SEMANTIC MATCH MAKING AND RANKING OF MEDICAL LITERATURE

By

PRIYA WADHWA

(Under the Direction of I.Budak Arpinar)

ABSTRACT

The amount of data and information in the field of medical sciences is increasing at a tremendous rate. PubMed comprises more than 21 million citations from biomedical literature. There are various new discoveries and researches carried out in the field of biomedical sciences almost every year. All this information is really important and useful for the patients as well as medical practitioners and should reach them on time so that they can carry on the correct treatment for a particular patient based on the new knowledge. The project is an attempt to carry out the match-making of a patient's profile with the various medical publications and ranking them based on the semantics involved.

INDEX WORDS: Semantics, Ontology's, PubMed, Match-making, Ranking, and Biomedical.

SEMANTIC MATCH MAKING AND RANKING OF MEDICAL LITERATURE

By

PRIYA WADHWA

B.E. COMPUTER SCIENCE, M.DU. ROHTAK, INDIA, 2008

A Thesis Submitted to the Graduate Faculty of the University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2012

SEMANTIC MATCH MAKING AND RANKING OF MEDICAL LITERATURE

By

PRIYA WADHWA

Major Professor: I. Budak Arpinar

Committee: Thiab Taha

Lakshmish Ramaswamy

Electronic Version Approved:

Maureen Grasso

Dean of the Graduate School

The University of Georgia

MAY 2012

DEDICATION

This thesis is dedicated to my parents who supported me throughout my life and have helped me in gaining my education. Thanks Mom and Dad.

ACKNOWLEDGEMENTS

I would like to thank Dr. Budak Arpinar for being an encouraging, motivating and supportive advisor over the past one and a half year and for inspiring me to work in the field of semantic web. I would also like to thank Dr. Thiab Taha and Dr. Ramaswamy for all their help and valuable suggestions. I would like to thank Asmita Rahman with whom I worked on this project .I would also like to thank the staff and faculty members of Computer Science department of University of Georgia for directly or indirectly supporting my work. I would also like to thank my friends and family for all their inspirational support.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 Introduction

The amount of data and information in field of medical sciences is increasing at a tremendous rate. PubMed comprises more than 21 million citations [1] for biomedical literature from MEDLINE, life science journals, and online books. These citations can be the links to full-text content from PubMed Central and publisher web sites. The number of citations in Medline grew by more than 700,000 in 2009.This huge amount of increase in data and information includes many articles or journals on discovery of new treatments, medications, clinical trials, research carried out on the already present drugs, their reactions or allergies. All this information is really important and useful for the patients as well as medical practitioners and should reach them on time so that they can carry on the correct treatment for a particular patient based on the new knowledge.

Since the amount of data and information related to biomedical sciences is huge and the available search engines are based on the keyword based search there is a possibility that the some of the citations might be missed by the user as there is a common tendency that user usually reads the first few articles and doesn't read the rest, because there is usually a notion that the first few articles would be more relevant as they have a higher rank as compared to the others. Hence, there is a possibility that one might miss the rest of the articles. So, a proper search system as well as a ranking system is required which is not only based on the keyword-based search but also takes into consideration the *semantics* involved in the medical literature.

The medical knowledge dissipation usually occurs through the conferences, articles, research papers and journals. The evolution of web has made it a bit easier for the users and medical practitioners as well researchers to get the information through various medical search engines but this process is passive and has a lot of limitations such as:

- The user has to search the online databases periodically to keep him/her updated about new medical discoveries and knowledge.
- The modes of dissemination of medical literature are very limited. The user has to search a large number of web sites to get the information.
- The time by which the doctor gets aware of the new research depends on how often he accesses the databases.
- In combination with all the above mentioned points there is one more possibility that even if the doctor is aware of the new medical discovery; its again a time consuming process for a doctor to figure that the new medical information or discovery is related to which particular patient(s).

All these limitations mark the necessity of coming up with a proactive medical information dissemination system. The need is not only to pull out the relevant information for the patients' health records and the medical course followed by them from the database but, an intelligent system is required that could actually understand the relations among the different terms and text in the doctor's prescription and which would further help in retrieving the most suitable and relevant citations for a particular patient profile and rank the various medical publications based on the semantics involved in the medical data. Our project incorporates such a system which uses

semantics enabled framework for the retrieval, ranking and distribution of relevant medical information.

1.2 Goals

Our project has the following primary goals:

- Use of the semantic relationships : This is about making the match- making and ranking process not just the keyword- based but also including the semantic relationships among the terms used in the medical knowledge to get much more precise and relevant results.
- Reduce the information overload: The research in medical field is advancing and also there are lots of conferences held every year all over the world regarding the new discoveries made. This might make some of the practices, treatments and medication followed by the doctors obsolete. To be updated with the medical information and the ongoing research in the biomedical sciences the user whether it's a patient or a doctor has to download all the publications on their machine which may result in information overload. Our project proposes a system that reduces this hustle of the end user by providing them with the relevant medical knowledge based on the patients' profile.

- **1.3 Challenges**

The following are some of the notable challenges that we faced during our research:

- The Electronic Medical Health records are kept strictly confidential and are not available for public access. We generated our own sample health records after

looking at a few sample EMR's at the Google Health and Microsoft Health Vault.

- EMR's and the medical research publications have to be processed automatically to find the various entities and concepts in them and also the inter-relationships among the concepts and terms. We used the NCBO bioportal annotator for annotating the medical publications and the UMLS (Unified Medical Language System), a very large ontology for biomedical sciences concepts and terms to find out the semantic relationships among the various terms and concepts.
- The keyword- based search which is used in most of the medical knowledge databases such as PubMed is inadequate. We developed a search and ranking technique based on the deeper semantics involved in the medical information and data.

1.4 Motivation

The following example scenario which is also one of the test cases for our project provides the motivation for carrying out the research in the semantic enabled framework for medical knowledge dissipation system.[2]A few years back doctors used to prescribe a drug named Plavix to avoid heart attacks among the patients. According to the information published in an article named"Plavix Drug Information: Uses, Side Effects, Drug Interactions and Other Warnings"

Plavix was launched in 1998,and it has been used to help protect against future stroke or heart attack.Over 11 years, doctors have prescribed this anti-platelet medication to help over 100 million people worldwide, and thus Plavix plays a crucial role in reducing

the future risk of stroke or even heart attack. But according to a recent research Plavix may result in the second heart attack to the patients who have problems of acidity or are suffering from stomach ulcers.[3] The U.S. Food and Drug Administration (FDA) has placed a boxed warning to the label for anti-blood clotting drug Plavix by March 2010, as a measure to alert the consumer that this drug can be less effective for the individuals who cannot effectively metabolize it to its active form. In this case, patients who have been identified of “poor metabolizers” (who carry a variant CYP2C19 gene that affects the enzyme to convert Plavix into its active form) may need an alternate treatment. Besides, adding the new “black box” warning on its normal dose, which has a potentially deadly lack of effect in 2% to 14% of patients, FDA also wants doctors to discuss Plavix options with patients. To avoid these side effects doctors used to provide patients with drug named as “Prilosec”. But, then it was observed that after few years that the reaction of this both of the drug killed many patients. After searching through many websites related to health records and journals we found that actually the paper was published a long back on this information that the combination of both drugs is very harmful for the patients. Table 1 shows the side effects of Plavix in combination with various drugs. This information was not disseminated to doctors properly which caused many people to die as it was prescribed for few years after paper was published. Thus this lack of information among doctors became the reason of the death of several patients. If this information reached the doctors on time the doctors who had prescribed the patients with both the medications would have stopped that course of medication and have prescribed something else and this way the life of those patients could have been saved. This particular example tells us how important it is for the

medical knowledge and discoveries to be disseminated on time to all the end users. Our system is an attempt to provide the end user(a patient as well as his/her doctor) with the most relevant information related to his/her medical profile and also ranking the publications based on the semantics involved in the publication as well as the medical profile.

BODY SYSTEM EVENT	PLAVIX+ASPRIN(N=6259)	PLACEBO+ASPIRIN(N=6303)
Body as a whole chest pain	2.7(<0.1)	2.8(0.0)
Headache Dizziness	3.1(0.1) 2.4(0.1)	3.2(0.1) 2.0(<0.1)
Abdominal pain Dyspepsia Diarrhea	2.3(0.) 2.0(0.1) 2.1(0.1)	2.8(0.3) 1.9(<0.1) 2.2(0.1)

Table 1: Adverse events occurring in >=2.0% of Plavix patients in cure

Source: Plavix Drug Information: Uses, Side Effects, Drug Interactions and Other Warnings

1.5 System Overview

A very brief overview of the system is described in this section. A detailed workflow is described in the later part of the thesis. EMR's are processed and semantically enriched with named entities and relationships and the result is a semantic graph. The resulting semantic graph is used for clustering and query generation. Scientific literature goes

into a similar process and is annotated using named entities and relationships resulting into various semantic graphs. The queries are run in the semantic match-making phase and the relevant literature is ranked and the results are displayed to the end user.

CHAPTER 2

RELATED WORK

2.1 Google Page Rank

Page rank is the most popular document ranking algorithm which is used by the Google search engine to rank the pages on the web . Page Rank was developed at Stanford University by Larry Page and Sergey Brin as part of a research project about a new kind of search engine. It's a link analysis algorithm named after Larry page[4]. It assigns a numerical weight to each element of a hyperlinked set of documents or pages to measure the relative importance of each page in a given set. The algorithm can be applied to any number and any set of documents with reciprocal links and references. A reciprocal link is a mutual link between two objects or entities, commonly between two websites to ensure mutual traffic. The numerical weight that is assigned to a document or entity or page is referred to as its page rank.

The page rank is based on the mathematical algorithm which is applied to a graph basically the web graph that is created by the World Wide Web pages as nodes and hyperlinks as edges. The value of the rank for a particular page specifies its importance. If a page has a hyperlink then that will further count towards its rank. The rank of a page is calculated recursively and depends on the number and rank of all the pages that link to a particular page. A page that has links to the pages with high page rank will get a higher page rank. The page rank takes into consideration two important things: firstly it applies the standard citation technique to the web structure i.e. every link

can be considered as an academic citation so a major page like www.uga.edu will have several other links pointing to it which may be considered as the citations. Second thing is the link structure of the web. Every web page has several outgoing and incoming links. The all incoming links for a particular page cannot be determined but if a page has been downloaded we can know its entire outgoing links. Page rank takes these two things into consideration to calculate the importance of a particular web page.

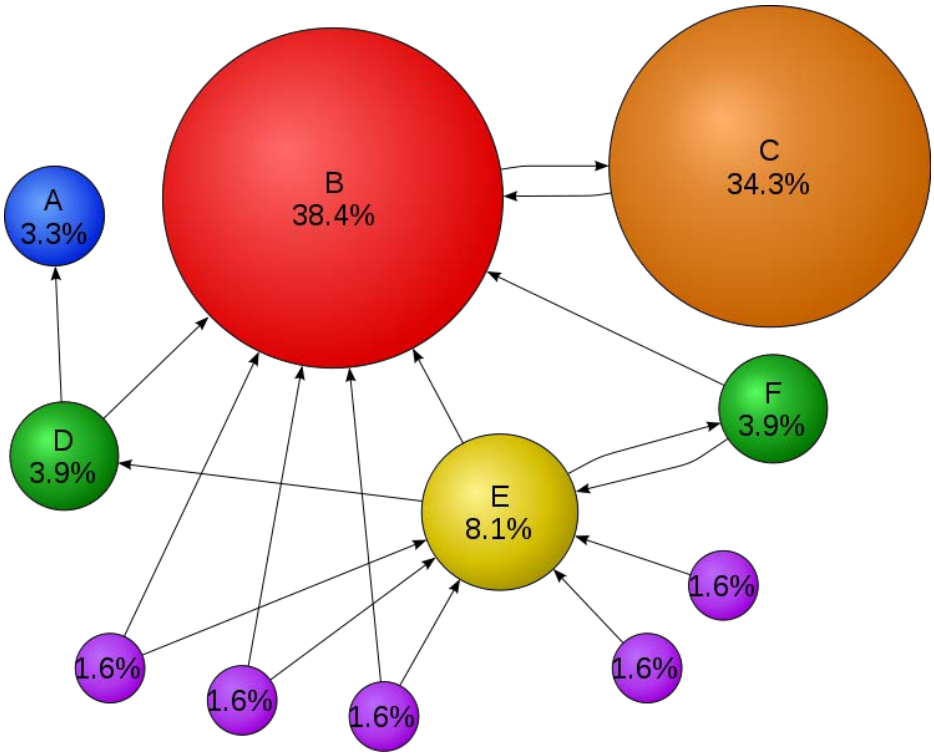


Figure 1: The page rank for various pages

Source: http://en.wikipedia.org/wiki/PageRank

The figure 1 above shows the page rank for various pages A, B, C, D, E and F out of 100.From the figure we can see that the page rank for page C is higher than the page rank for page E even though page E has more incoming edges than page C but the rank of the page with which C has link to has more rank than the combined rank of the

various pages to which page E is linked with. Thus the overall ranking of a particular page not only depends on the popularity of the page but also the ranks of the page that are referring to that particular page.

2.2 TRUST RANK

As we all know that now days the number of spam on the web is increasing at a tremendous rate and hence something is needed to avoid getting the spam as one of the results for a search query and displaying it to the user. [5]Most of the spams are created with the intention of misleading the search engine. These pages, chiefly created for commercial reasons; use various techniques to achieve higher-than-deserved rankings on the search engines' result pages. While human experts can easily identify spam, it is too expensive to manually evaluate a large number of pages. There are a number of web spamming techniques like adding some keywords to a web page which is not actually related to those keywords and hence when user inputs a query which has any of those keywords search engine would display that page as the result. Another method is by creating some meaningless links to a page and hence as the page rank gives a higher rank to page with more number of incoming links that particular page will receive a good rank although it has nothing important information in it. Detecting spam is not an easy task for the computer. Many search engine companies have employed staff to detect the web pages that are spam and as soon as a spam is detected the search engine stops crawling it and it is no longer indexed.

Trust rank is again one of the link analysis techniques for semi automatically separating the useful web pages from the spam by assisting human experts who detect the spam. The algorithm does not operate in isolation but involves human assistance. The

algorithm first selects a small seed set of pages whose spam status is to be determined. The human experts then identify the pages as spam or not spam (good ones). Finally the algorithm identifies the other pages as the good ones based on their links with the good seed pages.

2.3 PAGE RANK AND TRUST RANK Vs SEMANTIC RANKING

The above mentioned ranking algorithms are the ones that are used by the most popular search engines. The criteria used for ranking the pages on the web used by the algorithms involve the link structure of the web. The ranking doesn't take into consideration the semantics involved in the various pages or documents on the web. Compared with keyword-based search, semantic search and ranking seeks to improve search and ranking accuracy by understanding searcher intent and the contextual meaning of terms as they appear in the searchable data space within a closed system, represented as an ontology model, to generate more relevant results. The various search engines and the information retrieval systems are focusing on providing an efficient and intelligent system for answering user queries that takes into the consideration the semantic concepts and the various semantic relations involved and not just the keyword based search. So the ranking of documents is required in such a way that not only the syntactic information is considered while ranking but also the semantics of the various terms in the query as well the related document are brought into the picture and hence the documents ranked based on that. Our ranking algorithm takes the various possible semantics involved in the patient's profile and the medical publications and hence ranks the documents accordingly for a particular patient profile. Thus now the ranking is not only based on just the frequency of terms in the document

that are involved in the query but on the semantic concepts and the relationship of the terms are given importance while ranking. The annotations of the medical publication as well as the patient profile help in getting the semantic relationships involved and hence a better rank is assigned to each medical paper based on the user's health record.

2.4 SEMANTIC RANKING AND RESULT VISUALIZATION FOR LIFE SCIENCES

The domain of life sciences is experiencing an unprecedented growth. This ever increasing amount of data and information in life sciences requires the development of new semantically enriched data management that facilitates ease of scientific information retrieval and hence provide efficient results. Literature search is the most important task in scientific research. One of the most widely used database for articles on life sciences is PubMed with over millions of articles and this number keeps on increasing with time. The articles in the PubMed are annotated by a staff of indexers with terms from the Medical Subject Headings (MeSH) controlled vocabulary. MeSH organizes term descriptors into a hierarchical structure,allowing searching at various levels of specificity. The MeSH terms are basically organized into IsA hierarchy.

