
If the S_{length} is null then increment the length value by 1
While till the S_{length} is null and length len than max value

7. Concept mapping is done by graph construction assuming 'E' edge comprises concepts
8. Construct the sub-query graph generation

SGQ= {(query, weight)}
Where weight is ranking score and
Query is generated from edges in concept mapping
End

Algorithm steps 1-3 assign the concept mapping with graph theory concept. Initial assignment of vertex, edge and mapping initialization as null and choose a random concept assign to $S_{concept}$. In step 4-5, Depth First Search (DFS) is performed for fin**ding the concepts. In step 6, concept mapping is completed with the help of graph construction (simulating concept mapping by graph generation).In step 7, the graph assigned with weight. Here weight is assigned based on the ranking of the query. Based on this ranking, a sub graph or sub query is generated and the top weighted graph is the top most sub query and the remains, nodes are ranked in the same manner. After construction of sub query graph. We get a derived query list with ranking. Query transfiguration and syntactic web based domain knowledge are illustrated in the Table 1 below

Heterogeneous Knowledge Search: Our proposed heterogeneous knowledge search generates the derived sub query list with re-ranking from the sub query with the base ranking. The first ranked document in base ranking query list is opted for first ranked document in re-ranking query list which becomes the closest to the user query. The remaining query list is ranked in selecting the next local-best document as that of the second re-ranked query. The following will be used to calculate the local-best document.

Local-best document= Tradeoff Factor X Matching (document,query) + (1- Tradeoff). Originality of query

Here, Tradeoff factor belongs to either 0 or 1 which means it lies between matching and originality factors. The originality of the document is calculated from whether the user input query is satisfied by the derived query list.

Table 1: Expansion Presentation Using Heterogeneous Knowledgeable Approach

Given Query	Query Conversion	Candidate Concept Toning	Substitute Query Generation
Frequent Urination Thirsty	Two Terms are Identified 1. Frequent Urination 2. Thirsty Terms Expansion # Frequent Urination (Same as PolyUrea) # Thirsty (Same as Polydipsia)	concepts are notorious for #1: Biopolymers [D05.750.078] "Colestipol [D05.750.200]" "Cyanoacrylates [D05.750.259]" Dendrimers [D05.750.327] "Fluorocarbon Polymers [D05.750.395]" "Hexadimethrine Bromide [D05.750.470]" "Organically Modified Ceramics [D05.750.593]" "Plastics [D05.750.716]" Polyanetholesulfonate [D05.750.722] Polyanhydrides [D05.750.725] Polyesters [D05.750.728] "Polyethylene Glycols [D05.750.741]"	3 substitute queries was generated Diabetics mellitus Paroxysmal, nocturnal Kidney failure

		"Polygeline [D05.750.780]" "Polyphloretin Phosphate [D05.750.795]" "Pyran Copolymer [D05.750.830]" "Siloxanes [D05.750.900]" concepts are notorious for #2: "Aging, Premature [C23.888.069]" "Asthenia [C23.888.089]" Body Temperature Changes [C23.888.119] "Body Weight [C23.888.144]" "Cardiac Output High [C23.888.176]" "Cardiac Output Low [C23.888.192]" Chillss [C23.888.208] Cyanosiss [C23.888.248] Eedema [C23.888.277] "Eye Manifestations [C23.888.307]" Fail to Thrive [C23.888.338] "Fatigue [C23.888.369]" Feminization [C23.888.378] Fetal Distress [C23.888.380]	

Experimental Evaluation: Existing explicit search retrieval approaches are performing better than the implicit search method. Here, we are comparing the proposed approach with the explicit search methods like 1A-select, xQuad with combination of LAG (log based technique). The following are the outcome of this comparison :

(i) The proposed approach has natural Language Processing, medical ontology and re-ranking the derived query list based on matching and originality factors. The resulted output satisfies the user query.

(ii) Medical practitioners feel easy to access this search engine that ensures helping to enhance the health care quality.

