BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **18**, No 1
Sofia • 2018 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2018-0014

A Semantic Multi-Field Clinical Search for Patient Medical Records

E Umamaheswari Vasanthakumar, Francis Bond

Linguistics and Multilingual Studies, School of Humanities, Nanyang Technological University, Singapore
E-mails: vasanthuma28@gmail.com fcbond@gmail.com

***Abstract*:** *A semantic-based search engine for clinical data would be a substantial aid for hospitals to provide support for clinical practitioners. Since electronic medical records of patients contain a variety of information, there is a need to extract meaningful patterns from the Patient Medical Records* (*PMR*). *The proposed work matches patients to relevant clinical practice guidelines* (*CPGs*) *by matching their medical records with the CPGs. However in both PMR and CPG, the information pertaining to symptoms, diseases, diagnosis procedures and medicines is not structured and there is a need to pre-process and index the information in a meaningful way. In order to reduce manual effort to match to the clinical guidelines, this work automatically extracts the clinical guidelines from the PDF documents using a set of regular expression rules and indexes them with a multi-field index using Lucene. We have attempted a multi-field Lucene search and ontology-based advanced search, where the PMR is mapped to SNOMED core subset to find the important concepts. We found that the ontology-based search engine gave more meaningful results for specific queries when compared to term based search.*

***Keywords*:** *Semantic similarity, application to NLP, SNOMED ontology, information extraction and text simplification.*

1. Introduction

Nowadays clinical Information Retrieval (IR) based systems have become an active research area which deals with diverse collection of medical resources that contains information about hospital records of patients, previous medical history, research articles, social media and the web to support general users and medical practitioners. However only limited clinical search engines are

available and most of them are specific to certain diseases. The writing style of the Patient Medical Records (PMR) varies from one medical practitioner to the other. Short forms and abbreviations need to be handled carefully, in order to correctly identify the disease of a patient.

Clinical Practice Guidelines (CPGs), on the other hand, are designed to guide clinical practitioners in making decisions regarding diagnosis, management, and treatment in healthcare. Each guideline consists of two parts: conditions and instructions, such that if all the conditions of a guideline are met for a patient, the doctor should follow the instructions in the guideline. Many governments provide and maintain CPGs for their own citizens. The CPGs are continuously updated to reflect recent knowledge discovery or disease outbreaks. As a result, doctors may not know all the available CPGs, particularly outside their specialties. The goal of this work is to aid doctors in identifying the appropriate CPGs based on a patient's electronic medical records.

B o d e n r e i d e r [2] have argued that context based clinical summaries will reduce the time for clinicians to diagnose a particular patient's problems. The reason is that CPGs contain much information spread over many documents. Finding the best guidelines for particular symptoms and medical history in the corpus with diverse collection of documents leads to barrier in performance and accuracy. A feasible approach to reduce these barriers is automatic summarization of multiple sources in the context of a particular information need of a clinician.

In this paper, we analyze both the CPGs and PMRs in order to retrieve relevant results for given patient information with limited manual effort. Since the patient medical records contain properties such as symptoms/complaints, previous medical history, discharge-instructions and medication, the search interface is designed with multiple fields to search and similarly the indexing mechanism also aids multi-field search and ranking. Ontology based query processing has also been proposed to further reduce the unimportant terms in the query processing stage and to find semantically relevant clinical guidelines.

This work is different from a general purpose IR system in that it provides specialized document processing tasks that automate XML based CPG representation and stored them as multiple fields in Lucene Index including domain specific ontologies. Hence the offline process considerably reduced the manual efforts and the search with the index is highly efficient in terms of time and computation effort. As explained earlier the clinical notes of PMRs are unlimited in size, do not follow standard representation, short form texts with medical abbreviations and more symbols. Most of the existing clinical search system [10, 1, 13] face a major challenge in handling such queries and the computation efforts are more to find the most appropriate result for the given PMRs. The query filtering mechanism used in this proposed work aids in filtering unimportant terms and the ontology mapping helps to find semantically relevant terms.

This paper is organized as follows. Section 2 discusses the related work in clinical search and rank. Section 3 focuses on the methodology used to develop this multi-field clinical search and rank. Section 4 shows the test reports of this proposed work and Section 5 ends with conclusions and future work.