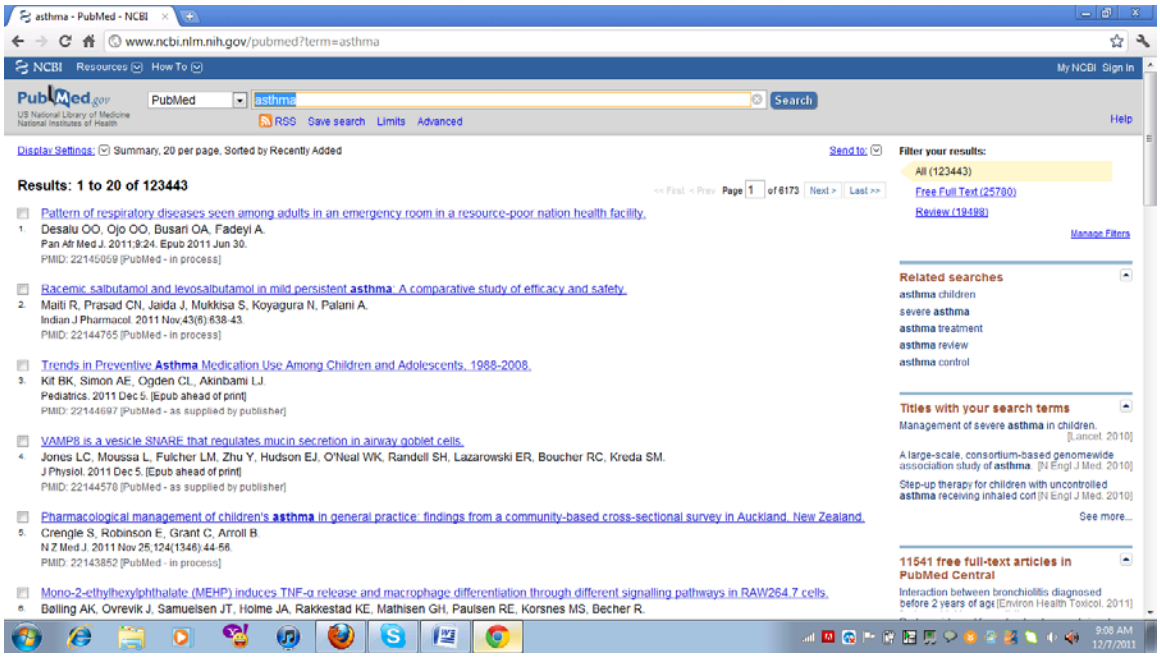


Figure 2: Snapshot of PubMed

Figure 2 is the snapshot of the query submitted in the PubMed and the results obtained for the disease Asthma.

The paper “Semantic Ranking and result visualisation for life sciences publication” is a semantic approach for ranking the PubMed articles. In this work, several ways to measure semantic relevance of a document to a query are proposed, and also how their semantic relevance can be computed efficiently on the scale of PubMed and MeSH is demonstrated.

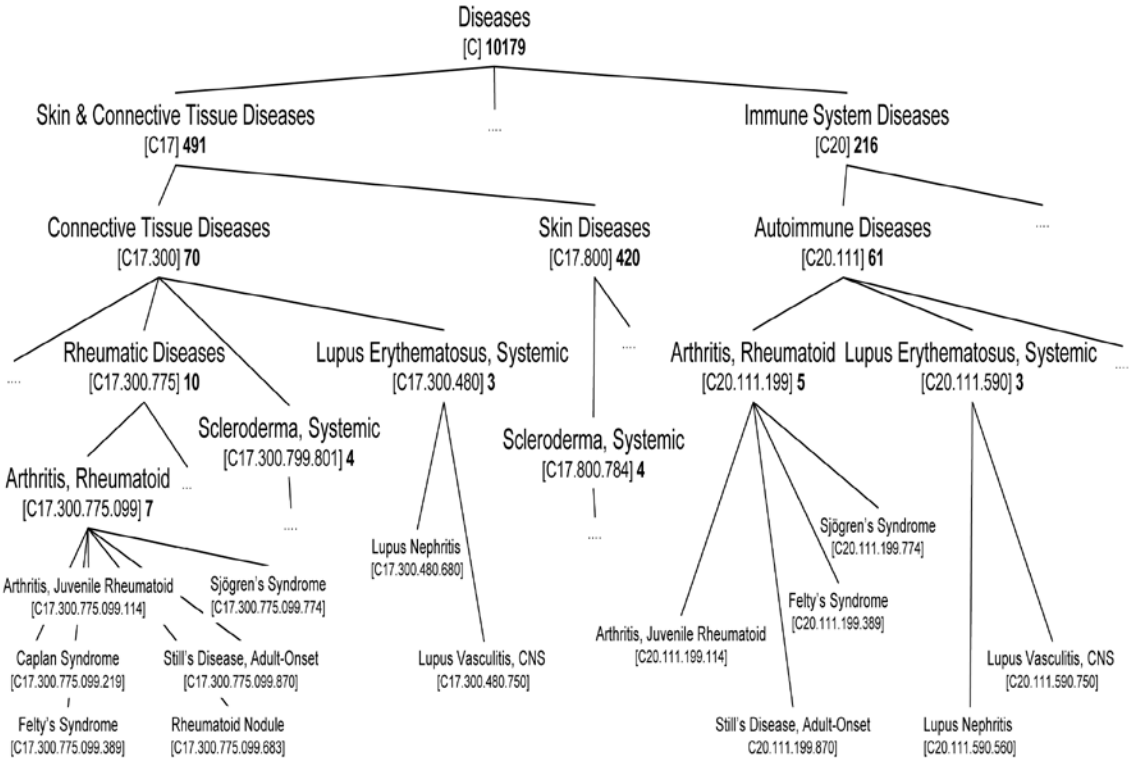


Figure 3: Mesh Polyhierarchy

Figure 3 explains the organization of terms in the Mesh vocabulary and describes how the terms and concepts are arranged in a tree and related to each other based on the hierarchies of the terms. Each term is represented as a concept in the tree structure and each concept has a unique concept id that starts with alphabet C followed by the number. The algorithm in the paper exploits this tree structure for ranking of the documents.

The following are the semantics of query relevance that are taken into consideration in the work proposed by the authors Julia Stoyanovich 1, William Mee 2, Kenneth A. Ross from Columbia University ,New York USA of this publication.

Motivation: A score is assigned to the documents whose MeSH terms overlap with the query terms. So the first similarity measure counts the number of terms that are common in the query and the concepts and sub concepts of the MeSH terms. If a query is {A,B} in Figure 2, and the document contains MeSH terms C and D, then both C and D contribute to the overlap because they are sub concepts of A and B.

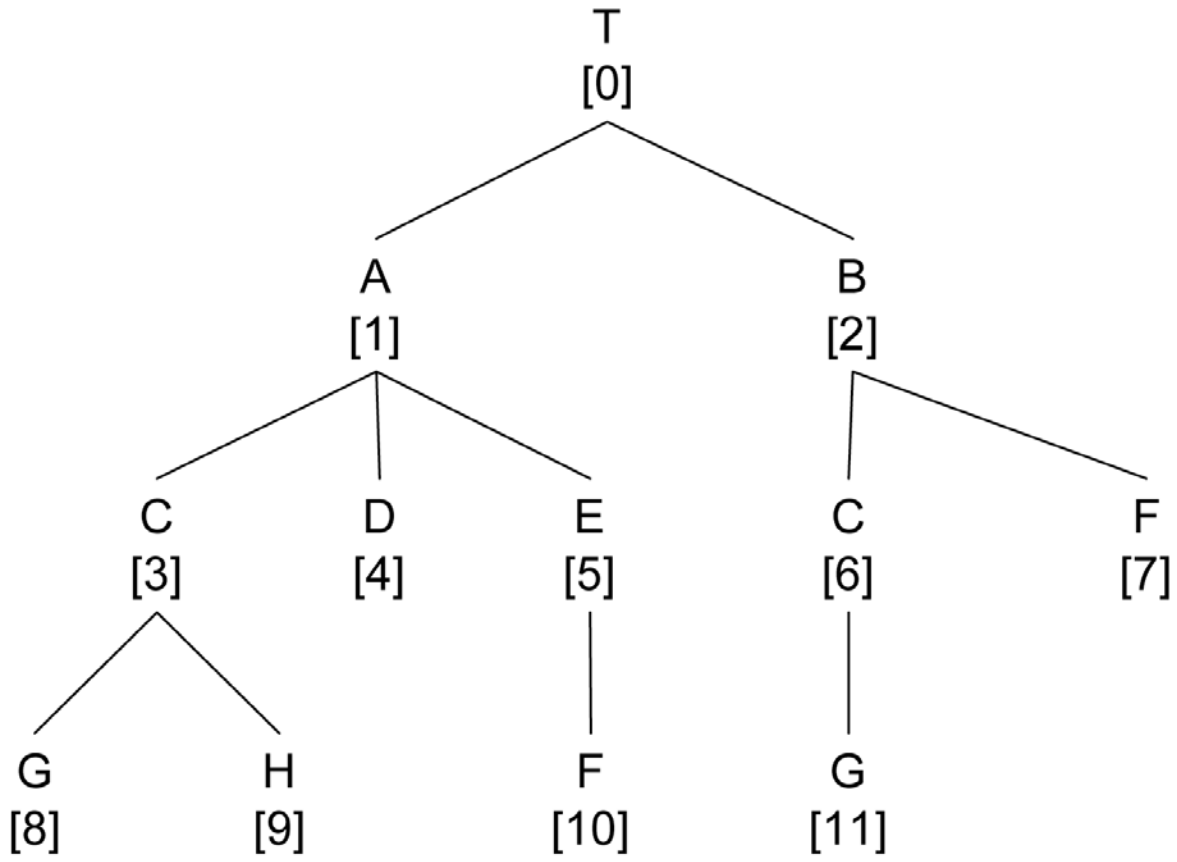


Figure 4: Scoped hierarchy.

The figure 4 explains the hierarchy of the various concepts that is taken into consideration while ranking the documents by the algorithm presented in the paper.

In the paper the focus is on queries that are conjunctions or disjunctions of MeSH terms, and rely on the query processing provided by Entrez to retrieve query matches.

Note that, while the query semantics is Boolean, it incorporates ontology expansion, blurring the line between strictly Boolean and set oriented processing. So, a document D will match a query Q = {q1, q2} if D is annotated with at least one term in the term-scope of each of q1 and q2.

.The algorithm receives as an input the similarity measure for a document based on some criteria, a sorted list of documents sorted on the basis of publication date, a query Q and an integer k that tells the number of skyline contours to be computed and assigns a rank to that document.

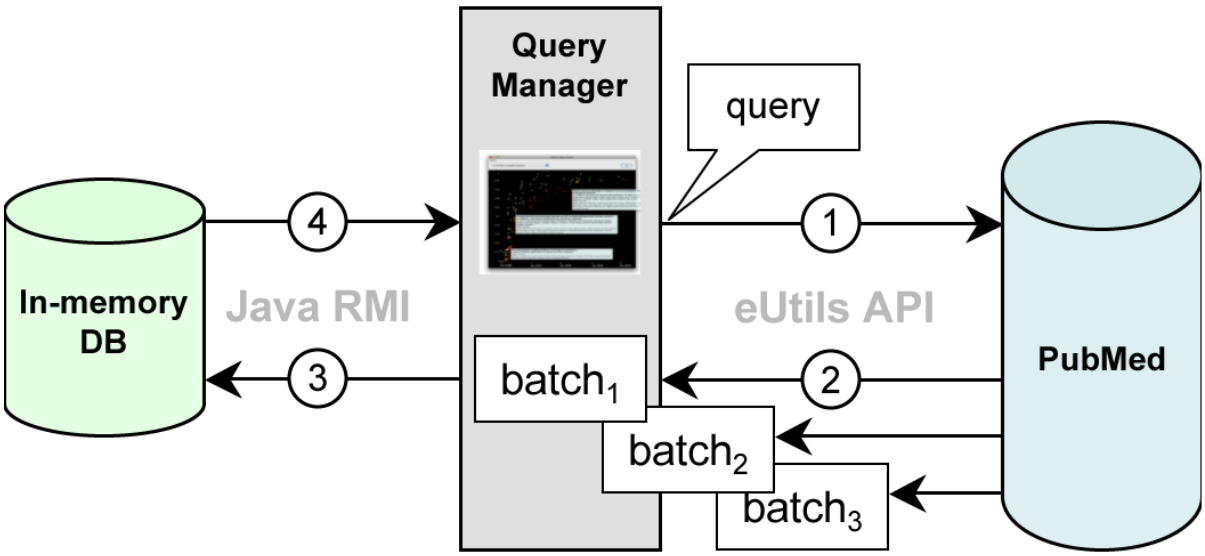


Figure 5: System Architecture

Figure 5 gives a big picture of how the documents are processed and ranked using the algorithm presented in the paper.

The authors have attempted to capture the semantics of a term by looking at all of the term's descendants, across the whole hierarchy of the MeSH terms. The algorithm developed three similarity measures that relate sets of terms based on the degree of overlap between the sets of their descendants. The question of how contributions of different terms, or different meanings of the same term, are reconciled in the final score is central to the above approach. Hence the term based similarity, the synonymy are used for the ranking of the documents based on the above approach.

2.5 SEMANTIC RANKING AND RESULT VISUALIZATION FOR LIFE SCIENCE Vs OUR SEMANTIC RANKING ALGORITHM

The work presented in the above paper takes into consideration the concept hierarchies involved in the Mesh vocabulary for ranking the medical publications and the results were compared with the PubMed ranking although they were better but still the algorithm they use doesn't exploit the various other semantics that may be present in the medical publications for ranking the documents. Our algorithm takes into consideration the various concept hierarchies obtained from the annotation of the medical publications and the various semantic relationships among the terms involved that the annotations provide such as synonymy of the terms. We also consider the publication date of the papers for ranking them as the research in the medical field is growing at a tremendous rate and hence the recent publications should be given a higher rank than the older once so that the user is aware of the latest advances in the medical sciences. We rank the documents based on the user's health record and hence involve the other semantics that may relate a medical paper to the patient's health

record such as the medications, symptoms etc to present the user with the more relevant results and not just doing the keyword based search and ranking.

2.6 DISCOVERING AND RANKING SEMANTIC ASSOSCIATIONS OVER A LARGE RDF METABASE

The paper describes how the semantic associations among the entities can be found and then ranked. The semantic associations were found in the SWETO (Semantic Web Technology Evaluation Ontology) which has 800,000 entities and 1.5 million explicit relationships among them. The user query about the semantic associations between two entities may result in hundreds of results and hence the paper presents an algorithm to rank these results or associations before presenting them to user so as to provide relevant results. The criteria used for ranking the associations in this paper are: Path length, context, subsumption (from more specialized ones to the general entities) and trust. The system has a web interface where a user can select the two entities and define the context of his query and hence the results of the ranked path are displayed according to the inputs provided by the user.[7]

The work presented in the paper is a good approach towards ranking of the semantic associations. The work that we have done is different from this paper as we are going to rank the medical publications on the whole after considering the semantic relationships among the patient's profile and the relevant match in the publications and also the various semantic relationships involved in the query terms and the annotations obtained from the medical papers. So it's not just ranking the associations but the document on its whole by considering the semantic relationships involved in the document. So our

work presents the semantic ranking of the documents in our case the domain of interest is medical publications so we take into consideration the semantics that can be relevant to the medical publications and the patient's profile.

2.7 GOOGLE HEALTH

Google health is the service provided by the Google to the users in 2008.The service facilitates the user to manage their health records either manually or by logging into their accounts at partnered health services providers – into the Google Health system, thereby merging potentially separate health records or creating one centralized Google Health profile. [8]

The information can include the health condition, allergies, medications, symptoms etc. The Google health uses the information entered by the user and provides the user with a merged health record, information on conditions, and possible interactions between drugs, conditions, and allergies.

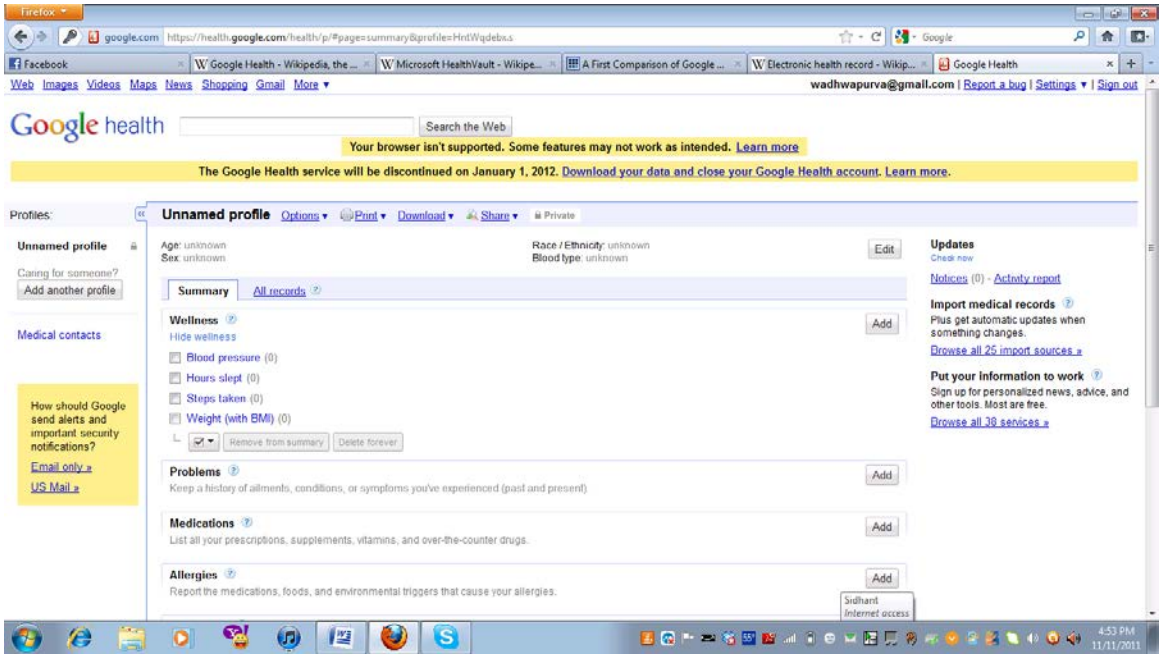


Figure 6: A screen shot of Google Health web interface.

