International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 12 (2017) pp. 3500-3506

A Secure Electronic Medical Information Retrieval System in EMIRS

R.Aravazhi[1] and Dr. M.Chidambaram[2]

[1]Ph.D., Research Scholar, A.V.V.M Sri Pushpam College (Autonomous), Poondi, Thanjavur, Tamilnadu. India.

[2]Assistant Professor in Computer Science, Rajah Serfoji Government College (Autonomous), Thanjavur, Tamilnadu, India.

[1]ORCID: 0000-0002-5694-5180

Abstract

Information retrieval in medical domain is now sharing major part of the web search. Now a day's most of the people especially adults are browsing health care and medical information at their homes using internet. Electronic Medical Information Retrieval System (EMIRS) through search engines providing positive information to the user based on the fixed questionnaires. In this paper we build a model for naïve users, who are having minimal knowledge to feedback the system by opting listed relevant questionnaire. Along with the framework, we also built an Intelligent Medical Search Engine (IMSE) for searching medical information on World Wide Web (WC3). The implementation setup of IMSE uses medical Ontology and questionnaire to facilitate naive internet users to search for medical information. IMSE introduces and extends expert system technology into the search engine domain. IMSE uses several key techniques to improve its usability and search result quality.

Keywords: Medical Information Retrieval; Medical Ontology; Knowledge Network; Diagnosis Reports;

INTRODUCTION

Convention of the internet becomes so popular in almost every country. People use internet to share the information needs, means users may search for their information needs or they may share the information with trusted parties. Every day billions of people use internet for their personal information needs. As per the survey of H&HN DAILY magazine 80 percent of Internet users look online for health information, making it the third most popular online activity among those tracked by the study, trailing only e-mail and using search engines. Roughly 44 percent of Internet users look online for information about doctors and other health professionals; 36 percent look up information on hospitals and other medical facilities. Similarly the PWC survey found that only 14 percent of Americans currently access their medical records electronically, the upward trends in online health engagement suggest those numbers will climb dramatically in the next few years. By observing these facts we proposed to develop enhanced and user friendly (i.e. trusted) medical search engine. In order to capture major share of the users who periodically search their medical information needs, we build this novel search engine[8]. As a road map we started looking at various Medical search engines so as to find the enhancements and requirements of the naïve users.

In this paper, we propose a novel approach to achieve the diversity-aware retrieval of medical records, where the semantic-based IR and search result diversification are combined together to tackle inherent ambiguity of the medical search. Different from existing diversifying strategies relying heavily on large amounts of query logs, the proposed approach employs a medical ontology that comprises rich medical knowledge to disambiguate the original query into multiple sub-queries (or query aspects). Each sub-query represents one aspect of the implied intents of the original query. Based on the modeled aspects of the sub-queries, we gives a novel strategy that exploiting the query disambiguation results for the diversity-aware medical search. The performance of the proposed approach is demonstrated on a real-world medical dataset. Experiment results show that the proposed approach fits well for the medical search environments and outperforms existing methods on both diversity and accuracy[4]. The contribution of this paper can be summarized as follows: (1) A novel approach for exploiting the ambiguity in a medical query for diversity-aware medical search is proposed, which first employs the medical domain knowledge for query understanding to construct multiple sub-queries from the original query and then the medical record relevance and novelty are combined together to handle the uncertainty in the information needs[3]; (2) The empirical experiments on the real-world dataset are reported, which demonstrate the effectiveness of the proposed approaches; (3) A pilot study is described for the application of the proposed medical search approach in a real-world usage scenario.

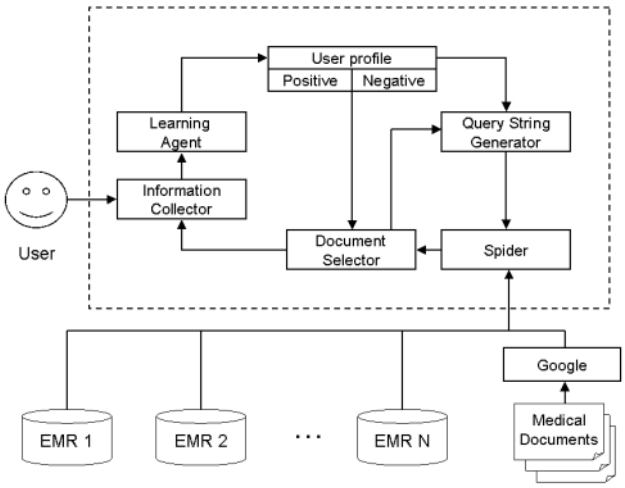

Figure 1: EMR System Framework

BACKGROUD THEORY

In this section, we present how diversity is currently considered in existing tools and approaches for web search. Then, we present the two notions of diversity that are considered in this work: 1) different aspects considered and 2) factual vs. affective content.

