Journal of Theoretical and Applied Information Technology
31st December 2013. Vol. 58 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SEMANTIC-BASED MEDICAL RECORDS RETRIEVAL VIA MEDICAL-CONTEXT AWARE QUERY EXPANSION AND RANKING

[1]NURFADHLINA MOHD SHAREF, [2]HIZMAWATI MADZIN

[1]Department of Computer Science, Faculty of Computer Science and Information Technology, University Putra Malaysia, MALAYSIA

[2]Department of Multimedia, Faculty of Computer Science and Information Technology, University Putra Malaysia, MALAYSIA

E-mail: [1]fadhlina@upm.edu.my, [2]hizmawati@upm.edu.my

ABSTRACT

Efficient retrieval of medical records involves contextual understanding of both the query and the records contents. This will enhance the searching effectiveness beyond merely keyword matching and is assisted by analyzing its semantics notion such as by the utilization of the MeSH thesaurus. The query is annotated and expanded by information from the deep medical contextual understanding. This is because typically medical records contain medical terminologies which may not be included in the user query but is important for accurate search hit. Besides, the terminologies have synonyms which should be utilized for richer and expanded query. The main contribution of the paper is the semantic-based retrieval technique by utilizing context-aware query expansion and search ranking method. Medical domain is chosen as a proof of concept and a medical record retrieval application was developed. The source of medical records are obtained from the ImageCLEF 2010 dataset which also houses a series of evaluation campaign such as photo annotation, robot vision and Wikipedia retrieval. This paper addresses the following problems: (i) semantic-based query expansion technique which increase the content awareness ability, (ii) MeSH-manipulated indexer which entails medical terminologies and their synonym, (iii) adoption of extended Boolean matching to measure similarity between query and documents, and (iv) ranking method which prioritizes matched expanded query size. The results were measured using precision, recall and mean average precision (MAP) score. Comparing against other approaches, our method has several achievements including; (i) more efficient access of MeSH thesaurus through the manipulated indexer compared to its original form; (ii) enrichment of query expansion using synonym term can improve mean average precision (MAP) value as opposed to standard query expansion; (iii) our comprehensive ranking method achieved high recall. According to MAP score we are in the top five run system amongst submitted run systems in ImageCLEF2010 medical task.

Keywords: *Clinical Record Retrieval, Semantic, Query Expansion, Ranking Method,*

1. INTRODUCTION

The enormous volume of clinical data motivated by the evolution of storage and networking technologies has created challenges for the information retrieval (IR) discipline (Alba C 2010, Volk M 2002). This is factored by the high usage of medical terminologies which are commonly synonymous and ambiguous. This scenario poses limitation in conventional IR because equal concern is given to all the terms in the query, despite the occurrence of the terminologies.

Semantic analysis of the query content and the understanding of the medical record contents can ensure higher relevant matching between the query and the clinical corpus The Medical Subject Heading (MeSH)[1] is an example of standard medical semantic representation tool that can assist in the analysis and annotation of the medical terminologies in the clinical records. The performance of the IR processes is improved through the identification of the terminologies, their hierarchies and classes. Ambiguity from acronym and abbreviation used in the query is solved because sometimes users may use slightly different format in the query which may not exist in the document (Mark Stevenson 2010).

[1] The Medical Subject Heading (MeSH)[1] is a semantic medical knowledge resource (thesaurus) developed by the National Library of Medicine (NLM) of the United States which have been exploited by several researchers in order to bridge the gap between surface linguistic form and meaning.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Medical reports are naturally enriched with medical terminologies, which are not necessarily familiar to the user. The ignorance of this is in current information retrieval technique has resulted to low precision. In this paper we adopt MeSH to annotate the medical terminologies with the synonymous terms. This allows more potential matching when a user query is obtained. We reduce query ambiguity through extended Boolean matching model and MeSH-enriched query expansion method. A MeSH indexer is created by manipulating the data from MeSH thesaurus to increase efficiency in accessing the medical and synonymous terms. A new ranking method is introduced which gives priority on (i) the size of matched expanded queries and (ii) the total occurrence of the matched expanded queries. These methods are applied in our medical records retrieval model called IMS.