These are the features provided by the Google Health:

- Google Health helps by offering a single place to organize and store a user's health information online. Track one's wellness metrics, gather and organize one's medical records, or import the health data directly into your account from connected doctors, hospitals and retail pharmacies.

- Google Health allows a better way to track a user's goals for weight, blood pressure, or other wellness metrics. Provides a feature to track the sleep patterns, record how much a user walks during the day. With Google Health one can set personalized goals online and monitor them regularly.

- Create custom trackers for things one wants to monitor like daily sleep, how much coffee one drinks in a day, or how many times one exercises a week. One can also take notes or keep a diary on how one is doing with a particular medical condition or a personal goal one sets.

2.8 MICROSOFT HEALTH VAULT

Microsoft health vault is another web based platform to store and maintain health information. A Health Vault record stores an individual's health information. The access to a particular record is through a Health Vault account. The information is kept confidential and no one else can access the health record of a particular person. However some accounts are authorized to access records for multiple individuals, so

that a mother may manage records for each of her children or a son may have access to his father's record to help the father deal with medical issues.[9]

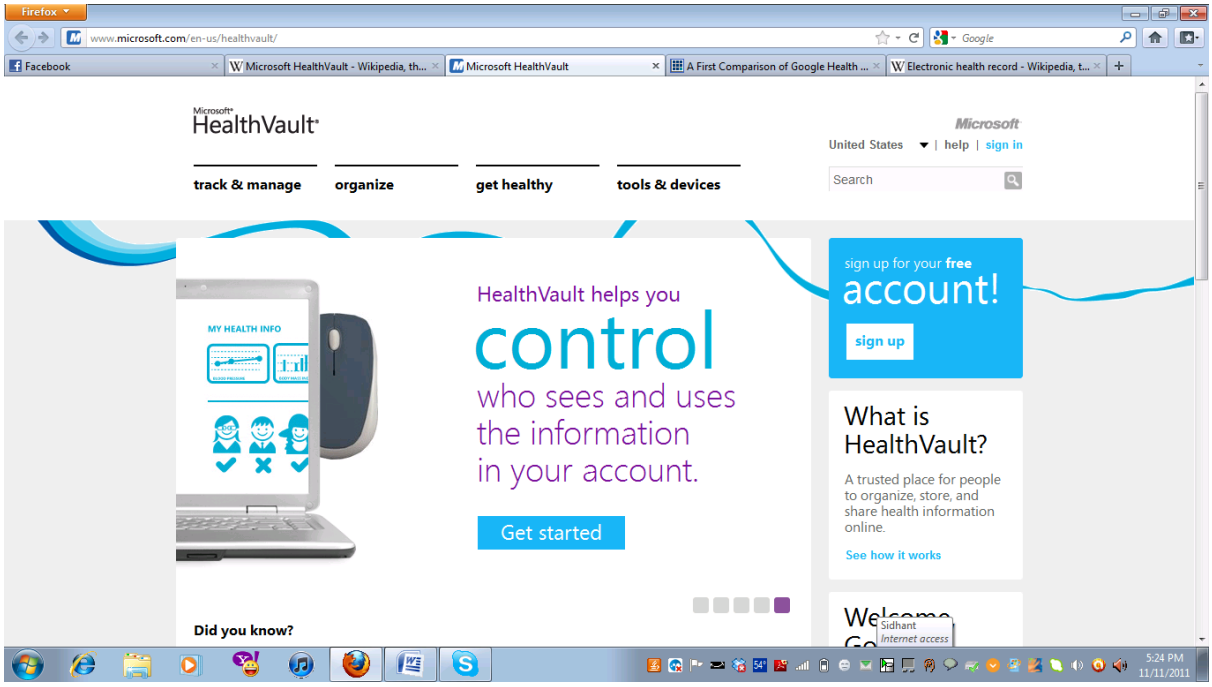


Figure 7: Screen shot of Microsoft Health Vault

An individual can interact with his health vault record through the Health vault web interface or through an application that supports the health vault platform.

Both the services Google Health and Microsoft Health Vault are an attempt to store a user's health records online and set their fitness schedule and manage and monitor them regularly. But other than that they don't provide any feature for the users to get some additional and new information in medical domain based on their health record. So our system can be used as an application on top of these systems to retrieve the

various articles and publications for a particular patient based on his health record. Our system is one step beyond these existing systems.

2.9 PubMed and beyond: a survey of web tools **[23]**

The paper presents a review on 28 different tools and web services which help users to search and retrieve relevant publications for their health related problems. The paper compares different tools to the PubMed and with one another, highlights their innovations and also scope of further improvements. They have also developed a website that is dedicated to online biomedical literature search systems

Literature search is a process in which people use tools to search for relevant literature based on their needs. In our case the domain of literature search is biomedical and the various search criteria can be disease name, symptoms, medication etc. PubMed is the primary tool for searching biomedical publications ever since it was developed as it has a huge database of over 20 million citations. Although PubMed provides an up to date and efficient search interface but still there is a problem of information overload associated with the PubMed search results.

PubMed has two strategies for displaying result for a particular query. The first one is matching the input query terms to MeSH database and displaying the result not only with the original query terms but also with the matched terms from MeSH. The second is its choice of ranking and displaying the ouput in reverse chronological order i.e. the publication date of the various citations.

COMPARISON OF RANKING TECHNIQUES USED BY DIFFERENT TOOLS:

The various tools discussed in the paper use their own different ranking techniques which are discussed below in brief:

RefMed ranks the documents based on machine learning algorithm which first displays the results to the user depending on his query and then takes their feedback and in the second iteration the ranking is based on that feedback. Quertle, another search engine uses concept categorization for ranking of the documents. MedlineRanker first takes as an input a set of articles on a particular topic and learns the various common words in those articles and then scores the newly published articles based on the learning set of words. MiSearch again uses the user's feedback for ranking the documents by tracking the browsing history of the user. Semantic MEDLINE uses semantics involved in the literature by incorporating some biomedical vocabularies. MScanner uses MeSh terms for ranking of the documents. PubFocus uses various factors for ranking scientific publications: journal impact factor, volume of forward references, reference dynamics, and authors' contribution level. There are a few search engines that prevent the information overload by providing the users with the facility of clustering the results based on different categories. Anne O'Tate does the post processing of the search results by grouping them either based on the MeSH terms, important words, author names etc. ClusterMed can cluster the results based on six subcategories like title and abstract, title and MeSH terms, MeSH terms, author names, publication dates etc. MedEvi gives priority to those citations which have the exact terms as in the user query. MEDIE provides semantic search by taking into consideration the semantic relationship

in the input query terms for example the result for the query “what causes breast cancer” would give an output a citation which would answer this question.

The above paragraph describes briefly the various ranking approach used by different search engines for ranking of the medical literature. The ranking approach used by different tools is either based on machine learning algorithms which involves user’s feedback or some words or terms clustering and categorization. A few of the tools use semantics in the query but that too limited to the specific vocabularies from biomedicine. There are some tools which use the impact factor, volume of the references etc as part of their ranking algorithm. All these tools use one or the other ranking techniques which are good enough in their own way but the tools above don’t use the semantic relationships involved in the concepts and terms in the query as well as in the medical literature. Our project proposes a ranking algorithm which ranks the documents by taking into consideration the patient’s profile as a whole and it not only looks for the terms involved in the query but goes a step further to add the semantic relationship from the various annotations obtained from the various publications and patient’s profile and looks for the synonymy, hierarchy and concept or terms categorization obtained from the UMLS and other different ontology’s that NCBO Bioportal Annotator uses. In addition to this the ranking algorithm takes the publication date, the presence of various other factors like the symptoms, medication etc present in the patient profile and also in the medical literature for ranking a literature for a particular patient. Thus, our approach ranks the documents based on the various important information of the EMR’s of the patients and also includes the semantic relationship among the different concepts and terms involved in the literature by annotating the text.

2.10 Similar Electronic Health Records Retrieval

Physicians have to make important decision for their patients especially when they are faced with an untypical case. Then they often use the information from their previous cases. However, the information on health records is really large and this makes the exhaustive search unfeasible. The paper proposes a technique to resolve this issue. The paper proposes a method for retrieving similar Electronic Health Records using UMLS concepts and representing the health records as semantic graphs.

The paper does a semantic match making on the health records by mapping the text onto UMLS concepts and creating a graph. The method achieves relatively high precision and recall, which are also well balanced, which indicates that even though some relevant records are not ranked in the top positions, most retrieved documents are relevant. [22]

CHAPTER 3

SYSTEM WORKFLOW

The workflow of the system can be explained in the following steps:

1. Creating sample health records
2. Parsing the health records and getting the useful information from them.
3. Adding the health records into the ontology.
4. Annotating the health records using NCBO Bioportal Annotator.
5. Adding the annotations obtained from the annotator to the ontology.
6. Downloading medical publications from the PubMed and adding them to the ontology.
7. Annotating the papers and adding the annotations for corresponding papers into the ontology.
8. Running a patient specific query.
9. Carrying out the match making.
10. Ranking the matched results.
11. Displaying the output.

The detailed working of the system is explained below:

1. Creating sample health records:

The following is an example of a health record we created in XML format:

<?xml version="1.0" encoding="UTF-8"?>

<Patient>

<Name>RobinHood</Name>

<Address>1563SouthMiltonst</Address>

<City>Tuscon</City>

<State>AZ</State>

<Zip>92009</Zip>

<Country>UnitedStates</Country>

<Id>1235</Id>

<Age>25</Age>

<KnownDisease>Asthma</KnownDisease>

<Medications>Aerobid,Alvesco</Medications>

<Gender>Male</Gender>

<symptoms>vomiting</symptoms>

<PrimaryPhysician>DrSmith</PrimaryPhysician>

<PhysicianId>dc1247</PhysicianId>

<PrimaryPharmacy>Walgreens</PrimaryPharmacy>

<PrimaryPharmacyId>247Phar</PrimaryPharmacyId>

</Patient>

2, 3.**PARSING THE HEALTH RECORDS:**

After parsing the health records we got the following information and stored them in the ontology.

Patient Details:

Name: Robin Hood

Symptoms: vomiting

Id: 1235

Age: 25

Gender: Male

Known Disease: Asthma

Medications: Aerobid, Alvesco

4, 5. Annotating the patients' data and storing the annotations after parsing the annotation output file into the ontology.

Example of an annotated file:

ObaResultBean [

ResultBean [

resultID = OBA_RESULT_8c82

statistics = [(CLOSURE, 0) , (MAPPING, 0) , (MGREP, 35)]

parameters = [longestOnly = false, wholeWordOnly = true, filterNumber = true, withSynonyms = true, withContext = true, ontology'sToExpand = [], ontology'sToKeepInResult = [], isVirtualOntologyId = false, semanticTypes = [], levelMax = 0, mappingTypes = [null], stopWords = [], withDefaultStopWords = true, isStopWordsCaseSenstive = false, text to annotate = asthma

ontology's = [[SNOMED Clinical Terms, nbAnnotation: 6, score: 78, (46116, 2010_07_31, 1353)], [MedDRA, nbAnnotation: 2, score: 40, (42280, 12.0, 1422)], [ICPC-2 PLUS, nbAnnotation: 2, score: 36, (42297, 2005, 1429)], [eVOC (Expressed Sequence Annotation for Humans), nbAnnotation: 2, score: 20, (44302, 2.9, 1013)], [Logical Observation Identifier Names and Codes, nbAnnotation: 2, score: 20, (44774, 232, 1350)], [NCI Thesaurus, nbAnnotation: 2, score: 18, (45400, 11.01e, 1032)], [Human Phenotype Ontology, nbAnnotation: 1, score: 10, (45774, unknown, 1125)], [Family Health History Ontology, nbAnnotation: 1, score: 10, (38631, 1.0, 1126)], [MedlinePlus Health Topics, nbAnnotation: 1, score: 10, (40397, 20080614, 1347)], [Galen, nbAnnotation: 1, score: 10, (4525, 1.1, 1055)], [International Classification of Primary Care, nbAnnotation: 1, score: 10, (40393, 1993, 1344)], [COSTART, nbAnnotation: 1, score: 10, (40390, 1995, 1341)], [Read Codes, Clinical Terms Version 3 (CTV3) , nbAnnotation: 1, score: 10, (42295, 1999, 1427)], [RadLex, nbAnnotation: 1, score: 10, (45589, 3.4, 1057)], [National Drug File, nbAnnotation: 1, score: 10, (40402, 2008_03_11, 1352)], [WHO Adverse Reaction Terminology, nbAnnotation: 1, score: 10, (40404, 1997, 1354)], [ICD10, nbAnnotation: 1, score: 10, (44103, 1998 , 1516)], [Medical Subject Headings, nbAnnotation: 1, score: 10, (44776, 2011_2010_08_30, 1351)], [Human disease, nbAnnotation: 1, score: 10, (45769, unknown, 1009)], [CRISP Thesaurus, 2006, nbAnnotation: 1, score: 10, (44432, 2006, 1526)], [Online Mendelian Inheritance in Man, nbAnnotation: 1, score: 10, (45553, 2010_04_08, 1348)], [International Classification of Diseases, nbAnnotation: 1, score: 10, (45221, 9, 1101)], [Experimental Factor Ontology, nbAnnotation: 1, score: 10, (45659, 2.12.1, 1136)],

[ICD10CM, nbAnnotation: 1, score: 10, (44860, 2010_03, 1553)], [Bone Dysplasia Ontology, nbAnnotation: 1, score: 10, (46301, 1.0, 1613)]]

annotations = [AnnotationBean [

score = 20

concept = [localConceptId: 46116/155574008, conceptId: 21567348, localOntologyId: 46116, isTopLevel: 1, fullId: http://purl.bioontology.org/ontology/SNOMEDCT/155574008, preferredName: Asthma, definitions: [], synonyms: [Asthma, Asthma (disorder)], semanticTypes: [[id: 25504782, semanticType: T047, description: Disease or Syndrome]]]

context = [MGREP(true), from = 1, to = 6, [name: Asthma, localConceptId: 46116/155574008, isPreferred: false],]

], AnnotationBean [

score = 20

concept = [localConceptId: 46116/155574008, conceptId: 21567348, localOntologyId: 46116, isTopLevel: 1, fullId: http://purl.bioontology.org/ontology/SNOMEDCT/155574008, preferredName: Asthma, definitions: [], synonyms: [Asthma, Asthma (disorder)], semanticTypes: [[id: 25504782, semanticType: T047, description: Disease or Syndrome]]]

context = [MGREP(true), from = 1, to = 6, [name: Asthma, localConceptId: 46116/155574008, isPreferred: true],]

], AnnotationBean [

score = 20

concept = [localConceptId: 42280/10003553, conceptId: 15946621, localOntologyId: 42280, isTopLevel: 0, fullId: http://purl.bioontology.org/ontology/MDR/10003553, preferredName: Asthma, definitions: [], synonyms: [Asthma], semanticTypes: [[id: 19419051, semanticType: T047, description: Disease or Syndrome]]]

context = [MGREP(true), from = 1, to = 6, [name: Asthma, localConceptId: 42280/10003553, isPreferred: false],]

], AnnotationBean [

score = 20

concept = [localConceptId: 42280/10003553, conceptId: 15946621, localOntologyId: 42280, isTopLevel: 0, fullId: http://purl.bioontology.org/ontology/MDR/10003553, preferredName: Asthma, definitions: [], synonyms: [Asthma], semanticTypes: [[id: 19419051, semanticType: T047, description: Disease or Syndrome]]]

context = [MGREP(true), from = 1, to = 6, [name: Asthma, localConceptId: 42280/10003553, isPreferred: true],]

], AnnotationBean [

score = 18

concept = [localConceptId: 42297/R96001, conceptId: 16269522, localOntologyId: 42297, isTopLevel: 1, fullId: http://purl.bioontology.org/ontology/ICPC2P/R96001, preferredName: asthma, definitions: [], synonyms: [Asthma], semanticTypes: [[id: 19761081, semanticType: T047, description: Disease or Syndrome]]]

context = [MGREP(true), from = 1, to = 6, [name: Asthma, localConceptId: 42297/R96001, isPreferred: false],]