a) Diversity in Web Search

Research on information retrieval in the biomedical domain has focused on the retrieval of biomedical research publications. These highly specialized text mining applications incorporate natural language processing capabilities, particularly specialized for the biomedical domain, complex algorithms and rules based on scientific vocabularies. Research in that area has focused on improving search results by taking domain knowledge into account[1]. To reduce an overload of low-quality pages in search results, specialized search engines have been set up including Curbside.md b for physicians or Health line c for patients. These search engines only retrieve content from professional sources or at least from verified sources. In contrast, we explore information retrieval from medical social media content, which is so far still a relatively unexplored domain and are considering diversity of search results. Social media content can provide additional insights into a medical topic that are not available in biomedical literature. Its authors recapture research results, enrich them with their own practical experiences which might be from a patient's perspective or from a professional perspective. Several studies have shown that users usually prefer diversified search results, i.e. results that are dissimilar. Thus, diversity is introduced as measure for dissimilarity[3]. Result diversification is realized by finding the best tradeoff between diversity and similarity. It targets finding the right balance between having more relevant results of the 'correct' intent and having more diverse results in the top positions. However, the notions of diversity that have been taken into account so far are still restricted to certain kinds of general content or category similarity, though a large range of more specific types of diversity exist. Besides classical search engines that provide flat lists of search results and that consider diversity only in the ranking of results, there exist faceted search that allows to explore a search result by filtering along some facets. Such systems assign multiple classes to one object which allows ordering in multiple ways.

The classes capture the different facets, i.e., dimensions or features, relevant to a collection. Diederich and Balke considered faceted search as an alternative for keyword search for biomedical literature. However, their facet analysis methods group text only according to topics. Additional dimensions of diversity remain unconsidered. Similarly, the French portal CISMeF (Catalog and Index of French-speaking Health Resources) allows to filter search results along two dimensions: document type (recommendations and guidelines, pedagogical resources, documents concerning patients) and target audience (professionals. students, patients)[9]. Diversity is provided at the result set level while our approach looks into the content of documents and thus, considers diversity on document- or content-level. Hliaoutakis et al. introduce MedSearch, a specialized search engine for medical information that provides diversified search results. For result diversification, web pages are clustered together when they describe the same topic. In the ranking, each cluster contributes at most one page. In our work, we focus on two diversity dimensions related to the content of a document that are analyzed with domain specific features of the medical domain and consider them for ranking. These two dimensions are aspects considered and type of information content and will be introduced in the next section.

To the best of our knowledge, diversity analysis has not yet been considered for the medical domain with respect to such notions of diversity.

b) Diversity in Medical Texts

Definitions and measures in this paper, we focus on two diversity dimensions related to the content of a text: aspects considered and type of information content[2]. To explain these two notions consider the following example: Assuming that there is a blog written by a patient suffering from depression. In some of her posts, she is writing about her daily life, i.e. about experiencing depression, feeling lost and sad. She is providing her experiences in living with that disease. In other postings she presents information on the medical treatments, diagnostic aspects and medications related to this disease[5]. The type of information content of the single posts differs, changing between information and experience. Further, the postings consider different aspects of the disease, which are aspects of the diagnosis, treatment or medication. In

general, we would assume that these two dimensions of diversity are independent from each other. However, it might be that some aspects are discussed rather from a personal view point than others. Future work needs to assess whether these dimensions are orthogonal or whether there are dependencies between aspects considered and the type of information content. To quantify diversity, four measures have been initially introduced and used for analyzing the diversity of medical web content. We can distinguish two categories of information content: factual and affective. They occur due to varying author intents. Correspondingly, two measures have been defined to quantify the type of information content, $degree_{fac}$, and $degree_{aff}$.

The diversity of aspects considered is seen as the variety in medical concepts or their semantic categories, respectively[4]. As medical concepts we consider concepts that are contained in biomedical vocabularies or ontologies such as the Unified Medical Language System (UMLS, http://www. nlm.nih.gov/research/umls/). The UMLS consists of around 1.7 Million biomedical concepts, where each concept is assigned to at least one of the 135 specified semantic types. The semantic types are grouped in turn into 15 semantic groups. These semantic types and main groups are exploited for measuring the diversity of some input text[14]. Two measures are used to describe the diversity of aspect considered, div_{type}, and div_{group}.