2. Related work

Jonnalagadda et al. [14] proposed a semantic based information retrieval system to extract relevant sentences from Medline abstracts automatically. They have tested their results for depression and Alzheimer diseases. Kilicoglu et al. [10] have proposed a rule based clinical decision making system based on Unified Medical Language System (UMLS). [2] represents the document semantics as a SemMedDB [10] repository using Subject-Object-Predicate triplets. UMLS concepts and associations have also been used with SemRep [10], to predict the semantics of the documents [2]. Here they have attempted Google PageRank based Text Ranking [1] an algorithm which computes the semantic similarity based on the number of conceptual links between the query sentence with the document sentence.

Oh, Jung and Kim [11] proposed a multi-stage re-ranking method for clinical documents to search relevant documents based on different ranking parameters in different levels. The main objective of this work [11] is to improve the ranking by analyzing the initially retrieved documents by computing score between the query and the documents. The levels of ranking are query expansion with abbreviations and discharge summary, cluster, centrality and Pseudo Relevance Feedback. This work [11] utilizes various existing techniques, instead of using external semantic resources to retrieve the relevant documents. Oh, Jung and Kim [11] proved that abbreviation and discharge summary based query expansion can improve the relevance of ranking even in the absence of specialized semantic representations and techniques. Similar work has been attempted by Zhu et al. [7] using Markov Random Field (MRF) [6], Mixture Reference Model (MRM) [9] improving, and MeSH-based[8] improving query expansion. They have used several external resources and an open source clinical NLP annotation tool called MedTagger to extract the contextual information from the PMR. Diaz and Metzler [9] recently processed a SPUD language model [4] clinical to support clinical decision making. Here the documents are modelled by finding a word's burstiness (analysis of the behavior of an uncommon word which may appear many times in a single document) by identifying the dependencies between recurrences of the same word-type. They proved that this model is suitable to process scientific texts. All the above models require either a sophisticated language model or external resources to build a clinical decision system. The work described in this paper requires less computational effort by properly segmenting the clinical documents using regular expression based approach and a lightweight ontology based concept mapping during query processing stage that helps to retrieve semantically similar documents even though the exact terms of the PMR are not present in the clinical document index. The Lucene-based multi-field indexer [16] helps to retrieve the clinical guidelines in multiple perspectives.

3. Methodology

In this work, we have developed a solution for identifying the semantics of PMR and properly extracting the CPGs for semantic matching between the two clinical texts in order to recommend relevant CPGs for a given Patient Medical Records (PMR). In this section we present the methods used for searching and ranking the Clinical Practical Guidelines (CPGs). We have attempted two types of search such as Basic Term based and Concept based search using ontology.

3.1. Dataset

We have used the Singapore CPG documents of Dental, Medical, Nursing and Pharmacy of 72 documents, available from Ministry of Health, Singapore (2016) (**https://www.moh.gov.sg/content/moh_web/heathprofessionalsportal/doctors/guidelines/cpg_medical.html**). There are 124.2 MB in all. We have used 200 Patient Medical Records obtained through cooperation with Khoo Teck Puat Hospital (**https://www.ktph.com.sg/main/home**) for developing this system. The CPGs are publicly available, while the PMRs are confidential.

3.2. Document processing and indexing

Since the CPG are PDF documents, we need to extract the text content and the important information related to the clinical guidelines with the help of regular expression based PDFMiner (**https://euske.github.io/pdfminer/**) using Python. The guidelines are represented in XML as shown in Example 1. Hence each guideline is associated with the list of properties such as Guidelines Category, Filename, Topic, Year, ID, Session Page Number, Session Title, Classification grade, Level and Full String. In order to identify these properties from the PDF documents, we have used regular expression rules. The properties such as Guidelines Category, Filename and Session Page Number have been identified easily using documents folder, filename and page number respectively. Topic, Year and ID can be extracted with the help of Python PDFparser which is a built-in pdfminer. The other properties such as guideline Category, Session Page Number, Classification Grade, Level and Full string are identified using the regular expression rules as explained below. Regular expressions allow users to create complicated queries and have potential uses in document annotations.

Regular expression Rules for Identifying CPG fields:

1. **Reference Page Number:** Find expressions that contains pg \d(1, 2).
2. **Session Title:** Remove stop words from the sentence that appear before the Full String pattern and check whether the first letter of all words are capital ^[A-Z].
3. **Classification Grade:** Text with Grade [A-D] patterns.
4. **Level:** Text with Level [1-9] [\+]* patterns.
5. **Full String:** Find the sentences that starts with a single capital letter between A to D that belong to the regular expression (^ [A-D]) and ends with the pattern (Grade [A-D], Level [1-9][\+]*). The regular expression that are used to separating sentence with "." as delimiter is *[\.][\'"\)\]]* *.