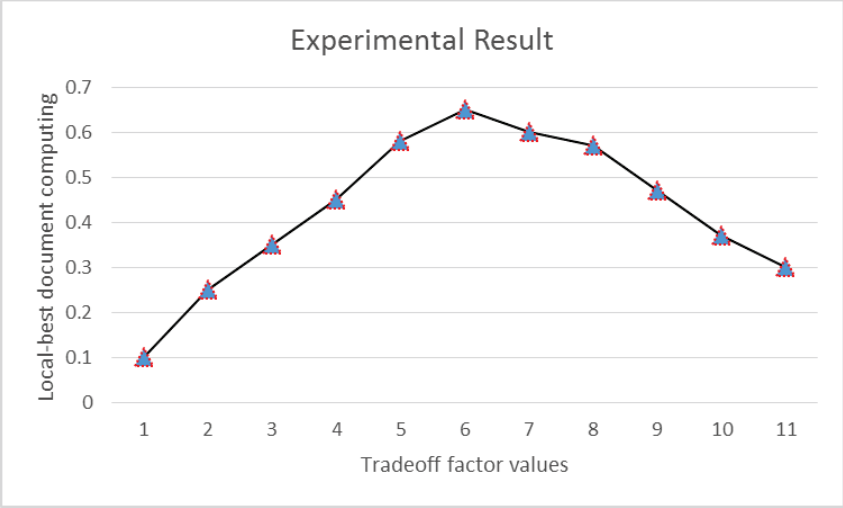


Fig: 2: Experimental Result For Balancing The Matching And Originality Of User Query

The graph represents the Tradeoff factor between originality and matching factor. The local-best document computing, on the medical dataset for choosing the correct value of Tradeoff factor which is between 0 and 1, may be, 0.1, 0.2..... till 1.When the Tradeoff factor value sets to 0.6, it achieves a peak value in local-best document choosing in medical data set.

Conclusion: This paper combines the technologies of syntactic based approach and website search engine domain knowledge understanding method for the fine retrieval of data from the electronic medical records. The implementation is done in two steps. The initial one is query transfiguration. It includes natural language processing and medical ontology to modify the user input query to many sub queries. The following step is re-ranking the derived queries where the sub queries with base ranking undergoes concept mapping and choose the local-best document as top ranked derived query and remaining queries are re-ranked with a similar procedure. The results are matched with the original user query requirement. Since, it has a medical ontology for synonym collection and ranking mechanism in it. The results are more precise and effective when compared with other existing methods. A sample study has been done with the medical data set. In future, we can extend this implementation for cloud data and can utilize in a mobile environment for handy use.

References:

1. B.Xu, J.Bu, C.Chen, C.Wang, D.Cai, X.He, "EMR: A Scalable Graph-Based Ranking Model for Content-Based Image Retrieval", IEEE Trans. Knowledge and Data Engineering, Vol.27, Issue.1, Jan.2015, PP.102 – 114.
2. Z.Zhang, B.Wang, F.Ahmed, I.V.Ramakrishnan, R.Zhao, A.Viccellio, K.Mueller, "The Five Ws for Information Visualization with Application to Healthcare Informatics", IEEE Trans. Visualization and Computer Graphics, Vol.19, Issue.11, Nov. 2013, PP.1895 – 1910.
3. M.Lesk, "Electronic Medical Records: Confidentiality, Care, and Epidemiology", IEEE Security & Privacy, Vol.11, Issue.6, Nov.-Dec. 2013, PP.19 – 24.
4. A.Tamersoy, G.Loukides, M.E.Nergiz, Y.Saygin, B.Malin, "Anonymization of Longitudinal Electronic Medical Records", IEEE Trans. Information Technology in Biomedicine, Vol.16, Issue.3, May 2012, PP.413 – 423.
5. D.C.Leonard, A.P.Pons, S.S. Asfour, "Realization of a Universal Patient Identifier for Electronic Medical Records through Biometric Technology", IEEE Trans. Information Technology in Biomedicine, Vol.13, Issue.4, July 2009, PP.494 – 500.
6. B.Y.Kang, D.W.Kim, H.G.Kim, "Two-Phase Chief Complaint Mapping to the UMLS Metathesaurus in Korean Electronic Medical Records", IEEE Trans. Information Technology in Biomedicine, Vol.13, Issue.1, Jan. 2009, PP.78 – 86.
7. S.Duca, A.Depeursingea, I.Eggela, H.Mullerab, "Mobile Medical Image Retrieval", Proceedings of the SPIE, Volume 7967, 2011.
8. E.S.Hall, D.K.Vawdrey, C.D.Knutson, J.K.Archibald, "Enabling remote access to personal electronic medical records", IEEE Engineering in Medicine and Biology Magazine, Vol.22, Issue.3, May-June 2003, PP.133 – 139.
9. R.S.Ledley, L.B.Lusted, "The Use of Electronic Computers in Medical Data Processing: Aids in Diagnosis, Current Information Retrieval, and Medical Record Keeping", IRE Trans. Medical Electronics, Vol.ME-7, Issue.1, Jan.1960, PP.31 – 47.
10. H.Zhu, D.Liu, I.Bayley, A.Aldea, Y.Yang, Y.Chen, "Quality model and metrics of ontology for semantic descriptions of web services", Tsinghua Science and Technology, Vol.22, Issue.3, June 2017, PP.254 – 272.
11. H.A.Mubaid, H.A.Nguyen, "Measuring Semantic Similarity between Biomedical Concepts within Multiple Ontologies", IEEE Trans. Systems, Man, and Cybernetics, Part C (Applications and Reviews), Vol.39, Issue.4, July 2009, PP.389 – 398.
12. B.Cheng, S.Zhong, J.Chen, "Context Ontology-Based Reasoning Service for Multimedia Conferencing Process Intelligence", IEEE Trans. Systems, Man, and Cybernetics: Systems, Vol.47, Issue.12, Dec. 2017, PP.3219 – 3232.
13. R.Kavitha, E.Kannan and S.Kotteswaran, "Implementation of Cloud based Electronic Health Record (EHR) for Indian Healthcare Needs", Indian Journal of Science and Technology, Vol.9, Issue.3, January 2016.
14. G.Leroy, K.M.Tolle, H.Chen, "Customizable and Ontology-Enhanced Medical Information Retrieval Interfaces", UA Campus Repository, 1999.