The paper is organized as follows: The first section introduces the background problem. Section 2 briefs the background of the study while Section 3 explains the general framework of IMS system. Here the description of MeSH thesaurus and the creation of MeSH-indexer are detailed out. This section also describe the methodology of MeSH-enriched query expansion process. This is followed by the details of ranking method in Section 4. The system is experimentally evaluated and we report the result is Section 5. In Section 6, we address the raised issues and finished by conclusion and future work.

2. BACKGROUND

The components of a medical document retrieval system are (i) the corpus index which organizes the content of the documents in the corpus (ii) distance measurement between query and documents or among documents (iii) the ranking of documents according to their similarity to the query. Boolean standard model and vector space model are the examples of conventional retrieval method. Vector space model is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers, such as, for example, index terms (G. T. Salton 1975). The vector space model utilizes the linear algebraic property from the documents being represented instead just the binary representation in Boolean model. This allows the computation of continuous degree of similarity between queries and documents. Documents are ranked according to their relevance besides partial matching score.

The Levenshtein distance is one of the most used partial matching techniques which is also known by matching with error (Joaquín Adiego 2004). The Levenshtein distance allows the minimum number of edit operations being insertion, deletion, or substitution of a single character needed to transform one string into the other. The extended Boolean model allows the combination of standard Boolean matching (binary representation or present/absent) with the partial matching and term weighting (Salton and Edward A. Fox 1983).

Approaches for similarity computation include (i) Boolean retrieval which is characterized by the occurrence of the query terms in the documents (David A. Grossman 2004), (ii) dice coefficient which may be defined as twice the shared information (intersection) over the sum of cardinalities (Rada Mihalcea 2006), (iii) cosine similarity which measures the similarity between two vectors by based on the cosine of the angle between them (Patel 2007), (iv) TFIDF which evaluates the importance of a word in a document or corpus according to the number of the word's occurrence in the documents (Hiemstra 2000). However, all these heavy index-based approaches create problems like large indexing storage, high computational overhead and sparse data representation. Furthermore, these methods do not cover the disambiguation and semantic notion capability which is needed to understand the user's query.

Semantic closeness between objects is typically computed by mapping terms to ontology and by examining their relationships in that ontology; to determine the intended meaning of an ambiguous word. Other similar methods have introduced online dictionary and encyclopedia such as WordNet, word sense disambiguation, taxonomy structure information and specific domain thesaurus such as Medical Subject Headings (MeSH)[3] and Unified Medical Language System (UMLS)[6], and semantic distance formulas. Further studies then follow with the focus to record and organize medical knowledge such as UniProt[2] (R.Apweiler 2004), GO[3] (R.Stevens 2000) and UMLS[4] (Bodenreider 2004). These methods allow the similarity computation on the lexically similar terms and terms that are conceptually similar. Ontology and thesauri were adopted in retrieving biomedical information from large volumes of documents in (Bodenreider 2004, Nelson SJ 2004) (A.R.Aronson 2001, S.Myhre 2006).

[2] http://www.ebi.uniprot.org/index.shtml
[3] http://www.geneontology.org/
[4] http://www.nlm.nih.gov/research/umls/

Journal of Theoretical and Applied Information Technology
31st December 2013. Vol. 58 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Successful medical retrieval is also factored by query ambiguity reduction. Query expansion is an effective technique (G. 1997, Mitra, Singhal and Buckley 1998), mainly due to the extension of the extra knowledge on the original query to make it more significant to the documents in the domain. This has motivated the technique to expand original query terms using ontology (J. Bhogalb 2007).

In this paper, we used 16 ad-hoc queries with relevance judgments containing relevant documents for each query from the medical documents in ImageCLEF 2010 (H. Müller 2010). These relevant judgments are assumed as the gold standard in our experiment. 70,000 medical reports and images are used as a source.

There are several other works that have tested the same data collection. Hong Wu (2010) used phrase extraction as indexing term by extracting the phrases and subphrase using MetaMap (Aronson 2006). MetaMap discovers concept and meaning in the query and medical documents by mapping biomedical text to UMLS Metathesaurus. The SINAI project (M.C. Díaz-Galiano 2010) utilized MeSH ontology using machine learning for decision making. (Ragia Ibrahim 2010) used caption segmentation based on natural language processing to select the most relevant sentence in the caption given. The IPL[5] group unifies medical documents with same title and caption into one record tagged by figure index. The combination of Medline term in PubMed website extracted from the PubMed and MeSH headings in MetaMap for MeSH term are used by OHSU[6] for CLEF data processing. OHSU has also adopted modality filter which classifies type of medical images according to categories such as computer tomography (CT) scans, X-ray and ultrasound. ISSR[7] has observed that paragraph extraction is less effective compared to sentence selection and that paragraph extraction did not improve the MAP score but increase recall. Query translation using the Google Machine Translation System and UMLS thesaurus are approached in the IR process by Bioigenium[8]. However this approach is computational expensive and time consuming and can cause slow performance. Table 1 shows the comparison of retrieval strategies between other studies that used the same medical ImageCLEF 2010 dataset.