Example 1. Guidelines Representation in XML

<guideline Category="Medical" FileName="cpg_Osteoporosis Summary Card-Jan 2009" Topic="Osteoporosis" Year="2009" id="164"><Session PageNumber="1" ReferencePage="17" SessionTitle="Clinical risk evaluation"/> <Classification Grade="C" Level="2+"/><FullString>Women identified as high risk using the Osteoporosis Self-Assessment Tool for Asians, should be recommended for bone mineral density measurement (see Table 3A and 3B)...</fullstring></guideline>

After converting the CPGs into an XML file format as shown in Example 1, the fields of guidelines including a Score (Lucene Default Score for guidelines) are indexed with the help of multifield Lucene index [16]. The Lucene Index stores all the fields of the guidelines and user can search the guidelines in different perspective with a reasonable time limit.

3.3. SNOMED Index

UMLS (Unified Medical Language System) [12] (NLM 2016) has a vocabulary with comprehensive coverage of biomedical terms, and is linked to a semantic classification system, called SNOMED [12] Clinical Terms (SNOMED_CT), which will be used as the basis for the semantic representation of the Patient Medical Records (PMR). In order to aid the PMR to find the semantically relevant CPGs, we have used the recent version of the SNOMED_CORE_SUBSET (**https://www.nlm.nih.gov/healthit/snomedct/**) with 6,358 ontological concepts. These concepts are also indexed in the Lucene multifield indexer to aid in effective search.

The fields are SNOMED_CID, SNOMED_FSN, SNOMED _CONCEPT_STATUS, UMLS_CUI, OCCURRENCE, USAGE, FIRST_IN_SUBSET, IS_RETIRED_FROM_SUBSET, LAST_IN_SUBSET, REPLACED_BY_SNOMED_CID. Here <SNOMED_FSN> field is used to retrieve the concepts from the ontology. Both the SNOMED and guidelines index helps in retrieving CPGs effectively.

3.4. Searching and ranking

This module consists of three major tasks: Query processing, Searching and Ranking. The queries are constructed from the Patient Medical Records (which are already converted to XML). The PMR queries in XML contain symptoms/complaints, previous medical history, discharge-instructions and ordered medication fields. Each field contains a description corresponding to the patients. Hence the main challenge is in dealing with the queries with multiple fields and keywords. There is no standard way to describe a patient condition and extracting common properties of patient medical records are difficult. Moreover these records contain lots of abbreviations, short forms and errors that belong to medical tests, diseases and patient medical conditions, etc.

We first removed stop words, words of length 2 and symbols from the PMR. We then used the Stanford English Parser [5] to identify NPs (Noun Phrases). The resulting pre-processed PMR contains a list of NPs which helps in extracting the

ontological concepts. At present the ontology matching is between the Fully Specified Names (FSN) with the NPs of the PMR. The top most concepts are taken as expanded concept for each NP of the PMR. As a result of query expansion, a set of expanded concepts with the terms of PMRs are considered for searching and matching. This is explained in the next section.

3.4.1. Search and Rank Algorithm

The PMRs are full of incomplete sentences (or fragments) without proper usage of full stops. We need to split the PMRs into fragments (e.g., noun phrase, verb phrase, and sentence) which will be used as inputs to the search module. The PMR terms are filtered by considering only NPs and VPs with the help of Stanford Parser [5] and the semantically similar terms are extracted from the SNOMED_CORE_SUBSET. Here PMR query terms represents the list of terms received from the PMR after symbol and stop words removal. SNOMED_CORE_SUBSET is a Lucene index that contains list of ontological concepts as explained in Section 3.3. CPGs contain all properties related to the guidelines as described in Section 3.3. Fig. 1 shows the search and matching procedure. PMR Category contains fields such as Complaints, Previous Medical History, Discharge Instruction and Ordered Medications.