], AnnotationBean [

score = 18

concept = [localConceptId: 42297/R96001, conceptId: 16269522, localOntologyId: 42297, isTopLevel: 1, fullId: http://purl.bioontology.org/ontology/ICPC2P/R96001, preferredName: asthma, definitions: [], synonyms: [Asthma], semanticTypes: [[id: 19761081, semanticType: T047, description: Disease or Syndrome]]]

context = [MGREP(true), from = 1, to = 6, [name: asthma, localConceptId: 42297/R96001, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 45400/Asthma, conceptId: 20312930, localOntologyId: 45400, isTopLevel: 0, fullId: http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#Asthma, preferredName: Asthma, definitions: [], synonyms: [], semanticTypes: [[id: 24188889, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 6, [name: Asthma, localConceptId: 45400/Asthma, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 38631/Asthma, conceptId: 13707724, localOntologyId: 38631, isTopLevel: 0, fullId: http://www.owl-ontology's.com/Ontology1172270693.owl#Asthma, preferredName: Asthma, definitions: [], synonyms: [], semanticTypes: [[id: 17088890, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 6, [name: Asthma, localConceptId: 38631/Asthma, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 40393/R96, conceptId: 14167628, localOntologyId: 40393, isTopLevel: 0, fullId: http://purl.bioontology.org/ontology/ICPC/R96, preferredName: Asthma, definitions: [], synonyms: [], semanticTypes: [[id: 17602645, semanticType: T999, description: NCBO BioPortal concept]]]

concept = [localConceptId: 46116/21341004, conceptId: 21630712, localOntologyId: 46116, isTopLevel: 0, fullId: http://purl.bioontology.org/ontology/SNOMEDCT/21341004, preferredName: Asthma, definitions: [], synonyms: [Asthma (disorder) [Ambiguous], Bronchial asthma, NOS, Asthma (disorder), Allergic bronchitis, Allergic bronchitis, NOS, Asthmatic bronchitis, NOS, Asthma, NOS, Bronchial asthma, Asthmatic bronchitis],

semanticTypes: [[id: 25571470, semanticType: T047, description: Disease or Syndrome]]]

context = [MGREP(true), from = 1, to = 6, [name: Asthma, localConceptId: 46116/21341004, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 45659/efo:EFO_0000270, conceptId: 19766146, localOntologyId: 45659, isTopLevel: 0, fullId: http://www.ebi.ac.uk/efo/EFO_0000270, preferredName: asthma, definitions: [], synonyms: [], semanticTypes: [[id: 23624445, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 6, [name: asthma, localConceptId: 45659/efo:EFO_0000270, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 40404/1367, conceptId: 14600511, localOntologyId: 40404, isTopLevel: 0, fullId: http://purl.bioontology.org/ontology/WHO/1367, preferredName: ASTHMA, definitions: [], synonyms: [ASTHMA AGGRAVATED], semanticTypes: [[id: 18072877, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 6, [name: ASTHMA, localConceptId: 40404/1367, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 46116/187687003, conceptId: 21602132, localOntologyId: 46116, isTopLevel: 1, fullId: http://purl.bioontology.org/ontology/SNOMEDCT/187687003, preferredName: Asthma, definitions: [], synonyms: [Asthma (disorder)], semanticTypes: [[id: 25541080, semanticType: T047, description: Disease or Syndrome]]]

context = [MGREP(true), from = 1, to = 6, [name: Asthma, localConceptId: 46116/187687003, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 40397/T2, conceptId: 13714247, localOntologyId: 40397, isTopLevel: 0, fullId: http://purl.bioontology.org/ontology/MEDLINEPLUS/T2, preferredName: Asthma, definitions: [], synonyms: [Bronchial Asthma], semanticTypes: [[id: 17095413, semanticType: T999, description: NCBO BioPortal concept]]]

], AnnotationBean [

score = 10

concept = [localConceptId: 44302/EV:0600009, conceptId: 16957055, localOntologyId: 44302, isTopLevel: 0, fullId: http://purl.org/obo/owl/EV#EV_0600009, preferredName: asthma, definitions: [], synonyms: [], semanticTypes: [[id: 20463686, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 6, [name: asthma, localConceptId: 44302/EV:0600009, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 46116/195967001, conceptId: 21611331, localOntologyId: 46116, isTopLevel: 0, fullId: http://purl.bioontology.org/ontology/SNOMEDCT/195967001, preferredName: Asthma, definitions: [], synonyms: [BHR - Bronchial hyperreactivity, Airway hyperreactivity, Asthmatic, Bronchial asthma, Bronchial hyperresponsiveness, Hyperreactive airway disease, Asthma (disorder), Bronchial hypersensitivity, Bronchial hyperreactivity], semanticTypes: [[id: 25550728, semanticType: T047, description: Disease or Syndrome]]]

context = [MGREP(true), from = 1, to = 6, [name: Asthma, localConceptId: 46116/195967001, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 40402/C1174, conceptId: 14125725, localOntologyId: 40402, isTopLevel: 0, fullId: http://purl.bioontology.org/ontology/NDFRT/C1174, preferredName: Asthma, definitions: [], synonyms: [Asthmas, Bronchial, Asthma [Disease/Finding], Bronchial Asthmas, Bronchial Asthma, Asthmas, Asthma, Bronchial], semanticTypes: [[id: 17559945, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 6, [name: Asthma, localConceptId: 40402/C1174, isPreferred: true],]

6, 7, 8. The papers are downloaded and annotated and then are stored in the ontology with the information as URL, author, publication date, title, abstract, annotation and strength.

9. Running the query.

10. Carrying out the match making process.

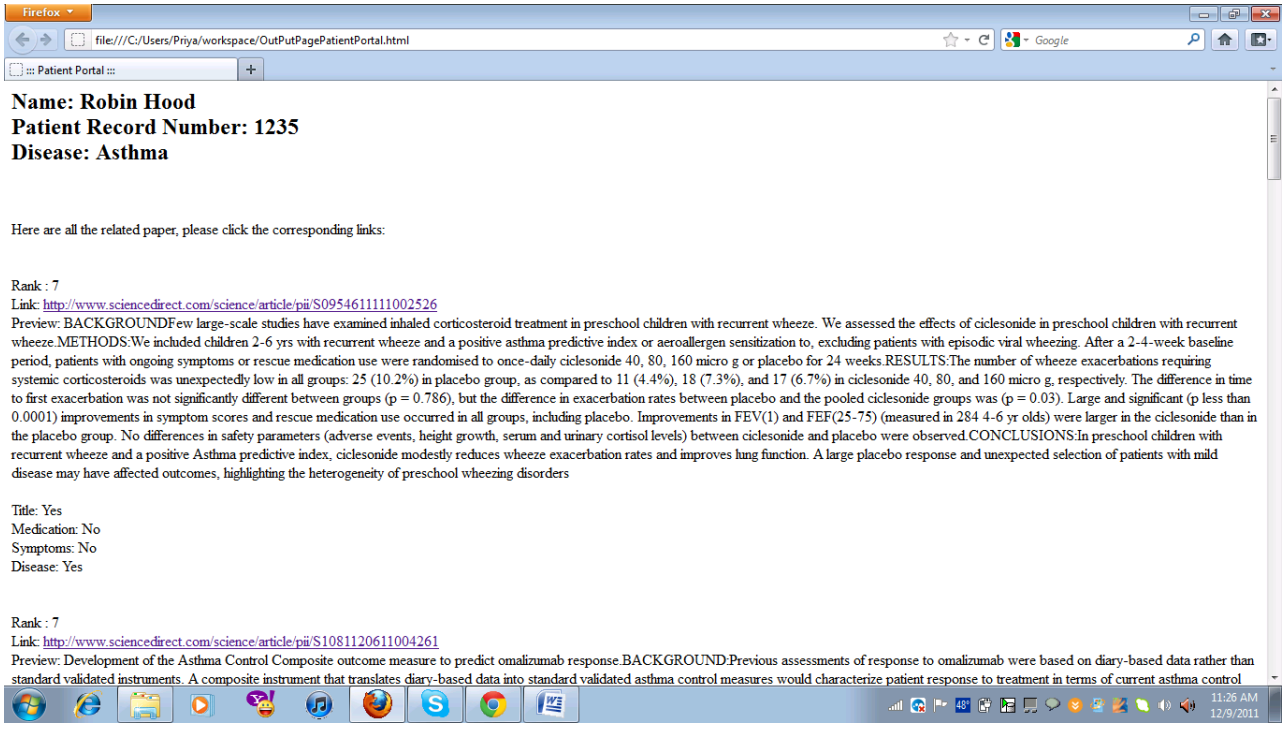
The match making process takes into consideration the following semantics for matching a patient's profile with the corresponding medical publications:

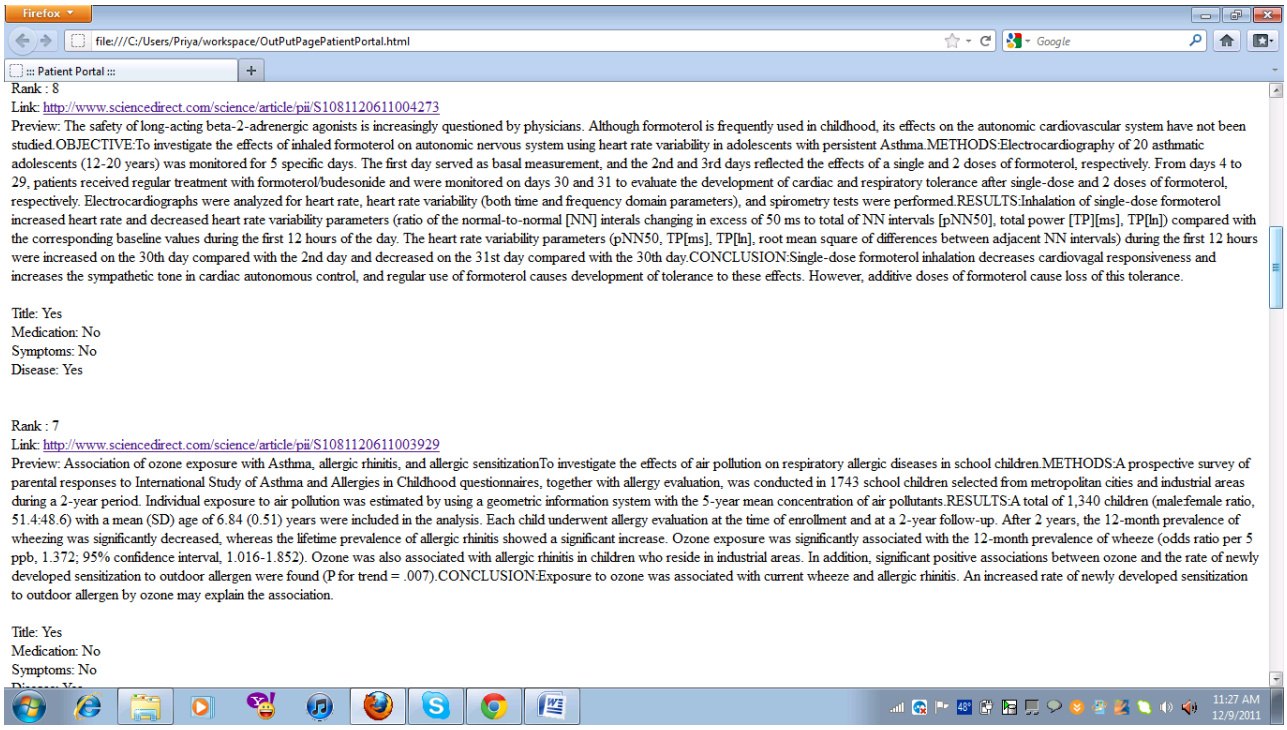
- Disease name or its synonyms: If any of the papers has a disease name in it then it's a match.
- Medication, symptoms or its synonyms: If any of the papers has any one of them or all of them in it then it's a match.

So now if the paper has any one of the above things or all of them in it is considered a match for that particular patient profile.

The match-making process that we carry out in our project is better than simple keyword based match because in the keyword based match the search engine would look for the papers which has only the terms in the query thus a query for *Asthma* would only give the papers that has Asthma as the keyword in it only and would not give the papers on the medications of Asthma, symptoms of Asthma etc. but our system would consider all the data related to a patient profile. Similarly if the paper has no direct name of the disease or any of the medication or symptom but, has the synonym of any of the

terms that are in the patient's profile in that paper, then that particular paper would be displayed as the result of the match. So the results would be more relevant to the query as it will involve the *Semantics* of the medical knowledge and not just the keyword based match. The following is the snap shot of the query run for a patient suffering from Asthma:



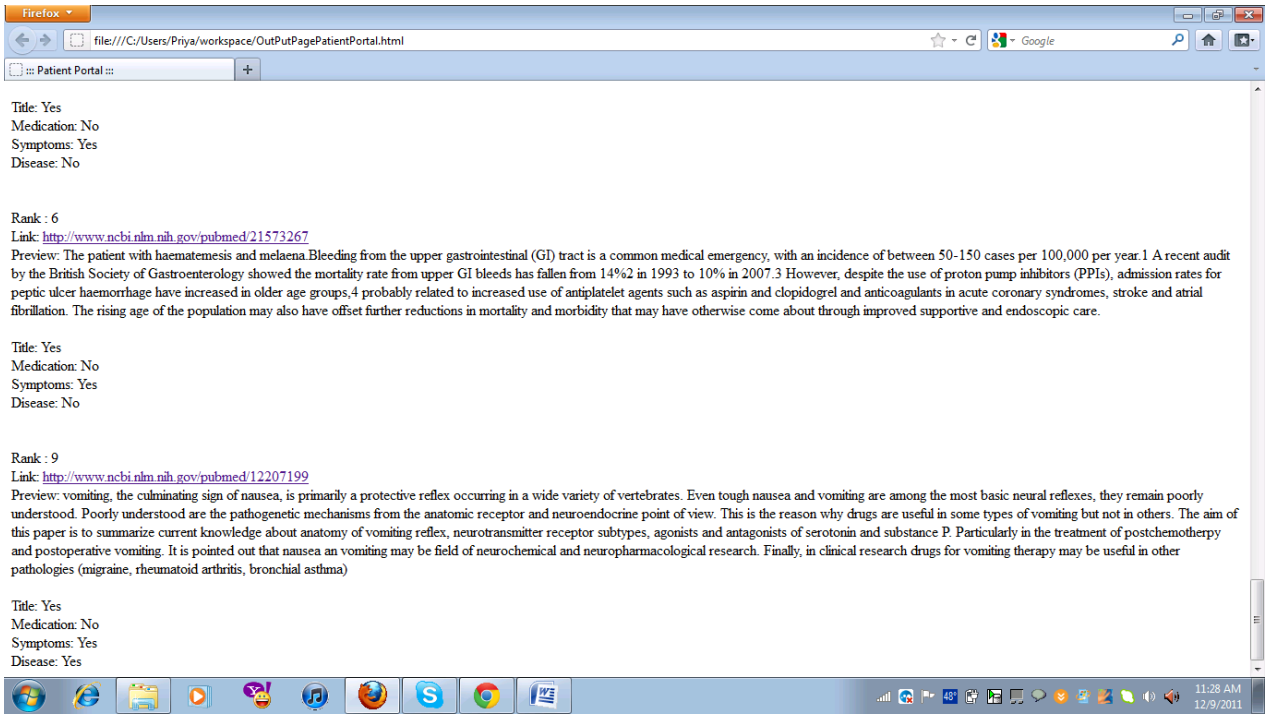


Rank : 8
Link: http://www.sciencedirect.com/science/article/pii/S1081120611004273
Preview: The safety of long-acting beta-2-adrenergic agonists is increasingly questioned by physicians. Although formoterol is frequently used in childhood, its effects on the autonomic cardiovascular system have not been studied.OBJECTIVE:To investigate the effects of inhaled formoterol on autonomic nervous system using heart rate variability in adolescents with persistent Asthma.METHODS:Electrocardiography of 20 asthmatic adolescents (12-20 years) was monitored for 5 specific days. The first day served as basal measurement, and the 2nd and 3rd days reflected the effects of a single and 2 doses of formoterol, respectively. From days 4 to 29, patients received regular treatment with formoterol/budesonide and were monitored on days 30 and 31 to evaluate the development of cardiac and respiratory tolerance after single-dose and 2 doses of formoterol, respectively. Electrocardiographs were analyzed for heart rate, heart rate variability (both time and frequency domain parameters), and spirometry tests were performed.RESULTS:Inhalation of single-dose formoterol increased heart rate and decreased heart rate variability parameters (ratio of the normal-to-normal [NN] interals changing in excess of 50 ms to total of NN intervals [pNN50], total power [TP][ms], TP[ln]) compared with the corresponding baseline values during the first 12 hours of the day. The heart rate variability parameters (pNN50, TP[ms], TP[ln], root mean square of differences between adjacent NN intervals) during the first 12 hours were increased on the 30th day compared with the 2nd day and decreased on the 31st day compared with the 30th day.CONCLUSION:Single-dose formoterol inhalation decreases cardiovagal responsiveness and increases the sympathetic tone in cardiac autonomous control, and regular use of formoterol causes development of tolerance to these effects. However, additive doses of formoterol cause loss of this tolerance.