We highlight our contributions compared to previous related studies (i) MeSH enriched query expansion technique, (ii) the manipulated MeSH architecture to speed up retrieval process (ii) the modification of traditional Boolean matching technique as similarity measurement method, and (iv) the scoring and ranking method. Our method is distinguished from the previous techniques that are mostly based on the vector space model (VSM) (Tamine 2010) or probabilistic models such as the Okapi model (Ros Stougiannis 2010) because these ranking methods sort documents based on their occurrence and relevance to the queries. On the contrary, we give priority to longest matching term where similarity is decided based on extended Boolean matching and the occurrence of matching queries in the documents. The online matching approach results to the reduction on processing time and optimizes space performance while the conventional methods need to be reconstructed when there is new document available which consume time and increase more space in the storage.

3. IMS FRAMEWORK

The *IMS* framework consists of two main components namely the query expansion and documents retrieval. The inputs of the framework are the query and the clinical records. The clinical records contain descriptions about cases based on the medical images in various modality which include magnetic resonance image (MRI), ultrasound, positron emission tomography (PET), computed tomography (CT) scan and X-ray data. The MeSH thesaurus is utilized to extract medical terms from the clinical records and to expand original queries with their synonymous terms. The output is the relevant clinical records found by launching queries with medical terms and the expanded queries.

3.1 MeSH Thesaurus

There are 26,142 medical subject heading terms naming descriptors in the XML-formatted MeSH thesaurus. The hierarchical structure permits searching of various related information at various levels of specificity such as the concept element which represent term synonymy as well as other information, such as relations between concepts. Generally MeSH has three-level structure which are 'Descriptor', 'Concept' and 'Term'. Each specific medical term known as 'Descriptor' may consist of

[5] IPL group represent Information Processing Laboratory from Athens University of Economics and Business

[6] OHSU is research group from Oregon Health & Science University (http://www.ohsu.edu/xd)

[7] ISSR represent Institute of Statistical Studies and Research at Cairo University

[8] Bioigenium represent Bioingenium Research Group, National University of Colombia

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

more than single concept element. 'Concept' corresponds to a class of terms which are synonymous with each other and 'Term' is the synonymous medical term in each concept.

3.2 MeSH-Indexer

MeSH thesaurus is heavily used in the query expansion in identifying medical terms and any synonymous terms. Moreover not all information in the MeSH are used for this research. Therefore, MeSH-indexer offline has been created to reduce searching time and increase the performance. The indexer is created by manipulating and filtering the MeSH thesaurus as an effort for more efficient thesaurus reference. Mesh-indexer contains three folders:

1. MeSH-Medical Descriptor (MMD) folder. Each file contains medical terms that start with the same alphabet. This medical term is taken from <DescriptorName> of each <DescriptorRecord> in MeSH thesaurus.
2. MeSH-Synonym Terms (MST) folder. Each file contains a set of data in the form of medical descriptor <DescriptorName> and list of concepts <ConceptName> and terms <TermList> (synonymous terms).
3. Medical Conceptual (MC) folder. This folder contains semantic type <SemanticType> and conceptual files (details in Table 4.5) for each medical descriptor, concepts and terms.

These three folders are organized according to alphabetical order. Therefore it will speed up the search between data and information in MeSH by only looking at the initial letter of the term.

3.3 Query Expansion Process

In the query expansion process (Figure 3) the original query, q_0 submitted by user is enriched with synonymous medical terms based on the MeSH thesaurus. This generates three types of outputs:

- Type 1: Exact match with the original query.(q_0)
- Type 2: Expansion of original query.(Eq_r)
- Type 3: Combination of expansion and MeSH-enriched query (Mq_r)

Retrieval using query generated from query expansion Type 1 will be based on standard Boolean matching by checking the occurrence of q_0 in the documents. Typically, it is possible to retrieve relevant medical documents by using only q_0. However, this traditional method is found to possess many limitations as it only returns results that contain exact matching, where typically there exist variation of the query used.