Algorithm 1: SEARCH and Rank Algoritm

Input: A PMR query concept set $C = \{c_1, c_2, \ldots, c_r\}$ of PMR_qt_r where r varies from 1 to n, that contains $ontoCon\&nonOntoCon \in C$

Output: A list of CPGs $cpg_1, cpg_2, \ldots, cpg_k$, such that $\sum_{i=1}^{k} cpg_i \in PMR\_qt_r$

```
cpg_i ← ∅
for i ← 1 to n do
    while c_i! = null do
        CPGs ← findmatch(c_i, cpg_i) if CPGs! = NULL then
            PMR_Category ← findPMRCategory(c_i)
            Word_Match_Count ← ComputeWMC(CPGs, c_i)
            CPGs ← RankCPGs(PMR_Category,
                            Word_Match_Count)
return CPGs
```

Fig. 1. Algorithm for searching and ranking

These fields are fixed templates help in searching the CPGs with different granularity. The SNOMED_CORE_SUBSET are not guaranteed to give expanded concepts for all the terms of the PMRs. Hence it is better to treat query terms with and without ontological expansion. Among these two categories, the query with ontological concepts is given higher preference in ranking.

The search query is a single patient's record (not the entire records of all patients). The number of terms in the query is unlimited in size. The query expansion part is online. The PMR contains much information pertaining to the patients and has on average slightly over 20 terms for each field. Hence there is a need to extract only the important information from the query to reduce the search

time. Since Stanford parser yields good performance even for ungrammatical text, we have used that to find the NPs and VPs of the given query PMR. The SNOMED_CORE_SUBSET is used to find the ontological concepts which help us to retrieve conceptual results.

Step 1. Remove stop words, symbols and extract NPs with the help of Stanford Parser.

Step 2. Match NPs with SNOMED_CORE_SUBSET and find the top most expanded terms for each terms of the given PMR.

Step 3. Separate the query with and without expansion.

The PMR Category indicates the different level of search (Complaints, Previous Medical History, Discharge Instruction and Ordered Medications, Overall) in CPG Index. These categories are identified from the query interface and used in ranking the results. The example is shown below.

The results are ranked based on the following levels:

Level 1. Query Concept Association with ontology.

Level 2. Patient Information (Specific and Overall Fields).

3.4.2. Level 1. Query concept association with ontology

This means that the concept C_i is associated with the term word used by the PMR query or expanded terms obtained from the SNOMED ontology. Here the Boolean variables ontoCon and nonOntoCon are used to identify whether the query terms are associated with or without ontological concepts. Among these two categories, the results of the query with ontological concepts are given higher preference in ranking. The expanded ontology concepts are shown in Fig. 2.

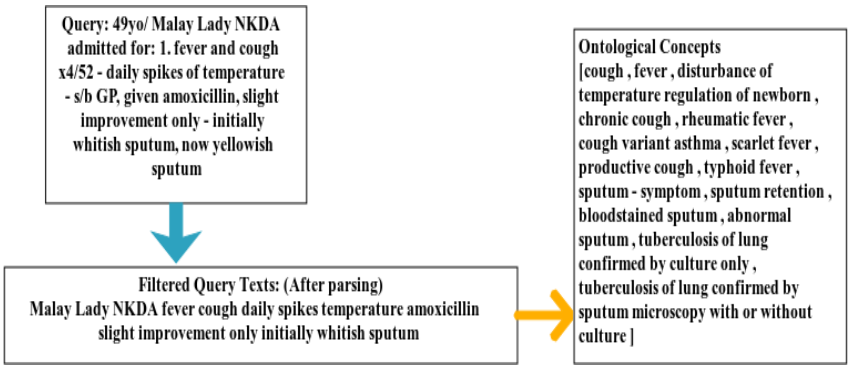

Fig. 2. Query Expansion using SNOMED_CT

As explained in Section 3.3, the SNOMED Clinical Terms are indexed in Lucene index. The expanded ontological concepts are obtained by Lucene multi-term search with the query terms in the index field SNOMED Fully Specified Names (SNOMED_FSN). The Lucene default score helps to rank the retrieved ontological clinical terms and the topmost ontological terms are considered. In the given example ontology-based query expansion helped to get “Signs and symptoms

of pneumonia when changes in sputum color" and "prevention and diagnosis and management of Tuberculosis" results within first two hits.