Title: Yes
Medication: No
Symptoms: No
Disease: Yes

Rank : 7
Link: http://www.sciencedirect.com/science/article/pii/S1081120611003929
Preview: Association of ozone exposure with Asthma, allergic rhinitis, and allergic sensitizationTo investigate the effects of air pollution on respiratory allergic diseases in school children.METHODS:A prospective survey of parental responses to International Study of Asthma and Allergies in Childhood questionnaires, together with allergy evaluation, was conducted in 1743 school children selected from metropolitan cities and industrial areas during a 2-year period. Individual exposure to air pollution was estimated by using a geometric information system with the 5-year mean concentration of air pollutants.RESULTS:A total of 1,340 children (male:female ratio, 51.4:48.6) with a mean (SD) age of 6.84 (0.51) years were included in the analysis. Each child underwent allergy evaluation at the time of enrollment and at a 2-year follow-up. After 2 years, the 12-month prevalence of wheezing was significantly decreased, whereas the lifetime prevalence of allergic rhinitis showed a significant increase. Ozone exposure was significantly associated with the 12-month prevalence of wheeze (odds ratio per 5 ppb, 1.372; 95% confidence interval, 1.016-1.852). Ozone was also associated with allergic rhinitis in children who reside in industrial areas. In addition, significant positive associations between ozone and the rate of newly developed sensitization to outdoor allergen were found (P for trend = .007).CONCLUSION:Exposure to ozone was associated with current wheeze and allergic rhinitis. An increased rate of newly developed sensitization to outdoor allergen by ozone may explain the association.

Title: Yes
Medication: No
Symptoms: No



Title: Yes
Medication: No
Symptoms: Yes
Disease: No

Rank : 6
Link: http://www.ncbi.nlm.nih.gov/pubmed/21573267
Preview: The patient with haematemesis and melaena.Bleeding from the upper gastrointestinal (GI) tract is a common medical emergency, with an incidence of between 50-150 cases per 100,000 per year.1 A recent audit by the British Society of Gastroenterology showed the mortality rate from upper GI bleeds has fallen from 14%2 in 1993 to 10% in 2007.3 However, despite the use of proton pump inhibitors (PPIs), admission rates for peptic ulcer haemorrhage have increased in older age groups,4 probably related to increased use of antiplatelet agents such as aspirin and clopidogrel and anticoagulants in acute coronary syndromes, stroke and atrial fibrillation. The rising age of the population may also have offset further reductions in mortality and morbidity that may have otherwise come about through improved supportive and endoscopic care.

Title: Yes
Medication: No
Symptoms: Yes
Disease: No

Rank : 9
Link: http://www.ncbi.nlm.nih.gov/pubmed/12207199
Preview: vomiting, the culminating sign of nausea, is primarily a protective reflex occurring in a wide variety of vertebrates. Even tough nausea and vomiting are among the most basic neural reflexes, they remain poorly understood. Poorly understood are the pathogenetic mechanisms from the anatomic receptor and neuroendocrine point of view. This is the reason why drugs are useful in some types of vomiting but not in others. The aim of this paper is to summarize current knowledge about anatomy of vomiting reflex, neurotransmitter receptor subtypes, agonists and antagonists of serotonin and substance P. Particularly in the treatment of postchemotherpy and postoperative vomiting. It is pointed out that nausea an vomiting may be field of neurochemical and neuropharmacological research. Finally, in clinical research drugs for vomiting therapy may be useful in other pathologies (migraine, rheumatoid arthritis, bronchial asthma)

Title: Yes
Medication: No
Symptoms: Yes
Disease: Yes

As we can see that the result also has a paper on vomiting which is last paper in the above figure because it is the symptom of Asthma but in PubMed we won't get a paper on vomiting for the query on Asthma.

Thus the Match making that our system does take into consideration all the semantics involved in the patient's profile for retrieval of the medical publications based on the patient's profile.

11, 12. **RANKING THE DOCUMENTS AND DISPLAYING THE RESULTS.**

The documents that are obtained as the result of match making are then ranked based on the algorithm described in detail in the later part.

System Workflow Diagram is shown on next page.

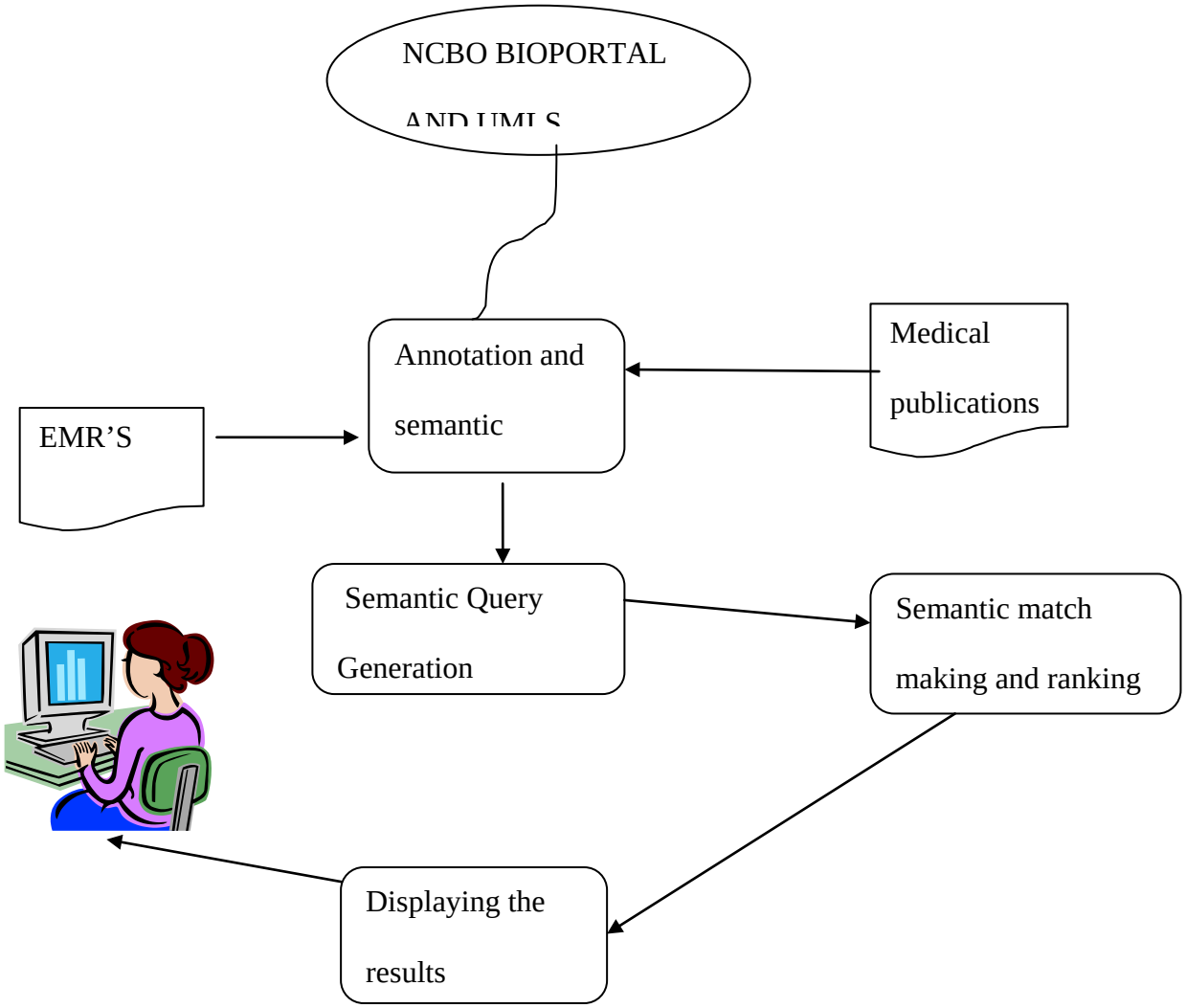


Figure 8: System Workflow

CHAPTER 4

NCBO BOIPORTAL ANNOTATOR

4.1 Semantic Annotation

Annotation of a text is providing additional information about the text by attaching names, concepts, descriptions and comments about the data present in the text to be annotated. The most important requirement for the semantic web is that the content should be well described using the semantics involved in the text. This is done by mapping the content with the ontology concepts. In general, an **annotation** is a note that is made while reading any form of text. [10] This may be as simple as underlining or highlighting passages.

The semantic annotation is different from general annotation as it takes into consideration the various semantics and the semantic relationships involved in the text. It makes the unstructured or semi-structured data rich and semantically meaningful with a context that is further linked to the structured knowledge of a domain. It also allows the results to be displayed after annotation that are not explicitly related to the original search but are semantically related to the text somehow or the other. [11]

Semantic annotations remove the ambiguity that may be present in the text and helps the computer to better understand the concepts and terms present in the text and hence relate them by providing the additional domain specific knowledge. This further makes the search and retrieval of information easy for a computer as the complex relations present in some textual data are processed through the annotations.

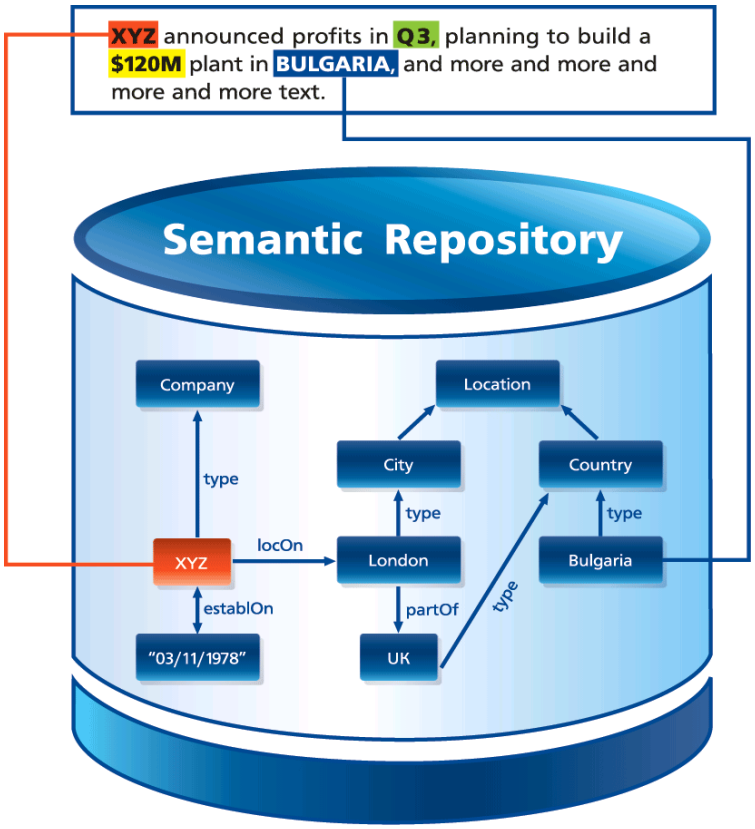


Figure 9: Annotations Example,

Source: http://www.ontotext.com/kim/semantic-annotation

The figure describes the semantic annotation of the text. As we can see how Bulgaria is described in the annotated text by associating the type as country to it which can then tell the computer easily that Bulgaria is a country.

The challenges posed by semantic web as today's web content is still often composed of unstructured text that is not completely re-usable by software agents or semantic engines. There are various ontology's that are present which can be used for annotating

the data. But annotating the data explicitly using the ontology's is still not very popular and is not brought into practice because (i) annotation often needs to be done manually either by expert curators or directly by the authors of the data; (ii) the number of ontology's available for use is large and ontology's change often and frequently overlap; (iii) users do not always know the structure of an ontology's content or how to use the ontology to do the annotation themselves; (iv) annotation can be a boring additional task without immediate reward for the user. Therefore, users need to annotate their data using automatic, easy to use, fast and accurate services that can be integrated into their processes. The annotation of biomedical data has become more difficult as the range of biomedical data is really large and the data is expanding at a faster rate which actually poses a problem for the researchers to efficiently extract the data that they actually need. The NCBO Bioportal annotator helps in solving the problem of annotation of the medical text.

The NCBO annotator was used to get the annotations of the various medical papers in our project. The annotation of the medical papers was actually a process of describing biomedical data that was present in the medical papers with the ontology concepts.

We also used the annotator for getting the annotations of the patient's profile and hence getting the semantic relationships among the various terms involved in the electronic medical health records of the various patients.

The NCBO annotator web service made it easy for us to get the annotations of the biomedical text with the mappings from one of the largest biomedical ontology's, the UMLS as well as the other bioportal ontology's present in the repository of the NCBO,

which further made the annotations concepts to be more rich and meaningful as the mappings were not just restricted to the UMLS.

4.2 NCBO Bioportal Annotator

The National Center for Biomedical Ontology (NCBO) annotator is ontology based web service which can be used for the annotation of biomedical text with the biomedical ontology concepts that are present in the NCBO bioportal repository.[12]

The NCBO annotator uses a set of more than 200 ontology's [13] the most important and the biggest of them all is UMLS (Unified Medical Language System).The annotations of the biomedical text through the ontology concepts and terms makes the unstructured free text data more structured and standardized which help in adding the semantics to the data and hence creating a biomedical semantic web that helps many computer scientists to carry out their projects which involves semantic integration of data.

The NCBO Bioportal offers the integration of various ontology's under one common ontology repository and also provides better functionality by linking the various concepts present in the ontology to the related online data repositories.

The NCBO annotator web service allows scientists to utilize most of the public biomedical ontology's for annotating their datasets automatically.

<u>The NCBO Bioportal Annotator workflow [14]</u>

The workflow of the annotator can be divided into two main steps:

- The free text to be annotated is given as an input to the concept recognition tool which has a dictionary. The dictionary has the list of strings which are actually the concepts that are defined in the ontology. The dictionary is made by accessing all the concepts, their synonyms, the various lowercase and uppercase terms for the concepts etc that all identify the concepts syntactically. The Annotator uses Mgrep2 to recognize concepts by using string matching on the dictionary.
- This primary set of direct annotations serves as input for the *semantic expansion components*, which expanse the annotations extracted from the first step using the knowledge represented in one or more ontology's.

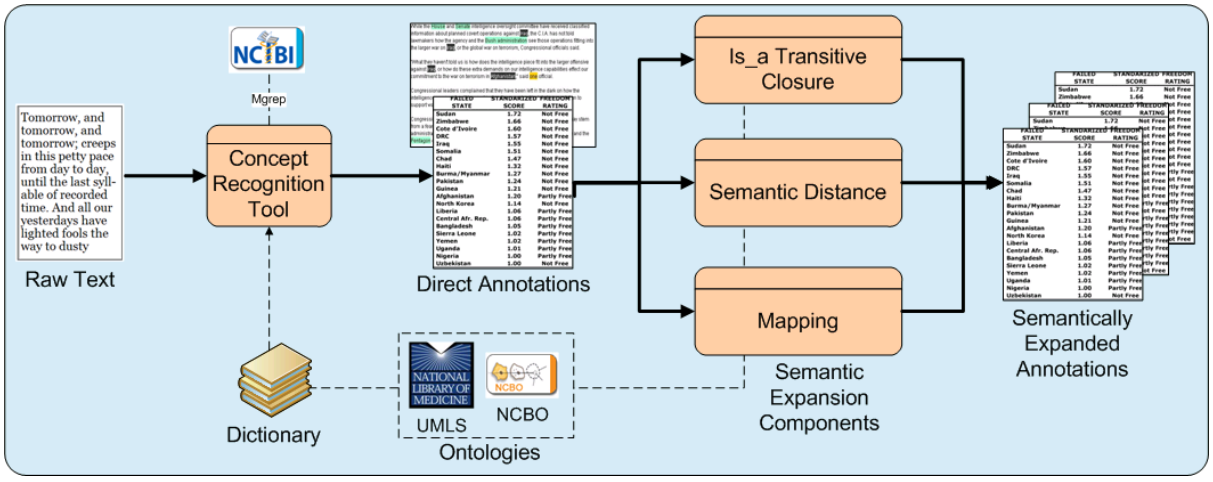


Figure10. NCBO Annotator workflow

Source:http://www.bioontology.org/wiki/index.php/File:OBA_service_workflow.png

The second step can be explained with the help of an example:

- The is a transitive enclosure exploits the parent child relationship of the ontology's, for example, if the text has the word melanoma this is a component generates the further annotated concepts of the melanoma like skin tumor or

neoplasm as NCI Thesaurus provides the knowledge that melanoma *is a* skin tumor and skin tumor *is a* neoplasm.

- The ontology mapping component can generate new annotations based on the mapping among different ontology's. The mapping also has the information on where, i.e. which ontology's the concepts mapping is generated. For example the concept melanoma may generate new annotation with the different concept id from the different ontology. For example the concept NCI/C0025202 (melanoma in NCI Thesaurus) can further generate the annotations from the different ontology's as SNOMEDCT/C0025202 (melanoma in SNOMED-CT) which a different ontology under the NCBO bioportal ontology repository.
- The semantic distance component uses the semantic similarity measures between related concepts and creates new annotation.