ELECTRONIC MEDICAL RECORD (EMR)

EMR is referred to as managing patient medical records electronically from a variety of sources. It deals with patient treatment, diagnosis, laboratory test, imaging, history, prescription and allergies that can be accessed from various sites within the organization with the protection of security and patient privacy[2]. Medical information retrieval is challenging because of the inherent ambiguity within the posed queries. Such ambiguity is manifested in different ways: (1) A query expresses a clearly defined sense, but the genuine needs under this sense may cover a broad range. Taking a common scenario where an ordinary user performs medical search for example, he feels uncomfortable (he has a high fever and rash erupts on his body) but is uncertain about his exact medical problems, so he inputs ‘‘fever’’ and ‘‘rash’’ as keywords into a search engine. In this case, as many diseases may cause these symptoms, the user may prefer to learn knowledge about all these diseases, so as to have a preliminary understanding about his situation and better prepare for the interview with doctors. (2) Query terms themselves are ambiguous, as most users have little medical knowledge. For instance, a pregnant woman feels pain in her abdomen, so she submits a query composed of ‘‘pain in the abdomen’’ and ‘‘pregnant’’. In this case, the term ‘‘pain’’ is ambiguous, which may mean ‘‘stabbing pain’’, ‘‘distending pain’’, ‘‘labor pain’’, etc. The user-cared reasons causing these different kinds of pain, however, may be totally different. In order to enable users to find their interested medical information, from technical point of view, traditional.

a) IR- Information Retrieval

Information Retrieval (IR) is the process of searching within a document collection for the information most relevant to a user’s query. It mainly uses keyword-based query as an input and returns a list of relevant documents as the output. Most searching systems running for traditional document collections use content-based approaches, e.g., the vector space model, Latent Semantic Indexing, or Nonnegative Matrix Factorizations. Since only the internal information of a document is employed to measure the similarity between queries and documents, they are not applicable to handle the complexity of the medical terminologies. Data retrieval (IR) systems utilize uncomplicated data model whereas DB systems is very complex

1) Information is well-ordered as a group of logs.
2) Records are randomly ordered, it is schema-less. IR is mused to retrieve or extract the relevant records based upon the user query. Such as keywords or concepts.

i) Keyword Search

In datatext retrieval, all the words in each of the query log are determined to be the unique keywords. It allows query expansion formed using keywords and the analytical connectives such as: AND, OR, and NOT. Documents are ranked based upon the determination of the pertinent of a query.

ii) Term frequency

Frequency of existence of the query keyword in log documents.

iii) Inverse document frequency

It is a weigh of how copious the word replicate that is, whether the key is familiar or sparse across all the log documents. If keywords in query arise close together in the document, the document has higher importance than if they occur far apart. Documents are returned in decreasing order of relevance score. Usually only top few documents are returned, not all.

iv) Similarity Based Retrieval

Similarity based retrieval retrieve documents similar to a given document. User selects a few closely connected documents from those retrieved by keyword query, and system finds other documents which is similar.

v) Vector space model

It is an n dimensional space, where n is the number of words in the document set. Vector for document d goes from origin

to a point whose i^{th} coordinate is TF (d, t) / n (t). The cosine of the angle between the vectors of two documents is used as a measure of their similarity.

vi) Precision vs Recall Tradeoff

In information retrieval contexts, precision and recall are defined in terms of a set of retrieved documents (e.g. the list of documents produced by a web search engine for a query) and a set of relevant documents (e.g. the list of all documents on the internet that are relevant for a certain topic), relevance. Measures of retrieval effectiveness:

Recall as a function of number of documents fetched,

$$Recall = \frac{t_p}{t_p + f_n}$$

Precision as a function of recall equivalently, as a function of number of documents fetched.