Meanwhile, in Type 2 process, Eq_r a set of additional queries will be automatically generated by using n-gram method (William B. Cavnar 1994) where n=[1, length of q_0]. For example given q_0="CT images containing fatty liver", the set of n-gram is:

1-gram*: *CT, images, containing, fatty, liver

2-grams: *CT images, images containing, containing fatty, fatty liver*

3-grams: *CT images containing, images containing fatty, containing fatty liver*

4-grams: *CT images containing fatty, images containing fatty liver*

However, not all terms produced from n-gram process can be used and have significant meaning. For example, the term "images containing" and "containing fatty" are not useful and meaningless. This leads to the issue to identify significant terms in the query. To answer this question the MMD from Mesh-indexer is utilized to filter out queries that contain medical terms. The set of filtered queries are "CT images", "fatty liver" and "liver". For retrieval process, queries generated using this query expansion type will be matched based on extended Boolean function.

Table 1: Example of Expanded Query Best Matching

Expanded Queries: Liver, Fatty Liver	
Original medical report being checked for matching term: … *Focal spared area in the fatty liver along the porta hepatis in a 73-year-old woman…*	
Matching query term "liver" (MEQsize=1)	… *Focal spared area in the fatty* ***liver*** *along the porta hepatis in a 73-year-old woman…*
Matching query term "fatty liver" (MEQsize=2)	… *Focal spared area in the* ***fatty liver*** *along the porta hepatis in a 73-year-old woman…*
Best match: … *Focal spared area in the* ***fatty liver*** *along the porta hepatis in a 73-year-old woman…*	

The query expansion in Type 3 performs query enrichment (Mq_r) with the synonymous terms recorded in the MST folder. Based on q_0, only "liver" and "fatty liver" terms have synonym which are "livers", "liver steatosis", "steatohepatitis, "steatosis of liver" and "visceral steatosis". However, before the query enrichment process, the query expansion as explained in query expansion strategy Type 2 will be performed. Therefore, the expanded queries generated in Mq_r are "CT images", "fatty liver" and "liver", "livers", "liver steatosis", "steatohepatitis, "steatosis of liver" and "visceral steatosis".

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.5 Ranking Method

Comprehensive ranking model ranks the documents based on the size of term matched in *RMR*. As previously mentioned, *RMR* list consists of Figure ID, Title, Caption, *MEQHit* and *meq* followed by *meq_hit*. We assume that, the more terms that match between query and document, the more relevant the document is deemed to be to the query. Figure 5 shows the flow chart of the comprehensive ranking model.

The input of this flow chart is retrieved *RMR* while the output is the result file of ranked *RMR*. Initially the system initializes *MEQSize* as number of item in *MEQ* that exist in each *RMR*. Then the next step is to sort *RMR* based on the largest size of *MEQSize*. To execute *RMR* ranking process, *CurrentRMR* is created as a temporary variable to place the current *RMR*. For each *RMR*, the *MEQSize* is identified and later *nameOfFile* is created based on *MEQSize* and *RMR* is copied by *CurrentRMR* into *nameOfFile* folder. For each folder *CurrentRMR* is sorted based on descending order of *MEQHit*. Finally, *nameOfFile* folders are combine and sorted in descending order based on *MEQSize.*

Figure 5, 6, and 7 show the example for q_0="CT images containing fatty liver" containing three *nameOfFile* folders namely MatchedSize-1, MatchedSize-2 and MatchedSize-3. For each *nameOfFile* folder, *CurrentRMR* is sorted based on descending order of *MEQHit*. This process continues for every item in *RMR* to generate the final result of the ranking process. The *nameOfFile* folders will be gathered and ranked and sorted in descending order of *MEQSize*. Finally Figure 8 is the example of *RMR* list output from comprehensive ranking model which similar with the output of retrieval strategy component but the order is based on *MEQSize* followed by *MEQHit*.

4. EXPERIMENTS AND RESULTS

In this section we experimentally evaluate our query expansion process and comprehensive ranking method. The result of this experimentation is compared with other conventional method. Our main goal is to investigate the effectiveness of our method in improving the IR system. The experimental set up and the result will be explained in detail in this section.