The NCBO annotator has 207 biomedical ontology's in total and this ontology's further offer a dictionary of 4,021,662 concepts and 7,637,125 terms. The annotations are scored based on their frequency as well as the context in which they appear in the text.

The following are the advantages of using NCBO annotator and also the reason we prefer to use NCBO annotator instead of any other annotator for the project:

- Large scale: Includes many resources and ontology's under one repository which are integrated and mapped together very well.
- Automatic: It keeps precision and accuracy.

- Easy to use and to access: It has web interface as well as a web service so we need not install it on our personal computer and hence no issues about the memory space problem.
- Customizable: The annotator can fit very specific needs as it provides recommendation for using various ontology's based on the text to be annotated.
- Various output formats available for the annotation such as: text, CSV and XML and hence the user may use the format they choose.

The appendix has the output of the annotation for the following piece of abstract from a paper from PubMed database."Current approaches to the diagnosis and management of asthma are based on guideline recommendations, which have provided a framework for the efforts. Asthma, however, is emerging as a heterogeneous disease, and these features need to be considered in both the diagnosis and management of this disease in individual patients. These diverse or phenotypic features add complexity to the diagnosis of asthma, as well as attempts to achieve control with treatment".

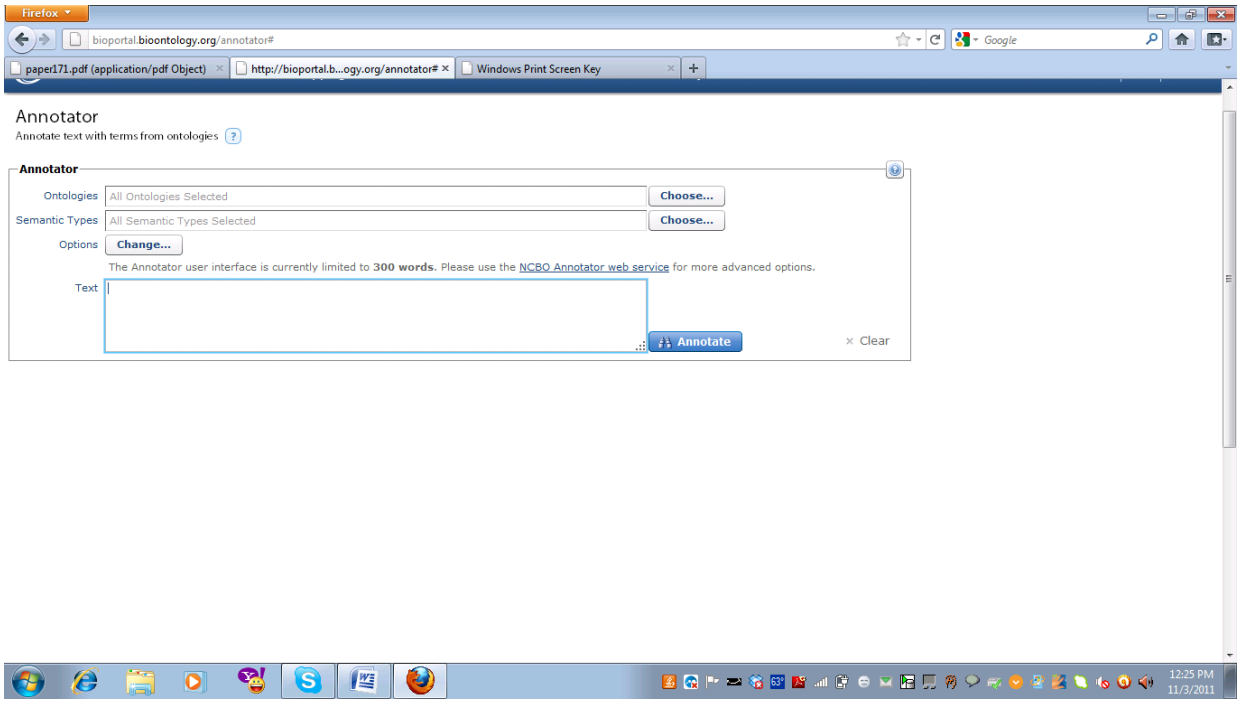


Figure 11: Screen shot showing the NCBO Bioportal Annotator

We can input a maximum of 300 characters for annotating.

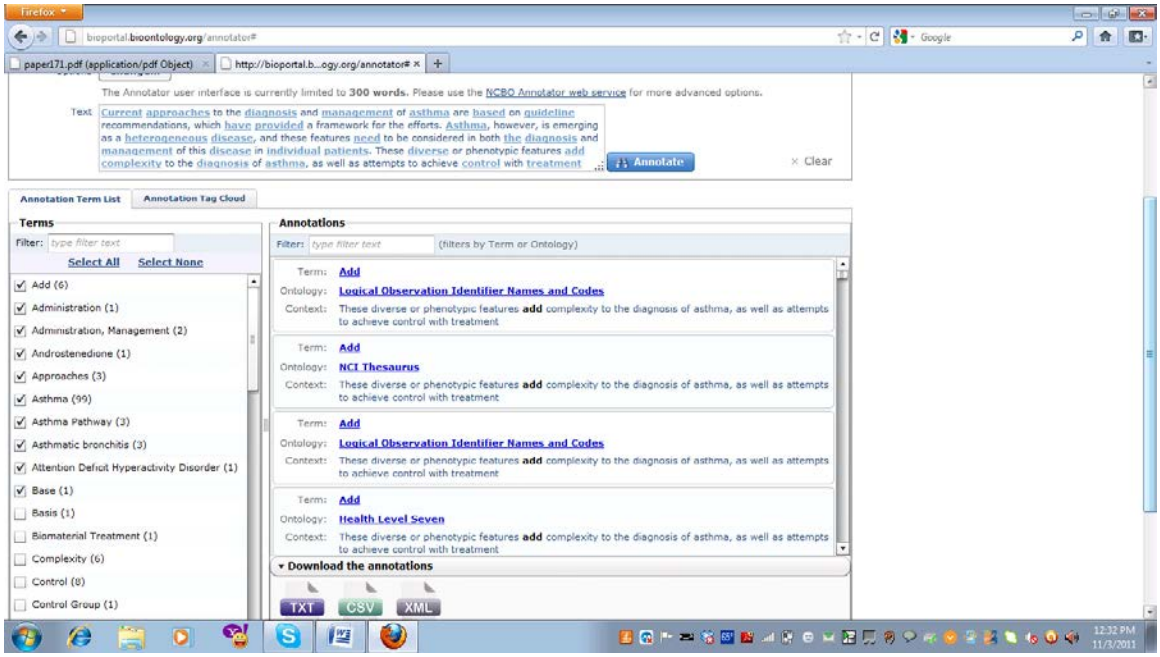


Figure 12: Screen shot of Annotation output

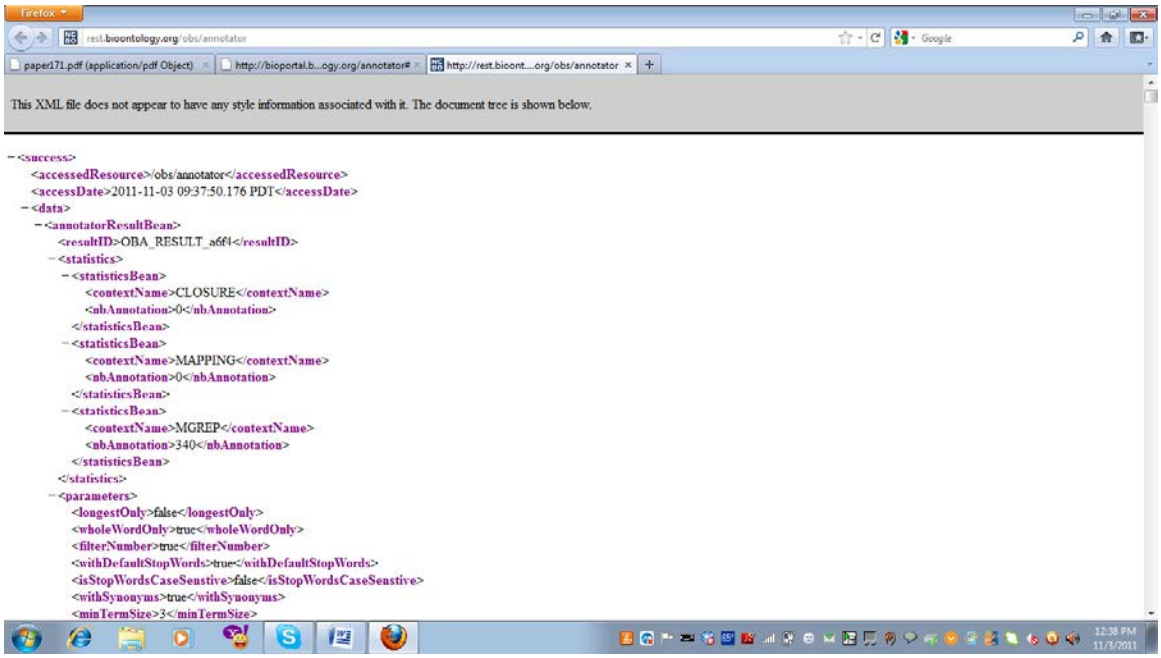


Figure 13: Snap shot showing the XML format of the annotation of the text. Figure 11,12 and 13 show the snapshot of the NCBO Bioportal Annotator web interface where user inputs a text and then clicks the annotate button and then can select from the various output formats available for the annotated file and save the file.

The annotations obtained from NCBO also have a field known as IsTopLevel which can be either true or false. This field tells whether a concept name in the ontology is a top level concept or not. The top level concepts are more important ones as compared to the non top level and we used this feature of the annotations in ranking of the medical papers by calculating the strength of the paper based on the number of top level concepts in that paper. The paper that had the more number of top level concepts is the one that has more number of biomedical terms included in it and has better rank than the others although the ranking doesn't differ much on the basis of strength because the overall ranking is based on several other semantic and syntactic factors.

The detail of how strength is calculated and what exactly a top level concept is explained in semantic ranking algorithm section of the thesis.

The NCBO Bioportal Annotator acted as one of the strong pillars for our project as the annotations obtained were really useful and had the semantic details which further helped us in generating the semantic query results for match-making and also in ranking of the documents by including the semantic relationships involved in the particular medical publication and using those relationships for ranking the documents for a particular patient profile.

CHAPTER 5

AN OVERVIEW AND ANALYSIS OF UMLS

UMLS (Unified Medical Language System) has the classification and coding standards and associated resources to promote creation of more effective and interoperable biomedical information systems and services. The UMLS has a defined ontology which includes terms related to medical sciences including medical health records. The Unified Medical Language System (UMLS) contains semantic information about terms from various sources; each concept can be understood and located by its relationships to other concepts: this is a result of the organizing principle of semantic locality. The various concepts are related by its synonym, hyponym and hypernym. [15] We exploited this feature of UMLS for our project and got really good results. We also used the Concept Hierarchy for the ranking of the medical papers for our project. In this I would explain how these concepts are stored in UMLS and a brief overview of its working and how and which of the relationships of the concepts we used for our project.

The Unified Medical Language System (UMLS) is the research and development project of US National Library of Medicine which was started in 1986 and whose main purpose is to integrate the various biomedical concepts form various distinct databases under one common database and to facilitate the development of a system that understands the language of biomedicine and health sciences. [16] The purpose of UMLS can also be regarded as to overcome a big obstacle that people usually face while dealing with

medical and health sciences information, data, terms and the various concepts. The UMLS overcomes two significant barriers of the effective retrieval of machine readable information:

- The first is the variety of ways the same concepts are expressed in different machine-readable sources and by different people.
- The second is the distribution of useful information among many disparate databases and systems.

There are three main UMLS knowledge sources: [17]

- The Metathesaurus, which contains over one million biomedical concepts from over 100 source vocabularies
- The Semantic Network, which defines 133 broad categories and fifty-four relationships between categories for labeling the biomedical domain
- The SPECIALIST Lexicon & Lexical Tools, which provide lexical information and programs for language processing

The users may use any one of these or all of them based on the type and amount of information they need from the UMLS. The later part of the thesis describes the organization of the concepts in the UMLS with an example and how this organization of information is exploited for our project.

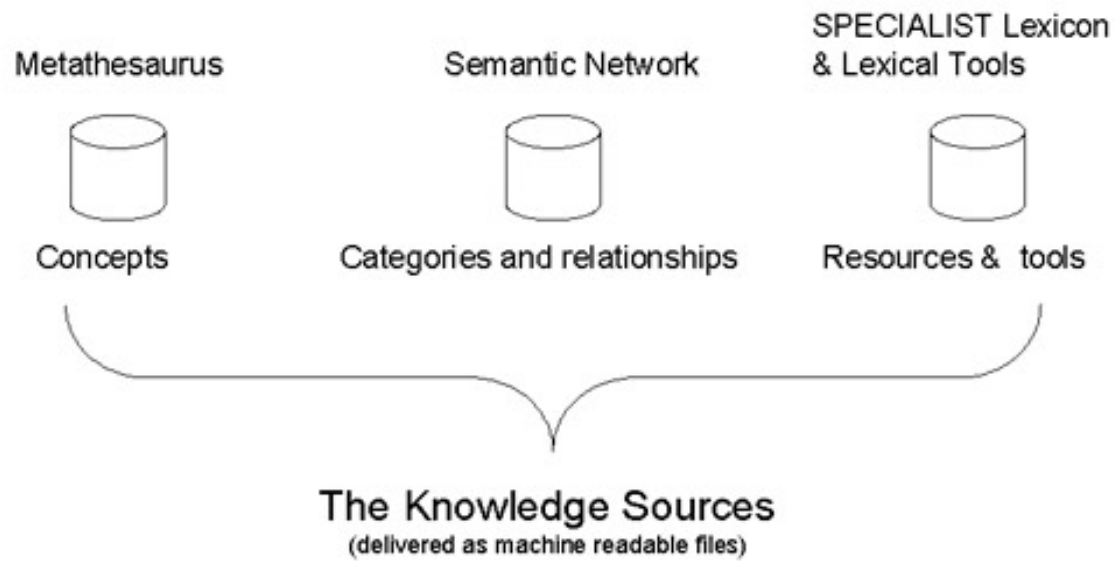


Figure14: The UMLS ORGANISTAION

Source: http://www.nlm.nih.gov/research/umls/new_users/online_learning/index.htm

THE METATHESAURUS

Methathesaurus is data base that has information stored in series of relational database tables and files. It's a large multipurpose database that has information about various biomedical and health sciences related concepts, terms, names and the relationships among them. The Metathesaurus is built from the electronic versions of many different thesauri, classifications, code sets, and lists of controlled terms used in patient care, health services billing, public health statistics, indexing and cataloging biomedical literature, and/or basic, clinical, and health services research. [18] The various terms and names are organized into concepts and assigned a unique identifier.

The Semantic Network

The semantic network describes various semantic types and relationships that can exist among the various terms and concepts that are stored in Metathesaurus. [19] Semantic types describe the various broad categories of biomedicine and health sciences in which the concepts can be categorized for example, the concept Breast cancer can be categorized as a Disease Name. Similarly there are various other semantic types like Clinical Drug, Disease Symptom, Syndrome etc. The relationships among the various terms cane be of the type A clinical drug is used to treat a Disease or a Disease has a Symptom. The semantic network is used to interpret the meanings of the various concepts.

SPECIALIST Lexicon and Lexical Tools

The lexicon consists of a set of lexical entries. Each entry represents a word (lexical item). The entry covers one or more spellings in a particular part of speech and describes the morphologic, orthographic and syntactic properties of a word. These all entries are from the Biomedical domain. The various sources for the lexical coding of these words are: the MEDLINE abstracts and *Dorland's Illustrated Medical Dictionary*. The Dorland's Illustrated Medical Dictionary is the dictionary that includes the meanings of all the medical terms that are in current usage. The lexical tools are nothing but a collection of Java programs that help in natural language processing of these words and terms. [20]

In our project, we basically exploited the Metathesaurs database of the UMLS.The concepts included in the database and the relationships among them were obtained through the annotation of the medical text. We used the NCBO bioportal annotator

which uses various medical ontology's to annotate the text and data provided to it for annotation and UMLS is one of biggest and major of all those ontology's. A lot many different vocabularies are included in Metathesaurs which are categorized into many different ways. The major categories of these vocabularies are: [21]

- Diagnosis
    - Logical Observation Identifier Names and Codes (LOINC)
- Procedures & Supplies
    - Current Procedural Terminology (CPT)
- Diseases
    - International Classification of Diseases and Related Health Problems (ICD-10)
- Comprehensive Vocabularies/Thesauri
    - Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT)

Other categories are anatomy, drugs, genetics, nursing and miscellaneous. The graph below shows the percentage of different vocabularies that are included in Metathesaurs.