3.4.3. Level 2. Patient Information (Specific and Overall Fields)

Here we have divided the search into specific and overall. In Level 2, the results are ranked based on the user preferred fields of the patient information. The users of this search system may want to search CPGs based on any fields related to the PMR. Each field of PMR can be searched specifically with one field or more than one field. The Level 1 ranked results are maintained the same as in Level 2 by separating the term based results and expanded concept based results. The Word Match Count computes the number of common terms between PMR and CPG. This score helps to rank the CPGs based on the maximum number of match between PMR and CPG at each PMR search category (PMR Category Search) which will be considered at the end of Level 1 and Level 2 ranking. If the search is a single specific field, then the Level 1 ranking is maintained and within each category of Level 1, the results are ranked based on the Word Match Count score. If the search query contains more than one field then a priority rank tag is set from 1 to 3 for Patient Complaints, Previous Medical History, Ordered medications and discharge instruction. The rank tag for Patient Information based search (Level 2) consists of Patient Complaints (PCom Match), Previous Medical History (PMH Match) and Ordered Medications and Discharge Instruction (OMDI Match).

The results are ranked in ascending order based on the rank tag when user searches with more than one field. The rank tags are assigned by experimenting sample results and we found that patient complaints based search mostly covers the diseases and symptoms; we gave value 1 in ranking. Previous Medical History provides additional information to the clinical practitioners, we have set the ranking value to 2 and similarly Ordered medications and discharge instruction also helps in finding the most suitable diagnosis to the patients hence it is set as 3. If there is no overlapping across the fields, the results are ranked in ascending order from 1 to 3, else the resulted CPGs from more than one field is given higher weightage in ranking. When the documents were tested with a small set of guidelines (nearly 500) for 50 PMR queries, we found the results obtained from PCom Match and PMH are more relevant when compared to OMDI match. This test is the empirical basis for the weights assigned in our work.

4. Result evaluation

This CPG search and rank method differs in considering various levels of search based on PMR and the indexing mechanism is also designed to aid the multi-field search. We have evaluated this search engine for a dataset of 3001 CPGs and 200 PMRs. The Precision Score [3] is computed at precision P@5 and P@10 to prove the relevance of the results. Since this method supports different levels of search and rank method, the results are also analyzed by separating them in different category.

4.1. Baseline

Since there is no standard state of art approach that uses Singapore Clinical guidelines as Dataset for comparing our approach with the existing work, the basic Lucene search framework has been considered as baseline. Apache Lucene is the most widely used search engine for term based search and indexes the documents similar to our work. The proposed work is compared by disabling the proposed ontology based search. The results are shown in Table 2.

In case of clinical search, the clinicians may not have time to look at large set of results and for them it is important to have relevant results in the top 10. In such cases P@5 and P@10 is appropriate measure. Hence we gave importance to P@5 and P@10. The Precision score have been computed for the following categories: Overall (Contains all PMR fields); Complaints based Results; Previous Medical History based Results; Discharge Instruction and Ordered Medications based Results. The results are given in Table 1. We have also tested the results with ontology and without ontology-based expansions. The results are shown in Table 2. A manual rating is assigned for each CPGs whether it is Fully Relevant (0.5), Partial (0.3), Few terms are present (0.2) and Not-Relevant (0.0). The averaged score is computed for top 10 and top5 CPGs to find the P@10 and P@5.

Table 1. Relevant Judgement for different levels of ranking

Levels	P@5	P@10
Overall Match	0.55	0.45
PCom Match	0.62	0.56
PMH Match	0.64	0.60
DIOM Match	0.31	0.33

We found improvements in generic queries that contain disease and symptom names. The PMR query terms which contains the general disease names such as “Hypertensive heart disease”,” Coronary artery disease”, “Parkinson’s disease”, “chronic medical renal disease”, “thyrocardiac disease” and “liver disease” gave semantically relevant results even though the terms are not present in the CPGs. The “Query Concept Association with ontology” mapping yields these conceptual results.

Table 2. Relevant Judgement with and without ontology

Levels	P@5	P@10
PMR Ontology Match	0.55	0.45
PMR non-Ontology Match	0.32	0.38

For example ”Hypertensive heart disease” gave results for ”Coronary artery disease”, “hypercholesterolemia”, ”Cardiac coarctation” which are semantic relevant results even though terms are mismatched. Fig. 3 shows ontology based search result for “Liver disease” and obtained chronic liver disease, liver function test acute hepatitis B infection as results. The general symptom-based queries that convey the disorder also gave relevant results. Say for example, when we give query a term ”anxiety” to the ontology it will give a list of concepts such as Generalized anxiety disorder, Mixed anxiety, and depressive disorder, etc. Since

this ontology core subset covers most of the disorder and finding-based concept, we found improvements in symptoms and disease based query terms. The procedure based queries such as ”dialysis”, ”disease screening” and various tests needs to be accessed with specific information, otherwise, it leads to incorrect test procedure related to the diseases. Similarly ordered medication and discharge instruction based search must be specific to the terms and found zero results in the ontology.