4.1 Experiment Setting

The dataset used in sourced by the medical data collection of ImageCLEF 2010 (H. Müller 2010) supplied by the Cross Language Evaluation Forum (CLEF) (Peters(Ed.) 2000). This data collection include of 70,000 medical documents and images. It also provides 16 ad-hoc queries with relevance judgments containing relevant documents for each query. These sets of relevant documents are assumed as the gold standard in our experiment and referred to as the relevant set in the remainder of this research article.

There are many evaluation methods proposed in IR system to evaluate the effectiveness of query expansion (Hersh W 2003). For this research paper we used precision, recall, and MAP value. The performance was first evaluated by recall (number of relevant documents in the collection retrieved by the query) followed by precision (number of relevant documents retrieved by the query). Since there are several queries involved in this experiment and the documents for each query might not be the same amount; therefore average precision is the precision values at the positions of each relevant documents is retrieved. Average precision computing a precision and recall at every position in the ranked sequence of documents.

Therefore the result is more precise where each document is measured based on its position in the list and we will know which medical document is the most relevant with the query. Then, MAP value is obtained by take the mean value of average precision from all the queries. Average precision is considering the order of returned relevant documents in the ranked sequence.

Objective 1: In order to verify the efficiency of our method, we implemented an experiment to evaluate the query expansion method as follows:

i. *Type 1: Direct matching with original query*
Evaluate proportion of relevant documents retrieved using original query from the user.

ii. *Type 2: Expansion*
Evaluate proportion of relevant documents retrieved based on medical terms extract from original query that match in MeSH-indexer

iii. *Type 3: Expansion and Enrich*
Evaluate proportion of relevant documents retrieved by combining both terms in expansion and enrich (synonymous term).

Objective 2: The retrieved medical documents obtained from the experiments above are ranked based on the comprehensive ranking method. Each result in the experiments is viewed in order to verify the expected improvement. The evaluation will be based on average precision of each query. We also compare our result with other researcher that used the same dataset. The evaluation is based on relevance judgment given obtained from the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ImageCLEF 2010 collection. MAP value is used for the comparison since this method is a standard measurement evaluation in most CLEF research. Moreover MAP more stable measurement compare to other measures such as precision and R-precision since it can assess more information (C. Buckley 2000).

Objective 3: To prove the effectiveness of MeSH-indexer, we execute time evaluation between MeSH-indexer and MeSH thesaurus.

4.2 Experiment Results

Table 3 shows the list of ad-hoc query and the average precision value of expansion, enrich and combination of expansion and enrich. It is shown that the query at ID 3, 4, 5, 9 and 14 have improvement up to 13% respectively because these queries have synonym terms in the MeSH thesaurus and these synonym terms were considered in the ImageCLEF relevance judgment result. In fact, the remaining queries have synonym terms but were disregarded in the benchmark data. Result has shown that the expansion and enrich query (Expansion Query Strategy Type 3) technique have higher average precision value compared to the others. There is also query that has zero precision such as ID 7,8 and 16. This is because of the constraints posed by our method that relies on the match between the query and the medical terms in the MeSH-indexer. Therefore, since the terms in the query do not exist in the MeSH indexer, zero precision is obtained. A similar situation is demonstrated in Figure 10.

The graph in Figure 5 shows the recall value for most of the queries is high. It proves that this method offers a promising solution since most of the relevant documents have been retrieved. Nevertheless there are few queries (with ID 3, 11, 13, 14, 15) that obtain low recall value. This is because our method has retrieved results which are not regarded as relevant by the benchmark. However, we argue that the results we obtained are relevant because the synonymous terms are identified. Therefore we assume that the benchmark data we referred to was not prepared by considering occurrence of synonymous terms and defend that our performance is competitive. It also observed that our method which prioritized MEQ has more strength in recall compared to precision score. This means that although most relevant documents are retrieved but the rank of relevant document is still in low positions which lead to low precision value as shown in Figure 11.

Table 4 shows the comparison of the similarity and representation method, query expansion technique and MAP values employed by other systems that used medical ImageCLEF 2010 data. Note that the MAP values from the listed approaches were originally generated in various run system but we have highlighted only their highest score. The IMS method has obtained the top ten ranking.