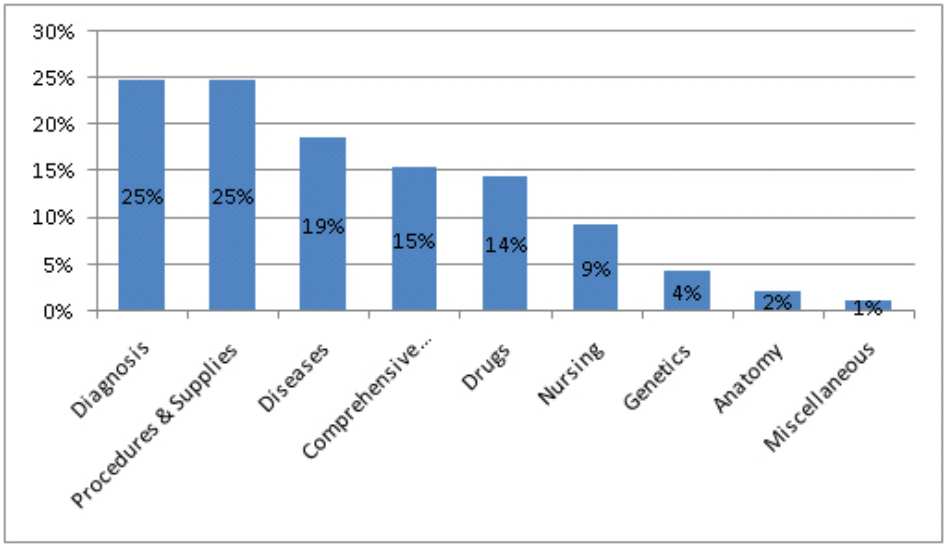


Figure15: Percentage of different vocabularies in Metathesaurus

Source:http://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_001.htm

As the Metathesaurus contains a lot many concepts in it users may select the subset of these concepts based on the application for which they are using these vocabularies and filter out the rest of the content stored in the Metathesaurs. The various subsets that are included or that can be chosen are based on:

- The language of the vocabulary
- The semantic type associated with the concepts and terms
- The terms related to a specific area of the biomedical and health sciences

There are two relational formats for Metathesaurus subsets that are available for selection by default while installing the UMLS Metathesaurus: [22]

- Rich Release Format (RRF)
- Original Release Format (ORF)

The format that is recommended by UMLS is RRF as it involves the detailed semantics of each of the vocabulary and provides well organized information about the concepts ,their relationships ,the various hierarchical categories included in the vocabularies and much more.

There is a name called “preferred term” for all the concepts that might have several different names in several vocabularies which is used to refer to a particular term in Metathesaurs. This preferred term naming can be better explained through this example:

The various terms that could identify the concept Hodgkin's disease are: Hodgkins disease, Hodgkin's disease, Hodgkin’s sarcoma, Hodgkin lymphoma and many more but the concept is specified using the preferred term which is Hodgkin Disease in the Metathesaurs vocabulary and all the other terms are related to it by one of the various semantic relation types which will be described later.

Each concept in Metathesaurs is given a unique identifier. There are four levels of unique identifiers:

- Concept unique identifiers are the identifiers that are attached with each concept that may have several names or synonyms but only one if the various names of the concept which is also the preferred term are given a concept unique identifier and the rest of the names of the concept are related to it using the various relationships. For example Hodgkin disease will be assigned a concept unique identifier.

- Lexical unique identifiers are the identifiers associated with each of the lexical variant of the concept for example Hodgkin lymphoma and Hodgkin's sarcoma each of them will be given different lexical unique identifiers.
- String unique identifier is given to any variation whether it is based on the upper case, lower case; punctuation difference for the same word is given a string unique identifier. For example hodgkin's disease and Hodgkin's disease and Hodgkin disease each will have their own string unique identifier.

A concept can occur in more than one vocabulary from which the Metathesaurs is built. So each occurrence of the same concept in different vocabularies there is an Atom Unique Identifier.

The concept unique identifiers link the concept data across the various files of the Metathesaurs.

The symbolic relationships that are present in the UMLS are:

- Hierarchical
    - Parent/child
    - Broader/Narrower
- Derived from hierarchies
    - Siblings
- Synonymy

- Similar
- Source asserted synonymy
- Possible synonymy

The examples of hierarchical relationships are:

Breast Cancer is a Disease, Animal is an Organism.

The examples of non hierarchical relationships are:

Chemotherapy treats Breast Cancer

Symptom diagnoses a Disease

Each Methathesaurs concept is assigned a Semantic type independent of its position in the hierarchy and Semantic relationship is a link that can exist between two concepts.

The diagram below explains the entire semantic network of the UMLS which include the semantic type of entities, the semantic type events and the relationships. The examples semantic type of entities are: Gene, Protein, Carbohydrate, Drug and the examples of semantic type events are: Social behavior, mental process etc.

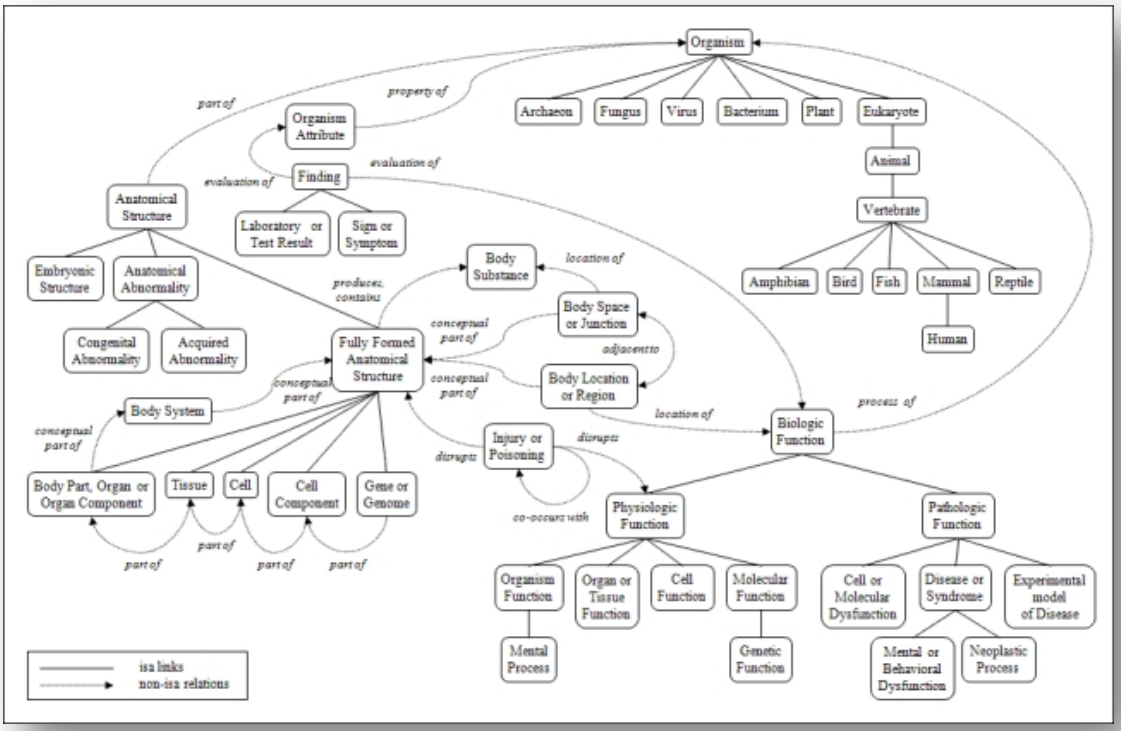


Figure 16: The semantic network of Metathesaurus

Source: http://www.ncbi.nlm.nih.gov/books/NBK9679/figure/ch05.F3/

The UMLS provides the health professionals and researchers to use the biomedical information from different sources. The terms present in different vocabularies that are same in meaning but might have different names are clustered into one unique concept and given a concept unique identifier while maintaining the original structure of each source vocabulary.

The advantages or features of UMLS that make it more powerful than other medical ontology's are:

- Integrates several source vocabularies under one common repository.

- Integrates over 2 million names for some 900 000 concepts from more than 60 families of biomedical vocabularies.
- Have more than 12 million relations among these concepts.
- NCBI taxonomy, Gene Ontology, the Medical Subject Headings (MeSH), OMIM, SNOMed and the Digital Anatomist Symbolic Knowledge Base are the vocabularies that are integrated.
- UMLS concepts are not only inter-related, but may also be linked to external resources such as GenBank.
- Metathesaurus is customizable based on the specific user needs and requirements.

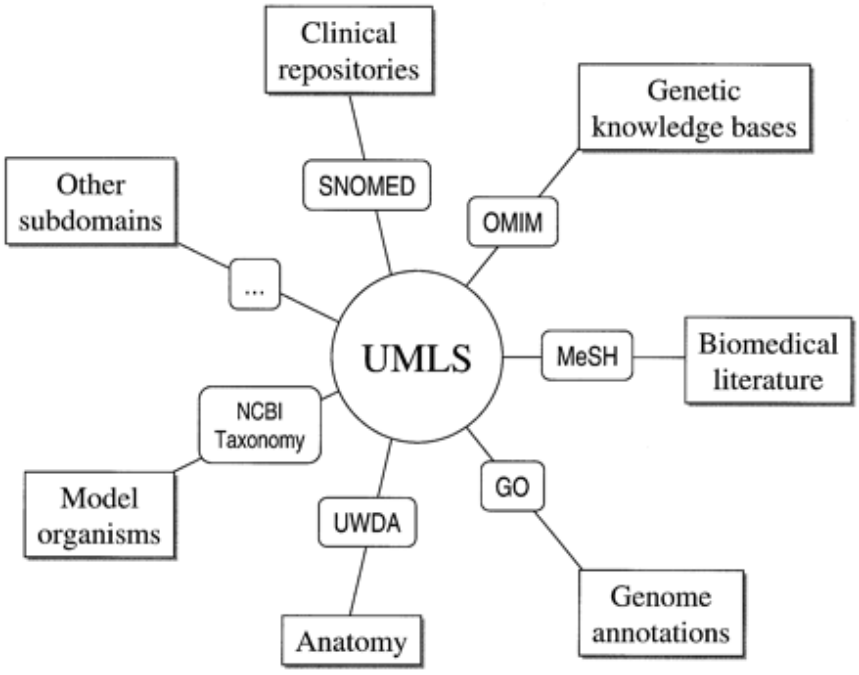


Figure17: The Sub domain integration in UMLS

Source:http://nar.oxfordjournals.org/content/32/suppl_1/D267/F1.expansion.html

We used the UMLS data to get the annotation for our patient's information. The various concepts and their semantic relationships were very useful for us to get the annotations of medical information and data for the various diseases, their medication, the symptoms and the synonyms of the various medical terms associated with the patient's profile. UMLS database was of great help for us to get the information related to a particular disease. We also used the UMLS along with the other ontology's for annotating the medical publications that were downloaded from PubMed and the annotation of the medical text was also used in the match making of the papers with a particular patient profile. The annotations that we got from the papers using UMLS had all the medical terms that were present in a particular paper along with the various synonyms of those terms which were really useful for us in efficient match making and then ranking the various papers for the particular patient's profile.

The annotations helped us in getting the various concepts present in the paper and thus made the process of match making really efficient as the patient's profiles were queried against the various publications that were present in our ontology and the match making of the papers was done based on the semantics involved in the patient profile and the medical publication and the semantics for the medical publications were basically obtained from the annotations of the medical text which used the concepts of the UMLS and also some of the other NCBO ontology's. As the annotations had all the terms whether in different case, punctuations, the synonyms of the concepts and much more information that made the task of match making and also the ranking of the matched publications much more relevant for a particular patient.

The hierarchical relationships were exploited for calculating the strength of the various medical papers. The output that we get for the annotation of the various medical publications has a Top level concept value which is either true or false so we calculated the strength of the medical papers based on the numbers of top level concepts present in the particular medical publication and then finally used that for the ranking of the medical papers. The more the top level concepts present the better the strength of the paper. The example of top level concepts and the formula for calculating the strength are explained in detail in the semantic ranking algorithm section of the thesis.

We used the NCBO bioportal annotator and UMLS ontology in the backend to get the results for our queries. The workflow shows how UMLS is used for getting the annotations of the medical information and the text. The annotations are created from the concept recognition based on the terms (concept names and synonyms) that are stored in the UMLS and other NCBO ontology's. The NCBO annotator along with the UMLS was really helpful for the match making as well as the ranking process.

Here as we can easily see from the figure that UMLS is the major and the important ontology that is used by the NCBO annotator along with some of the NCBO's own ontology for annotation the text .So UMLS played an important role for getting the annotations of the medical publications as well as the various synonyms of the terms that were included in the medical patient's profile and thus making the process of semantic match making and ranking of the documents relevant and efficient. Though we didn't only use UMLS for annotations and synonymy relationships, the annotator also used several other ontology's for getting the annotations which further made the

process of match making and ranking have more scope in it as the terms used are not restricted to just one ontology but various ontology's combined together ,the biggest of them all is UMLS.

The various relationships present in UMLS among the concepts made the match making really efficient and powerful as there are different medical papers which don't have the exact name of the disease that is there in the patient's health records or the name of the medication he is prescribed or the symptoms that he has but the annotations obtained from the UMLS had everything related to a particular term its various synonyms, if there is some medication then the salts of that medication etc. and if any of the papers had the name related to the names in the annotations obtained is also displayed as a result and is ranked based on the ranking algorithm thus the system we proposed gives better results as compared to key word based search engines and hence will make it easy for the users to get the most relevant medical publications related to their particular medical health condition.

CHAPTER 6

SEMANTIC RANKING OF MEDICAL PUBLICATION

The medical publications from the PubMed were downloaded and stored in the ontology with the following information which was set as the properties of each individual paper: the URL which uniquely identified each paper, the title of the paper, the author of the paper, the publication year, the abstract of the paper. The abstract and the title of each paper were given to the NCBO annotator as the input and the annotations were obtained for each paper and then stored in the ontology. The annotations that were obtained from the NCBO annotator also had a field named IsTopLevel which has Boolean value either 0 or 1.If the value is 1 that means a particular concept or term is top level in the particular ontology and hence that term is of significance in medical field as compared to the other terms included in the paper as the NCBO uses the medical data ontology's for annotating the text provided to it as an input. The ontology for the medical publications also has the field strength which is calculated based on the score of each paper. The formula for calculating the strength is explained in the later part of this section. So each paper in the ontology has the above mentioned properties which are used for the match making and ranking of the medical publication based on the patients' profile.

The patient profile is basically an Electronic Health Record of the patient which was parsed and the important information that was needed for the match making process and the ranking were stored into the ontology. The patients' data was given to the NCBO annotator and the annotations were obtained for each individual patient and from those annotations the synonyms for the various terms and concepts were stored into the ontology. The patient profile ontology had properties as: patient's id which uniquely identified the patient, patient's name, known disease, symptoms, medication and then the information from the annotations output which was the medication synonym, disease synonym, the symptom synonym which added the semantics to the process of match making and ranking of the medical information and knowledge. The query for a particular patient's profile is carried out and the results of the match making and ranking are obtained based on the semantics involved in the patient's medical data and the various medical publications. The match making process is described in brief in the system's workflow section and here we shall concentrate on the Semantic Ranking Algorithm in detail which is deployed in the project.