$$\frac{t_p}{t_p + f_p}$$

PROPOSED METHODOLOGY

The proposed model encompasses six major processes such as EMR Pre-processing, Diversification strategy, Meta Map, MeSH ontology, Vector Space Ranking Model and Neural Network based Classifier. Pre-processing strategy is used to analyses the stop word, synonym, and white space present in the user query and the appropriate keyword is extracted from the input medical query. The diversification strategy involves four steps: Query understanding, query transformation, candidate concept mapping and derived query generation. Meta map concept identifier is used to map the biomedical terms to MeSH (Medical Subject Heading) concepts. The extent of VSM is used to constitute the EMR log as vectors. Each group of words consists of multiple concepts and words. The documents are ranked based on its importance. The importance of Neural Network based Classifiers used to train and test the medical documents and queries to predict the output based on the similarity between the document vectors and query vectors.

a) The Diversity-Aware Medical Search Approach

Our proposed approach on diversify-ware retrieval of medical records includes two steps, i.e., (1) query understanding to discover the implied aspects of the original query as multiple sub-queries; (2) diversity-aware medical retrieval to exploit multiple sub-queries to for diversifying the medical search results. The following of this section gives a detailed description on each of the two steps.

i) Query understanding

Since the keyword query is a simple and user-friendly search model, it is prevailing in many practical search systems. Our research assumes to use a keyword-based interface for the users to express their information needs and returns a list of relevant EMRs as the output.

The list of keywords in the query can be interpreted diversely, we need to handle the ambiguity problem, i.e., understand the meanings of the concepts specified in the user's queries and discover the potential aspects of the given query. More specifically, given a query q containing a list of keywords, the task of query understanding is to transforms it into a set of derived queries to model different aspects of q. As medical ontology contains rich and accurate professional knowledge that is shared by domain experts, we use it as background knowledge to uncover the underlying aspects of information needs. The detailed query understanding process contains three sub-steps as below.

ii) Query transformation

This sub-step carries out two functions, i.e., keyword phrase identification and expansion. With the support of available semantic resources, e.g., WordNet and Consumer Health Vocabulary (CHV), the former uses the maximum matching approach to scan the keywords in the query sequentially and find the longest matching subsequences defined in the semantic resources as the keyword phrases. For example, given a query ''difficulty breathing headache'', the longest maximum matching approach can find ''difficulty breathing'' as a keyword phrase and ''headache'' as the other keyword phrase. For the latter, two types of expansions are conducted. On the one hand, the layman keywords input by lay persons should be mapped to professional medical terms, for examples, ''difficulty breathing'' is rewritten to ''dyspnea''. As previous researches demonstrated that professional terms were likely to achieve better search results than layman terms. We employ the CHV, which provides a mapping between medical terms and layman terms, to perform this expansion. On the other hand, the input keywords (even the professional medical terms) may have synonyms. For instances, the distinction between ''diagnosis'' and ''finding'' is not clear, and ''fever'' is a synonym of ''febrility''.

Meta Map MetaMap is a device formed by NLM that plots free script to medicinal ideas in the UMLS, or equally, it determines metathesaurus ideas in script. Meta Map is a method to recognize entities from raw text by mapping them to MeSH terms with a scoring system. Take "mammary cancer" as an example, Meta Map will not only map entities to the MeSH term Malignant Neoplasm of Breast, but also provide information on the source vocabularies from which the term is originated. In this case, it is the MeSH term Breast Cancer, therefore, one can use this Meta Map feature to identify hierarchically related entities, which is exactly the

main idea behind the first approach. However, it easier to further improve its performance.

Meta Map based MeSH includes following steps,

- Entity processing.
- Apply Meta Map to processed entities.
- Generate candidate mappings from Meta Map results.
- Choose final mapping from candidates.

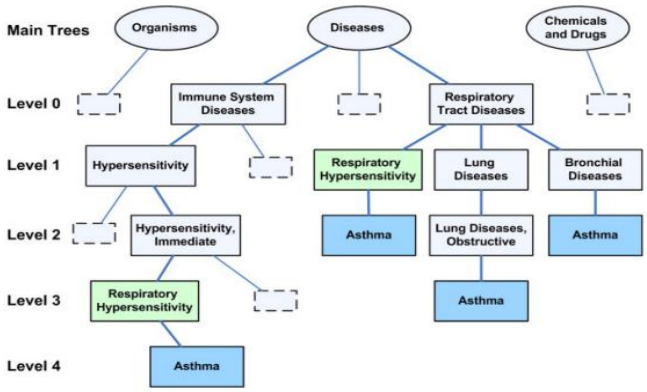

Figure 2: MeSH Tree

iii) Replacement of Greek Letters

Replace Greek letters by their full English names (e.g. α by alpha). Greek letters and their full names are used interchangeably in articles, but only their full names are used in MeSH.