As described earlier, MeSH thesaurus is used heavily in the retrieval process. We have developed a MeSH-indexer to organize the components of the medical thesaurus which focus were given to the MeSH heading term and the synonym term. An experiment was carried out to observe the computational performance between the manipulated MeSH and the original version. The result on the performance of the MeSH-indexer over the original MeSH is shown in Table 5.

5. CONCLUSION

The retrieval of clinical records is challenged by the condensed usage of medical terminologies which cannot be processed efficiently with traditional IR method. This paper highlights a semantic-based clinical records retrieval model by expanding the query with semantic information sourced from the MeSH thesaurus so that the query and the content of the records are better understood. A MeSH manipulated indexer is also introduced which organizes and filters the needed information. Our third contribution is the enhancement of the standard Boolean matching technique for more precise similarity matching.

Unlike statistical model, our Boolean model emphasizes on efficient process with advantages discussed earlier and which is only based on matching between query and documents and no term occurrence indexing required on the documents. Therefore new documents can easily be inserted in the data collection without having to re-index the length vector of the documents. A comprehensive ranking method is introduced where relevance of documents is judged based on the size of matched expanded query in addition to term occurrence.

Our contributions are distinguished from the other related works that are based on term weighting and offline content indexing. Instead, we offer online indexing of records which ensures faster processing time and optimizes memory usage. Observation results on recall, precision and MAP value has shown that our proposed method achieve top five run system amongst submitted run systems in ImageCLEF2010 medical task and the developed MeSH-indexer has improved

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

computational performance compared to the original MeSH version. A few future works can be suggested which are (i) investigation of ranking by giving priority to specific term that relate to clinical question such as patient characteristic, type of disease, disease condition, diagnostic test and etc; (ii) reorganization of the medical documents to remove redundant description; (iii) perform document expansion instead of query expansion and observe the retrieval performance.

ACKNOWLEDGMENT

The authors gratefully acknowledge the funding support received from the Malaysian Ministry of Higher Education under the Long Research Grant Scheme, Project No: LRGS/TD/2011/UITM/ICT/01/02

REFERENCES

[1] A.R.Aronson. "Effective mapping of biomedical text to the UMLS Metathesaurus:." Proceedings of the AMIA Symposium. 2001. 17–21.

[2] Alba C., Tamara S., Jose A. S., Vanessa A., Ana F., Julian D., Alejandro P. “Retrieval and management of medical information from heterogeneous sources, for its integration in a medical record visualisation tool” Int. J. of Electronic Healthcare, 2010 Vol.5, No.4, pp.371 - 385

[3] Aronson, AR. "MetaMap: Mapping Text to the UMLS Metathesaurus." AMIA Annual Symposium Proceeding. 2006.

[4] Bodenreider, Olivier. "The Unified Medical Language System (UMLS): integrating biomedical terminology." Oxford Journals Life Sciences Nucleic Acids Research Volume32, Issue suppl 1, 2004: D267-D270.

[5] C. Buckley, EM Voorhees. "Evaluating evaluation more srability." 23rd Annual International ACM SIGIR Conference on research and Development in Information Retrieval. 2000. 33-40.

[6] David A. Grossman, Ophir Frieder. Information retrieval: algorithms and heuristics. Springer, 2004.

[7] G., Salton. "Developments in automatic text retrieval." Science, 1997.

[8] H. Müller, J. Kalpathy-Cramer, I. Eggel, S. Bedrick, J. Reisetter, C.E.K. Jr., and W.R. Hersh. "Overview of the CLEF 2010 Medical Image Retrieval Track." CLEF (Notebook Papers/LABs/Workshops). 2010.

[9] Hersh W, Bhupatiraju RT, Price S. "Phrases, boosting and query expansion using external knowledge resources for genomic information retrieval." The 12th Text Retrieval Conference. 2003.

[10] Hiemstra, Djoerd. "A probabilistic justification for using tf×idf term weighting in information retrieval ." International Journal on Digital Libraries , 2000.

[11] Hong Wu, Changjun Hu , Sikun Chen. "UESTC at ImageCLEF 2010 medical retrieval task ." CLEF 2010. Padova, Italy, 2010.

[12] J. Bhogalb, A. Macfarlanea and P. Smitha. "A review of ontology based query expansion ." Information Processing & Management, 2007: 866-886.