15. S.Bandyopadhyay, K.Mallick, "A New Feature Vector Based on Gene Ontology Terms for Protein-Protein Interaction Prediction", IEEE/ACM Trans. Computational Biology and Bioinformatics, Vol.14, Issue.4, July 2017, PP.762 – 770.
16. A.Urrutia, E.Chavez, R.Motz, R.Gajardo, "An Ontology to Assess Data Quality Domains. A Case Study Applied to a Health Care Entity", IEEE Latin America Transactions, Vol.15, Issue.8, 2017, PP.1506 – 1512.
17. A.Rosier, P.Mabo, M.Chauvin, A.Burgun, "An Ontology-Based Annotation of Cardiac Implantable Electronic Devices to Detect Therapy Changes in a National Registry", IEEE Journal of Biomedical and Health Informatics, Vol.19, Issue.3, May 2015, PP.971 – 978.
18. N.Lasierra, Á,Alesanco, J.García, "Designing an Architecture for Monitoring Patients at Home: Ontologies and Web Services for Clinical and Technical Management Integration", IEEE Journal of Biomedical and Health Informatics, Vol.18, Issue.3, May 2014, PP.896 – 906.
19. H.Wang, Q.Zhang, J.Yuan, "Semantically Enhanced Medical Information Retrieval System: A Tensor Factorization Based Approach", IEEE Access, Vol.5, April.2017, PP.7584 – 7593.
20. J.Hirschberg, B.W.Ballard, D.Hindle, "Natural language processing", AT&T Technical Journal, Vol.67, Issue.1, January-February 1988, PP.41 – 57.
21. Muneo Kushima, Member, IAENG, Kenji Araki, Muneou Suzuki, Sanae Araki and Terue Nikama, "Text Data Mining of the Electronic Medical Record of the Chronic Hepatitis Patient", IMECS, Vol.I, March.2012.
22. R.K.Taira, V.Bashyam, H.Kangarloo, "A Field Theoretical Approach to Medical Natural Language Processing", IEEE Trans. Information Technology in Biomedicine, Vol.11, Issue.4, July 2007, PP.364 – 375.
23. P.G.Anick, "Integrating natural language processing and information retrieval in a troubleshooting help desk", IEEE Expert, Vol.8, Issue.6, Dec. 1993, PP.9 – 17.
24. M.P.Barnett, W.M.Ruhsam, "A Natural Language Programming System for Text Processing", IEEE Trans. Engineering Writing and Speech, Vol.11, Issue.2, Aug. 1968, PP.45 – 52.
25. G.Leroy, H.Chen, "Meeting medical terminology needs-the ontology-enhanced Medical Concept Mapper", IEEE Trans. Information Technology in Biomedicine, Vol.5, Issue.4, Dec. 2001, PP.261 – 270.
26. H.L.Tang, R.Hanka, H.H.S.Ip, "Histological image retrieval based on semantic content analysis", IEEE Transactions on Information Technology in Biomedicine, Vol.7, Issue.1, March 2003, PP.26 – 36.
27. Y.Wang, P.F.Li, Y.Tian, J.J.Ren, J.S.Li, "A Shared Decision-Making System for Diabetes Medication Choice Utilizing Electronic Health Record Data", IEEE Journal of Biomedical and Health Informatics, Vol.21, Issue.5, Sept.2017, 1280 – 1287.
28. J.Stojanovic, D.Gligorijevic, V.Rado savljevic, N.Djuric, M.Grbovic, Z.Obradovic, "Modeling - Healthcare Quality via Compact Representations of Electronic Health Records", IEEE/ACM Trans. Computational Biology and Bioinformatics, Vol.14, Issue.3, June 2017, PP.545 – 554.

***