iv) Extraction of non-English words

Some chemicals and proteins are described by non-English words, such as "3-chloro-1, 2-propanediol" and "IGF-1". Those words have exact matches in MeSH. It uses several features to identify non-English words, such as the presence of numbers, capital letters, and special characters (e.g. hyphen and comma).

b) The Diversity-Aware Retrieval Engine (DARE)

Briefly, the system works as follows: From a set of relevant online sources new content is collected regularly. The diversity of the single texts is assessed by calculating diversity measures. Given a search query provided by a user, the system retrieves medical blogs and other social media content matching the query from the previously collected and indexed data. The diversity measures are exploited when the result set is ranked and presented to the user.

The diversity-aware retrieval engine is implemented as a service-oriented architecture. The search interface allows the user to interact with the system. The server is responsible for triggering services in the correct order and for the communication between user interface and services.

We can distinguish four types of services:

- Collection Service: Content Collector and Indexer
- NLP Services: Domain Filter, Concept Annotator, Diversity Assessor
- Result Preparation Service: Ranking
- Presentation Service: Visualization

Collection Services collect content from the web. NLP Services aim at processing and analyzing the natural language from input documents. Result Preparation Services filter irrelevant results or rank the results according to user needs. Presentation Services are responsible for preparing and visualizing the results. Several resources are used by the system including the UMLS, a list of URLs of relevant sources and training material for the algorithms. The components are described in more detail in the following.

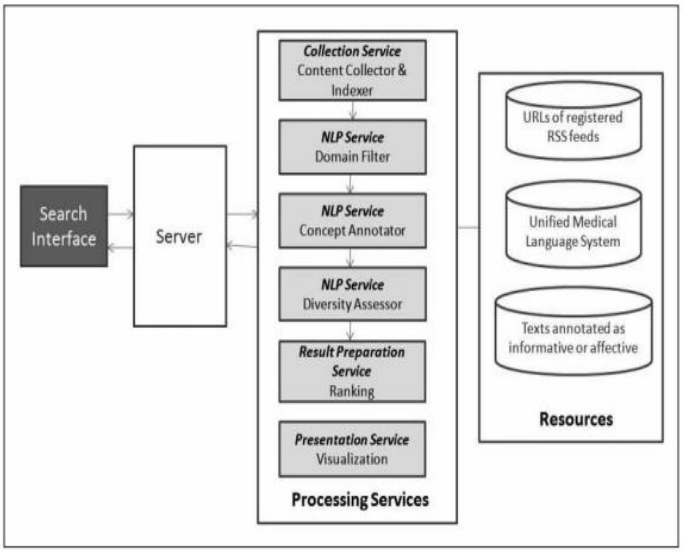

Figure 3: Processing of Diversity-Aware Retrieval Engine (DARE)

i) Diversity Assessor

The diversity assessor analyses the collected content with respect to the aspects considered and the diversity of information content. It exploits the number of semantic types and main categories determined by the Concept Annotator for calculating the diversity measures: div_{type}, div_{group}, $degree_{fac}$, $degree_{aff}$. A high diversity in aspect considered is reflected by a large variety in semantic types or main groups. This is considered in formulae that calculate the proportion of different semantic types (main groups) contained in a text to measure diversity. A value close to 1 indicates a high diversity, while a value close to 0 corresponds to a small diversity. The number "135" refers to the semantic types provided by the UMLS, while "15" is the number of UMLS semantic groups.

$$\mathrm{div}_{type} = \frac{types}{135}$$

$$\mathrm{div}_{group} = \frac{groups}{15}$$

This decision was made since we had in mind users that search for information on disorders. It is clearly possible to broaden the scope of the retrieval engine and consider also other UMLS categories when determining the medical content of a text.

Additionally, this service distinguishes factual from affective postings. Affective parts of a text are reflected by opinionated words. To count the opinionated words in a text, words that are neither medical content nor stop words can be looked up in SentiWordNet. The measures $degree_{fac}$ and $degree_{aff}$ are exploited together with the number of words and the number of stop words as input for a supervised machine learning algorithm. Through experiments with different machine learning algorithms implemented in the SimpleLogistic classifier has been chosen since it outperformed NaiveBayes and other algorithms. The algorithm performed with 86.5% accuracy in 10-fold cross validations on 750 factual and 750 affective blog postings. This text material has been classified manually and also provides the training material within our Diversity Assessor.

ii) Ranking

The results matching a query are ranked considering the diversity measures divtype and divgroup as boosting factors. The main assumption is that postings with a larger diversity in types and groups are of higher interest to the user than those with a smaller diversity. In our experiments, this assumption will be studied.

iii) Visualization

The user interface presents the ranked results. It consists of a single line text field for the query and a result section. Factual and affective texts can be shown separately. In addition, percentages are listed for the categories disease, treatment and medication. They show to what extent a posting considers the single aspects and thus allow to quickly judge upon the general theme of a posting with respect to a query.