[13] Joaquín Adiego, Pablo de la Fuente , Gonzalo Navarro. "Merging Prediction by Partial Matching with Structural Contexts Model." DCC '04 Proceedings of the Conference on Data Compression . 2004.

[14] M.C Diaz Galiano, M.T Martin Valdivia, L.A Urena Lopez. "Query expansion with a medical ontology to improve a multimodal information retrieval system." Computers in Biology and Medicine, 2009.

[15] M.C. Díaz-Galiano, M.T. Martín-Valdivia, A.M. Ráez, and M.A.G.Cumbreras. "SINAI at ImageCLEF 2010 Medical Task." Proceedings of CLEF (Notebook Papers/LABs/Workshops)'2010. 2010.

[16] Mark Stevenson, Yikun Guo. "Disambiguation of ambiguous biomedical terms using examples generated from the UMLS Metathesaurus." Journal of Biomedical Informatics, 2010: issue 5.

[17] Mitra, M., A. Singhal, and C. Buckley. "Improving automatic query expansion;." Mitra, M.; Singhal, A.; Buckley, C. SIGIR ‘98: Proceedings of the 21st annual international ACM SIGIR. 1998.

[18] N M Sharef, Trevor Martin, Yun Shen. "Incremental Evolutionary Grammar Fragments." Proceedings of the 2008 UK Workshop on Computational IntelligencePage 89. UK, 2008. 89-94.

[19] Nelson SJ, Schopen M, Savage AG, Schulman JL, Arluk N. "The MeSH translation maintenance system: structure, interface design, and implementation." 11th congress in Medical Informatics. 2004. 67-69.

[20] Patel, Sandeep Tata Jignesh M. "Estimating the Selectivity of tf-idf based Cosine

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Similarity Predicates." SIGMOD Record. 2007. (Vol. 36, No. 2).

[21] Peters(Ed.), C. "Cross-Language Information retrieval and Evauation Workshop of Cross Language Evaluation Forum." CLEF 2001. Springer , 2000. 387.

[22] R.Apweiler, A.Bairoch,C.H.Wu,etal. "UniProt:theuniversalprotein." NucleicAcidsResearch32, 2004: 115-119.

[23] R.Stevens, C.A.Goble,S.Bechhofer,. "Ontology-basedknowledgerepresentation." Brief Bioinformatics 2000 1, 2000: 398-414.

[24] Rada Mihalcea, Courtney Corley. "Corpus-based and knowledge-based measures of text semantic similarity." AAAI Conference on Artificial Intelligence . 2006. 775-780.

[25] Ragia Ibrahim, Waleed Arafa. "Medical Image Retrieval: ISSR at CLEF 2010." CLEF 2010. 2010.

[26] Ros Stougiannis, Anestis Gkanogiannis , Theodore Kalamboukis. "IPL at ImageCLEF 2010 ." CLEF 2010. 2010.

[27] S. Radhouani, G. Falquet, J.-P. Chevallet. "Description Logic to Model a Domain Specific Information Retrieval System." In Proc. 19th International Conference on Database and Expert Systems Applications (DEXA'08). Turin, Italy: Lecture Notes in Computer Science, Vol. 5181, Springer., 2008.

[28] S.Myhre, H.Tveit,T.Mollestad,A.Laegreid. "Additional Gene Ontology structure." Bioinformatics 22, 2006: 2020–2027.

[29] Salton, G., Tang, C.S., and Wong, A. "A vector-space model for automatic indexing." Communication of the CAM, 1975: 613-620.

[30] Salton, Gerard, and Harry Wu Edward A. Fox. "Extended Boolean information retrieval." Communications of the ACM, 1983: Volume 26, Issue 11.

[31] Tamine, Duy Dinh and Lynda. "IRIT at ImageCLEF 2010: Medical Retrieval Track." CLEF 2010. 2010.

[32] Volk M, Ripplinger B, Vintar S, Buitelaar P, Raileanu D, Sacaleanu B. "Semantic annotation for concept-based cross-language medical information retrieval." International Journal of medical Informatics, 2002: 97-112.

[33] William B. Cavnar, John M. Trenkle. "N-gram based text categorization." Proc. of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. 1994.