Total Hits:100

Guidelines

FileName:https://www.moh.gov.sg/content/dam/moh_web/HPP/Doctors/cpg_medical/current/2014/diabetes_mellitus/cpg_Diabetes%20Mellitus%20Booklet%20-%20Jul%20
CATAGORY:GPP TOPIC:Diabetes Mellitus FileName:cpg_Diabetes Mellitus Booklet - Jul 2014 YEAR:2014 CLASS_LEVEL:GPP GUIDLINES:Evaluation for non-**alcoholic** fatty
about puberty, menstrual irregularities and obstructive sleep apnea should be done at diagnosis and annually thereafter. score: 2.0

FileName:https://www.moh.gov.sg/content/dam/moh_web/HPP/Doctors/cpg_medical/current/2011/cpg-Chronic%20Hep%20B%20Infection%20Card%20-%20Mar%202011
TOPIC:Chronic Hepatitis B Infection FileName:cpg-Chronic Hep B Infection Card - Mar 2011 YEAR:2011 CLASS_LEVEL:GPP GUIDLINES:Patients whose HBsAg is positive
they may be patients with undiagnosed **chronic** hepatitis B virus infection even if the clinical criteria may not have been met yet. The appropriate follow-up actions should th
symptoms of **liver disease**, family history of **chronic** hepatitis B virus infection, any recent travel or high risk activity. Physical Examination: Examine the patients. Look for s
chronic liver disease, ascites, jaundice, etc. Investigation: Check blood for **liver function** test and alpha-fetoprotein level. If either the physical examination or the blood te
gastroenterologist. Consider admitting the patient to the hospital through A and E or direct access, if **acute** hepatitis B infection is suspected. score: 2.0

FileName:https://www.moh.gov.sg/content/dam/moh_web/HPP/Doctors/cpg_medical/current/2011/cpg-Chronic%20Hep%20B%20Infection%20-%20Mar%202011.pdf Page
TOPIC:Chronic Hepatitis B Infection FileName:cpg-Chronic Hep B Infection - Mar 2011 YEAR:2011 CLASS_LEVEL:GPP GUIDLINES:Patients whose HBsAg is positive for th
may be patients with undiagnosed **chronic** hepatitis B virus infection even if the clinical criteria may not have been met yet. The appropriate follow-up actions should then b
of **liver disease**, family history of **chronic** hepatitis B virus infection, any recent travel or high risk activity. Physical Examination: Examine the patients. Look for signs of **live**
disease, ascites, jaundice, etc. Investigation: Check blood for **liver function** test and alpha-fetoprotein level. If either the physical examination or the blood test results are
gastroenterologist. Consider admitting the patient to the hospital through A and E or direct access, if **acute** hepatitis B infection is suspected. score: 2.0

FileName:https://www.moh.gov.sg/content/dam/moh_web/HPP/Doctors/cpg_medical/current/2011/cpg-Chronic%20Hep%20B%20Infection%20-%20Mar%202011.pdf Page
TOPIC:Chronic Hepatitis B Infection FileName:cpg-Chronic Hep B Infection - Mar 2011 YEAR:2011 CLASS_LEVEL:GPP GUIDLINES:Patients whose HBsAg is positive for th
may be patients with undiagnosed **chronic** hepatitis B virus infection even if the clinical criteria may not have been met yet. The appropriate follow-up actions should then b
of **liver disease**, family history of **chronic** hepatitis B virus infection, any recent travel or high risk activity. Physical Examination: Examine the patients. Look for signs of **live**

Fig. 3. Ontology based result snapshot for Liver disease query

In summary, the ontology expansion definitely helped matching: the terms used in the patient medical records and the clinical guidelines are often different. Identifying the Patient Complaints and Previous Medical History and using them for search terms gave higher accuracy than using the entire medical record (62-64% vs 55%): it helps to build the queries from only the most relevant descriptions.