CONCLUSION

In this paper, an architecture for a diversity aware retrieval engine for medical web data has been introduced. We centered our assessments on retrieval of medical social media data, a still relatively unexplored domain. Diversity measures that consider medical concepts mentioned in a text and their categories are used to rank retrieval results. The evaluation results suggest that the diversity measure reflecting diversity of aspects considered are well suited for supporting ranking. It could be shown that users are satisfied with a result set when diverse texts are shown in the top N positions. Our assumption that we can increase user satisfaction by ranking texts that have a higher diversity in higher positions has been proven correct.

The described approach to diversity aware ranking (and retrieval) has been proven successful in improving user satisfaction. The users provided the feedback that some postings they had to judge were only advertisements with medical keywords. The filtering algorithms need to be adapted to filter out such non-sense postings in advance. The evaluation is limited in a way that for some queries only a small number of texts could be retrieved from the data set. Further, the data set was quite small. However, from the queries where more than 300 results were retrieved, we learned that user satisfaction with the ranking is even better than for small result sets; the annotator agreement is higher. This makes us confident that the approach will perform well on larger corpora which will be assessed in future research.

REFERENCES

[1] Betin.A, MedicoPort (2007): A medical search engine for all, Computer Methods and Programs in Biomedicine 86 (April) 73–86.

[2] Carbonell.J, Goldstein.J(1998) The use of MMR, diversity-based re ranking for reordering documents and producing summaries, in: Proc. of SIGIR'98, 335–336.

[3] Daumke P, Markó K, Propat M, Schulz S, Klar R. Biomedical information retrieval across languages. Med Inform Internet Med 2007; 32 (2): 131 –147.

[4] Daumke P, et al. Subword-based semantic retrieval of clinical and bibliographic documents. Methods Inf Med 2010; 49 (2): 141 –147 .

[5] Darmoni SJ, Leroy JP, Baudic F, Douyere M, Piot J, Thrion B. CISMeF: a structured Health resource guide. Methods Inf Med 2000, 39 (1): 30 – 35.

[6] Dakka W, Ipeirotis PG. Automatic extraction of useful facet hierarchies from text databases. In: Proc. of ICDE" 08. Washington, DC, USA; 2008. pp 466 – 475.

[7] Diederich J, Balke WT. Automatically created concept graphs using descriptive keywords in the medical domain. Methods Inf Med 2008; 47 (3): 241– 250.

[8] Denecke K. Diversity in medical social media data: Approaches, study and future challenges. International Journal of Computational Linguistics Research 2010; 1(1): 1– 11.

[9] Hearst MA. Clustering versus faceted categories for information exploration. Communications of the ACM 2006: 49 (4); 59 – 61.

[10] Järvelin K, Kekäläinen J. IR evaluation methods for retrieving highly relevant documents. In: Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval; 2000. pp 41 – 48.

[11] Jonquet C, Shah NH, Musen MA. The open biomedical annotator. Summit on Translat Bioinforma 2009. pp 56 – 60.

[12] Krallinger M, Valencia A. Text-mining and in - formation retrieval services for molecular biology. Genome Biol 2005; 6 (224): 1– 8.

[13] Luo G. Design and Evaluation of the iMed Intelligent Medical Search Engine. In: Proceedings of the 2009 IEEE International Conference on Data Engineering (ICDE" 09). IEEE Computer Society, Washington, DC, USA; 2009. pp 1379 –1390.

[14] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Stud Health Technol Inform 2001; 84 (Pt 1): 216 – 220.

[15] Radlinski F, Craswell N. Comparing the Sensitivity of Information Retrieval Metrics. Proc. SIGIR 2010. pp 667– 674.

[16] Vanhecke T, Barnes M, Zimmerman J, Shoichet S. Pubmed vs. highwire press: a head-to-head comparison of two medical literature search engines. Comput Biol Med 2007; 37 (9): 1252 –1258.