[34] Zhiyong Lu, Won Kim, W. John Wilbur. "Evaluation of Query Expansion Using MeSH in PubMed." Information retrieval Boston, 2009: 69-80.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: Results Of Average Precision On Ad-Hoc Query List

ID	Query	EXPANSION	EXPANSION & ENRICH
1)	CT images thoracic aortic dissection	0.2580	0.258
2)	A microscopic image of Acute Myeloid Leukemia	0.0050	0.005
3)	ECG images	0.2940	0.404
4)	x-ray showing congestive heart failure	0.0065	0.007
5)	CT images for brachial plexus nerve block	0.1760	0.244
6)	CT images containing fatty liver	0.2930	0.293
7)	x-ray images of a greenstick fracture	0.0000	0.000
8)	microscopic images streptococcus pneumonia	0.0000	0.000
9)	MR images papilledema	0.6820	0.903
10)	MR images pericardial effusion	0.6350	0.635
11)	All types images with atherosclerosis in blood vessels	0.0150	0.015
12)	Radiation therapy treatment plans	0.0710	0.071
13)	Images of dermatome	0.2000	0.200
14)	images showing sacral fracture	0.8130	0.938
15)	images coronary arteries	0.0456	0.082
16)	images dermatofibroma	0.0000	0.000
Mean Average Precision (MAP)		**0.218**	**0.253**

Table 3: Comparison Of MAP Results

No	Run System	MAP
1	IPL	0.316
2	OHSU	0.299
3	UESTC	0.279
4	IPL	0.278
5	UESTC	0.273
6	OHSU	0.261
7	ISSR	0.258
8	HES-SO VS	0.257
9	OHSU	0.256
10	***IMS Expansion & Enrich***	***0.233***
11	ISSR	0.231
12	*IMS Expansion Only*	*0.223*
13	ISSR	0.219
14	ITI	0.188
15	ITI	0.158
16	ISSR	0.147
17	HES-SO VS	0.131
18	Bioingenium Research Group	0.101
19	ISSR	0.098

Journal of Theoretical and Applied Information Technology
31st December 2013. Vol. 58 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4: Time Evaluation Between Mesh-Indexer And Original Mesh Thesaurus

	Expansion (milliseconds)	Expansion and Enrich (milliseconds)
Original MeSH	4896	4896
MeSH-Indexer	74	101
% Improvements by MeSH-Indexer	66.16%	48.48%

FigureID	Title	Caption	MEQHit	MEQSize=1
28055	ith multiphase breat	the suspected lesion (op	3	liver:3
28080	ance of the soft palat	the soft palate (arrow) a	1	fatty liver:1

Figure 5: Matchedsize-1 File

FigureID	Title	Caption	MEQHit	MEQSize =1	MEQSize =2
28045	vith multiphase breat	inium-enhanced mr angic	5	liver:4	ct:1
28056	trast medium on panc	0 ml of nonionic contrast	3	liver:1	ct:2
28057	trast medium on pan	n volume of 150 ml and	2	liver:1	ct:1

Figure 6: Matchedsize-2 File

FigureID	Title	Caption	MEQHit	MEQSize =1	MEQSize =2	MEQSize =3
59549	cs and pseudolesions	h obtained during arteria	8	fatty liver:1	liver:2	ct:5
112963	omparison of helical	angle 15&#176) and (c) sp	5	fatty liver:1	ct:3	liver:1
117448	focal lesions at micr	enous phase us image th	4	fatty liver:1	liver:2	ct:1

Figure 7: Matchedsize-3 File

FigureID	Title	Caption	MEQHit	MEQSize=1	MEQSize=2	MEQSize=3
59549	cs and pseudolesions	obtained during arteria	8	fatty liver:1	liver:2	ct:5
112963	omparison of helical	angle 15&#176) and (c) s	5	fatty liver:1	ct:3	liver:1
117448	focal lesions at micr	enous phase us image th	4	fatty liver:1	liver:2	ct:1
28045	ith multiphase breat	hium-enhanced mr angic	5	liver:4	ct:1	
28056	trast medium on pan	) ml of nonionic contrast	3	liver:1	ct:2	
28057	trast medium on pan	n volume of 150 ml and	2	liver:1	ct:1	
28055	ith multiphase breat	he suspected lesion (op	3	liver:3		
28080	ance of the soft palat	the soft palate (arrow) a	1	fatty liver:1		

Figure 8: Example Output In Comprehensive Ranking Model