The evaluation is divided into different levels as given in Table 1. Each level of search is evaluated to find the relevance of the results. We found good results for Patient complaints based results and medical history based results. Discharge Instruction and ordered Medication mostly contains abbreviations related to the medical tests conducted and their corresponding results which is unavailable in CPGs. Moreover most of the discharge instruction are not present in the CPGs, sometimes gave zero results for some Discharge Instruction and ordered Medication PMRs. Hence complaints based search and medical history based PMR fields are more important in the retrieval and ranking. Though we are able to achieve reasonably good results, the search ranking method can be extended for finding query intention for PMR queries and semantic representation of document. The abbreviations and multiword terms need to be separated and mapped with the CPG index identifier to further improve the results. When we test the search and rank

accuracy, for a limited term based PMRs, the relevance was good. As we move from smaller to bigger level queries, we found less relevant results. Hence in order to improve the results, we have started working on semantic based representation of both CPGs and PMR to improve the precision.

5. Conclusion

We have designed a basic ontology based search engine that supports multi−level Patient Medical Records search and rank. The document indexing task is simplified by automatically extracting important fields from the document and indexing them, unlike existing IR ranking systems that consider either terms or concepts. This proposed work simplifies the computational effort by dividing the huge query into different categories; field based indexing of CPGs and achieved good results for PMRs that contains higher number of NPs. We can further improve the results by extracting meaningful semantic relations from the PMR that would find the exact ontological concepts. This system is in preliminary version and will be extended to find the query intention with the help of semantic based representation of PMRs and CPGs.

References

1. Langville, A. N., C. D. Meyer. Google's PageRank and Beyond: The Science of Search Engine Rankings. Princeton University Press, 2011.
2. Bodenreider, O. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. – Nucleic Acids Research, Vol. **32**, 2004, Suppl. 1, pp. D267-D270.
3. Manning, C. D., P. Raghavan, H. Schutze. Introduction to Information Retrieval. Vol. **1**. Cambridge, Cambridge University Press, 2008.
4. Cummins, R. Clinical Decision Support with the SPUD Language Model. TREC, 2015.
5. Klein, D., C. D. Manning. Accurate Unlexicalized Parsing. – In: Proc. of 41st Annual Meeting on Association for Computational Linguistics, Vol. **1**, Association for Computational Linguistics, 2003, pp. 423-430.
6. Metzler, D., W. B. Croft. A Markov Random Field Model for Term Dependencies. – In: Proc. of 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2005, pp. 472-479.
7. Zhu, D., S. T.-I. Wu, J. J. Masanz, B. Carterette, H. Liu. Using Discharge Summaries to Improve Information Retrieval in Clinical Domain. – In: CLEF (Working Notes), 2013.
8. Zhu, D., B. Carterette. Improving Health Records Search Using Multiple Query Expansion Collections. – In: 2012 IEEE International Conference on Bioinformatics and Biomedicine (BIBM'12), IEEE, 2012, pp. 1-7.
9. Diaz, F., D. Metzler. Improving the Estimation of Relevance Models Using Large External Corpora. – In: Proc. of 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2006, pp. 154-161.
10. Kilicoglu, H., D. Shin, M. Fiszman, G. Rosemblat, T. C. Rindflesch. Semmeddb: A Pubmed-Scale Repository of Biomedical Semantic Predications. Bioinformatics, Vol. **28**, 2012, No 23, pp. 3158-3160.
11. Oh, H.-S., Y. Jung, K.-Y. Kim. A Multiple-Stage Approach to Re-Ranking Medical Documents. – In: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2015, pp. 166-177.
12. Fung, K. W., M. Ma, S. Srinivasan. The Umls-Core Project – A Study of the Problem List Vocabularies Used in Large Health Care Institutions, 2010.

13. R o b e r t s o n, S. E., S. W a l k e r. Some Simple Effective Approximations to the Poisson Model for Probabilistic Weighted Retrieval. – Readings in Information Retrieval, 1997, p. 345.
14. J o n n a l a g a d d a, S. R., G. D. F i o l, R. M e d l i n, C. W e i r, M. F i s z m a n, J. M o s t a f a, H. L i u. Automatically Extracting Sentences from Medline Citations to Support Clinicians' Information Needs. – Journal of the American Medical Informatics Association, Vol. **20**, 2013, No 5, pp. 995-1000.
15. L a v r e n k o, V., W. B. C r o f t. Relevance Based Language Models. – In: Proc. of 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2001, pp. 120-127.
16. X i a, Y., H. Z h a o, K. L i u, H. Z h u. Normalization of Chinese Informal Medical Terms Based on Multi-Field Indexing. – In: Natural Language Processing and Chinese Computing, Springer, 2014, pp. 311-320.

Received 31.05.2017; *Second Version 01.12.2017*; *Accepted 31.01.2018*