In general, for simple ranking of any documents for information retrieval based on the user query the things that are taken into consideration are the terms in the query and then those terms are matched with the relevant document, the frequency of the query terms in the relevant document, term proximity i.e. words that are close together in the query are close somewhere in the relevant document, the location of the query terms in the document, prefer documents that are more popular the idea behind page rank, prefer documents with short URL's and those which have query terms in the URL. These are the general criteria that are taken into consideration for the information

retrieval and hence ranking of the documents. These all factors are only based on the keyword match and then ranking the documents but doesn't involve the semantics of the document. Thus ranking that is done on basis of these factors no doubt will give some good results as compared to no ranking at all but still a lot of filtering of the information would have to be done manually by the user as the results would contain at least some of the documents which do not answer the user query properly but are displayed as the output as they might contain some of the words that were present in the user query. for example if a user types a query "some good Indian restaurants in Athens GA" he will get the result with all the pages that have any of the terms in above query with only 2-3 results that are precisely relevant to the above query and hence the user will have to filter the information from among the set of the output information but what if the search engine understands the semantics of the above query and displays only the good Indian restaurants in Athens GA as the result. This is when semantics play an important role in information retrieval. Similarly when a user types Symptoms of Asthma in PubMed search engine the result that the user obtains is ranked on the basis of publication date firstly and then is based only on the keyword match and not on the overall semantics. A snapshot of the above query is shown below:

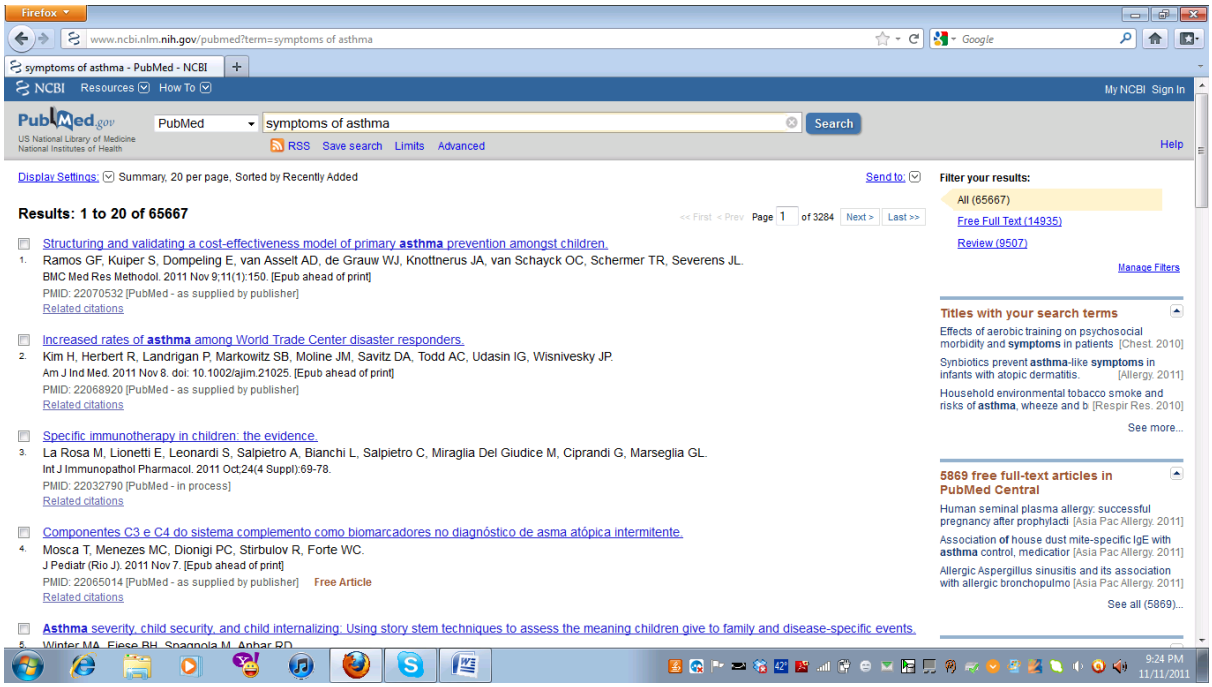


Figure 18: Snapshot of PubMed Query

As can be seen from the snap shot the result of the query does not involve the symptoms of asthma rather only displays the papers that have asthma in it but the user actually needs the papers on the symptoms of asthma.

So a system is needed that takes the semantics involved in the query and displays the results to the user not just based on the keyword matching.

Our algorithm takes into consideration some of the above factors in addition to the other factors that add the semantics to the information retrieval process and hence reduce the workload at the end user side of filtering the documents retrieved as a result of his/her query.

The following are the factors our algorithm considers that add up to the overall ranking of the medical publications:

- Frequency and occurrence of a term: The system searches for any of the terms present in the patient's profile in all of the papers in the database then that paper which is with the terms is retrieved as a match and is given a rank.
- Term Proximity/Location of the term: We check for the term present in the patient's profile is in the title and abstract of the paper or is only in the abstract. If the term is only present in the abstract but not in the title then that means that the paper is not specifically related to that particular concept or term and hence is given a lesser rank and if the term is present in both the title and abstract that paper is given a higher rank than the above ones and since there cannot be a publication that has a term in the title only and not in the abstract so we do not take this case into consideration. for example if a query is run for the patient who has a disease lung cancer and there is a paper in the database that has lung cancer or any of its medication or symptom or any of its synonyms in the title of the paper along with the abstract then that paper is given more rank than a paper which has any of the terms in only the abstract as that abstract might only have that term in relation to some other disease or condition. The test cases in the later section will describe the ranking in a more clear way.
- Timeline: The recent is the publication the newer is the discovery and hence it is given a better rank as compared to the old ones. We have given higher rank to the publications from last three years and a lower rank to the older publications.
- Semantics Involved: (Presence of symptoms, disease, medication)If a paper has symptoms, disease name as well as medication mentioned in it for a particular patient profile that paper is given a higher rank as compared to a paper

which has any two of them which is given a less higher rank or just one of them which has even lesser rank because a paper that talks about the symptoms, medications and the disease is the most relevant and most suited to the patient's profile and hence should be given a higher rank than the paper that only talks about any one or any two of them.

- The algorithm not only searches for the disease name, medication or symptom for ranking the papers but goes one level deeper in the hierarchy where if a paper doesn't have the direct match for the disease name, medication or symptom but has the synonym of the above mentioned terms and hence provides the paper a rank based on what all are the relationships a paper has with the patient's profile.
- Strength of the paper: Every paper is given a strength which is based on the score of the paper. The score of the paper is the total number of top level concepts present in the paper which is calculated from the annotation result and the strength of the paper is calculated by using the following formula:

  Strength=Score/Total number of concepts

  The value of strength is between 0 and 1 and in the ontology we have given the strength the value either 0.5 for every strength value between 0 and 0.5 and a value of 1 for any strength value between 0.5 and 1.

  The example of Top level concept is explained below in the annotation output with the annotations of two terms i.e. disease and a drug named Aerobid. Since disease is a common term and hence not very specific therefore it is not a top level concept in most of the ontology's whereas the term Aerobid is top level in all

the ontology's used by NCBO Bioportal Annotator. Hence a paper that has more number of top level concepts is given a better strength as compared to the paper that has lesser number of top level concepts.

The annotation output of Aerobid :

```
ObaResultBean [
ResultBean [
        resultID = OBA_RESULT_6eb6
        statistics = [(MAPPING, 0) , (MGREP, 2) , (CLOSURE, 0) ]
        parameters = [longestOnly = false, wholeWordOnly = true, filterNumber = true, withSynonyms = true, withContext = true, ontology'sToExpand = [], ontology'sToKeepInResult = [], isVirtualOntologyId = false, semanticTypes = [], levelMax = 0, mappingTypes = [null], stopWords = [], withDefaultStopWords = true, isStopWordsCaseSenstive = false, text to annotate = aerobid]
]
        ontology's = [[RxNORM, nbAnnotation: 1, score: 10, (44775, 10AA_100907F, 1423)], [Medical Subject Headings, nbAnnotation: 1, score: 8, (44776, 2011_2010_08_30, 1351)]]
        annotations = [AnnotationBean [
                score = 10
                concept = [localConceptId: 44775/215045, conceptId: 18847696, localOntologyId: 44775, isTopLevel: 1, fullId: http://purl.bioontology.org/ontology/RXNORM/215045, preferredName: AeroBid, definitions: [], synonyms: [], semanticTypes: [[id: 22357811, semanticType: T110, description: Steroid], [id: 22357812, semanticType: T121, description: Pharmacologic Substance]]]
                context = [MGREP(true), from = 1, to = 7, [name: AeroBid, localConceptId: 44775/215045, isPreferred: true], ]
```

], AnnotationBean [

score = 8

concept = [localConceptId: 44776/C007734, conceptId: 19464827, localOntologyId: 44776, isTopLevel: 1, fullId: http://purl.bioontology.org/ontology/MSH/C007734, preferredName: flunisolide, definitions: [], synonyms: [Inhacort, Ratiopharm Brand of Flunisolide, flunisolide hemihydrate, (6alpha,11beta,16alpha)-isomer, Roche Brand of Flunisolide, Apo-Flunisolide, Elan Brand 1 of Flunisolide, Syntaris, flunisolide, (6beta,11beta,16alpha)-isomer, flunisolide hydrofluoroalkane, Ivax Brand of Flunisolide, ratio-Flunisolide, RS-3999, Elan Brand 2 of Flunisolide, 6 alpha-fluorodihydroxy-16 alpha,17 alpha-isopropylidenedioxy-1,4-pregnadiene-3,20- dione, Rhinalar, Nasarel, Dermapharm Brand of Flunisolide, Boehringer Ingelheim Brand of Flunisolide, 6 alpha-fluoro-11 beta,16 alpha,17,21- tetrahydroxypregna-1,4-diene-3,20-dione cyclic 16, 17-acetal with acetone, Forest Brand of Flunisolide, AeroBid, Apotex Brand of Flunisolide, Nasalide, flunisolide HFA], semanticTypes: [[id: 23197067, semanticType: T110, description: Steroid], [id: 23197068, semanticType: T121, description: Pharmacologic Substance]]]

context = [MGREP(true), from = 1, to = 7, [name: AeroBid, localConceptId: 44776/C007734, isPreferred: false],]

]]

]

The annotation output of Disease:

ObaResultBean [

ResultBean [

resultID = OBA_RESULT_0c7f

statistics = [(MAPPING, 0) , (MGREP, 29) , (CLOSURE, 0)]

parameters = [longestOnly = false, wholeWordOnly = true, filterNumber = true, withSynonyms = true, withContext = true, ontology'sToExpand = [], ontology'sToKeepInResult = [], isVirtualOntologyId = false, semanticTypes = [], levelMax = 0, mappingTypes = [null], stopWords = [], withDefaultStopWords = true, isStopWordsCaseSenstive = false, text to annotate = DISEASE]

]

ontology's = [[ICPC-2 PLUS, nbAnnotation: 2, score: 36, (42297, 2005, 1429)], [PKO_Re, nbAnnotation: 1, score: 10, (40917, 1.1, 1409)], [SemanticScience Integrated Ontology, nbAnnotation: 1, score: 10, (45775, 0.8.12, 1532)], [Ontology for General Medical Science, nbAnnotation: 1, score: 10, (45302, 2011-02-21, 1414)], [Gene Regulation Ontology, nbAnnotation: 1, score: 10, (44629, 0.5, 1082)], [Event (INOH pathway ontology), nbAnnotation: 1, score: 10, (45404, unknown, 1011)], [Gene Regulation Ontology, nbAnnotation: 1, score: 10, (45127, 0.5, 1106)], [Brucellosis Ontology, nbAnnotation: 1, score: 10, (44723, 1.0.67, 1537)], [Host Pathogen Interactions Ontology, nbAnnotation: 1, score: 10, (45230, 1.0, 1569)], [NMR-instrument specific component of metabolomics investigations, nbAnnotation: 1, score: 10, (44836, unknown, 1033)], [Pilot Ontology, nbAnnotation: 1, score: 10, (40653, 0.1, 1399)], [Ontology for Biomedical Investigations, nbAnnotation: 1, score: 10, (45713, 2011-04-20, 1123)], [Vaccine Ontology, nbAnnotation: 1, score: 10, (45715, Vision Release; 1.0.457, 1172)], [National Drug File, nbAnnotation: 1, score: 10, (40402, 2008_03_11, 1352)], [Protein-protein interaction, nbAnnotation: 1, score: 10, (39508, 1.52, 1040)], [NIFSTD, nbAnnotation: 1, score: 10, (45355, 2.2 - December 20, 2010, 1084)], [SNOMED Clinical Terms, nbAnnotation: 1, score: 10, (46116, 2010_07_31, 1353)], [BIRNLex, nbAnnotation: 1, score: 10, (29684, 1.3.1, 1089)], [EDAM, nbAnnotation: 1, score: 10, (45158, beta11, 1498)], [Medical Subject Headings, nbAnnotation: 1, score: 10, (44776, 2011_2010_08_30, 1351)], [ExO, nbAnnotation: 1, score: 10, (45220, 1, 1575)],

[Human disease, nbAnnotation: 1, score: 10, (45769, unknown, 1009)], [Logical Observation Identifier Names and Codes, nbAnnotation: 1, score: 10, (44774, 232, 1350)], [Infectious disease, nbAnnotation: 1, score: 10, (46205, unknown, 1092)], [Experimental Factor Ontology, nbAnnotation: 1, score: 10, (45659, 2.12.1, 1136)], [Translational Medicine Ontology, nbAnnotation: 1, score: 10, (45369, 1.0, 1461)], [NCI Thesaurus, nbAnnotation: 1, score: 8, (45400, 11.01e, 1032)], [PHARE, nbAnnotation: 1, score: 8, (45138, 110114, 1550)]]

annotations = [AnnotationBean [

score = 18

concept = [localConceptId: 42297/A99001, conceptId: 16265023, localOntologyId: 42297, isTopLevel: 1, fullId: http://purl.bioontology.org/ontology/ICPC2P/A99001, preferredName: disease, definitions: [], synonyms: [Disease], semanticTypes: [[id: 19756536, semanticType: T047, description: Disease or Syndrome]]]

context = [MGREP(true), from = 1, to = 7, [name: disease, localConceptId: 42297/A99001, isPreferred: true],]

], AnnotationBean [

score = 18

concept = [localConceptId: 42297/A99001, conceptId: 16265023, localOntologyId: 42297, isTopLevel: 1, fullId: http://purl.bioontology.org/ontology/ICPC2P/A99001, preferredName: disease, definitions: [], synonyms: [Disease], semanticTypes: [[id: 19756536, semanticType: T047, description: Disease or Syndrome]]]

context = [MGREP(true), from = 1, to = 7, [name: Disease, localConceptId: 42297/A99001, isPreferred: false],]

], AnnotationBean [

score = 10

concept = [localConceptId: 45302/obo:OGMS_0000031, conceptId: 19925950, localOntologyId: 45302, isTopLevel: 0, fullId: http://purl.obolibrary.org/obo/OGMS_0000031, preferredName: disease, definitions: [], synonyms: [], semanticTypes: [[id: 23784528, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: disease, localConceptId: 45302/obo:OGMS_0000031, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 44836/obi:OBI_155, conceptId: 13385087, localOntologyId: 44836, isTopLevel: 0, fullId: http://obi.sourceforge.net/ontology/OBI.owl#OBI_155, preferredName: disease, definitions: [], synonyms: [], semanticTypes: [[id: 16637175, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: disease, localConceptId: 44836/obi:OBI_155, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 40402/C2140, conceptId: 14129412, localOntologyId: 40402, isTopLevel: 0, fullId: http://purl.bioontology.org/ontology/NDFRT/C2140, preferredName: Disease, definitions: [], synonyms: [Diseases, Disease [Disease/Finding]], semanticTypes: [[id: 17563632, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: Disease, localConceptId: 40402/C2140, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 44723/obo:OGMS_0000031, conceptId: 16632794, localOntologyId: 44723, isTopLevel: 0, fullId: http://purl.obolibrary.org/obo/OGMS_0000031, preferredName: disease, definitions: [], synonyms: [], semanticTypes: [[id: 20137670, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: disease, localConceptId: 44723/obo:OGMS_0000031, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 45769/DOID:4, conceptId: 20244866, localOntologyId: 45769, isTopLevel: 1, fullId: http://purl.org/obo/owl/DOID#DOID_4, preferredName: disease, definitions: [], synonyms: [], semanticTypes: [[id: 24120825, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: disease, localConceptId: 45769/DOID:4, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 39508/MI:0617, conceptId: 17515323, localOntologyId: 39508, isTopLevel: 0, fullId: http://purl.org/obo/owl/MI#MI_0617, preferredName: disease, definitions: [], synonyms: [], semanticTypes: [[id: 21021954, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: disease, localConceptId: 39508/MI:0617, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 45775/resource:SIO_010299, conceptId: 19935475, localOntologyId: 45775, isTopLevel: 0, fullId:

http://semanticscience.org/resource/SIO_010299, preferredName: disease, definitions: [], synonyms: [], semanticTypes: [[id: 23794053, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: disease, localConceptId: 45775/resource:SIO_010299, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 40917/PKO_Revamp:Disease, conceptId: 15936007, localOntologyId: 40917, isTopLevel: 0, fullId: http://www.semanticweb.org/ontology's/2009/10/25/PKO_Revamp.owl#Disease, preferredName: Disease, definitions: [], synonyms: [], semanticTypes: [[id: 19408373, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: Disease, localConceptId: 40917/PKO_Revamp:Disease, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 45404/IEV:0000075, conceptId: 19831513, localOntologyId: 45404, isTopLevel: 0, fullId: http://purl.bioontology.org/ontology/IEV/IEV_0000075, preferredName: Disease, definitions: [], synonyms: [], semanticTypes: [[id: 23689812, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: Disease, localConceptId: 45404/IEV:0000075, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 40653/Disease, conceptId: 15927404, localOntologyId: 40653, isTopLevel: 0, fullId: http://www.owl-ontology's.com/2009/9/24/Ontology1253802770.owl#Disease, preferredName:

Disease, definitions: [], synonyms: [], semanticTypes: [[id: 19399770, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: Disease, localConceptId: 40653/Disease, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 44629/GRO:Disease, conceptId: 14064378, localOntologyId: 44629, isTopLevel: 0, fullId: http://www.bootstrep.eu/ontology/GRO#Disease, preferredName: disease, definitions: [], synonyms: [], semanticTypes: [[id: 17495544, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: disease, localConceptId: 44629/GRO:Disease, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 45659/efo:EFO_0000408, conceptId: 19766282, localOntologyId: 45659, isTopLevel: 0, fullId: http://www.ebi.ac.uk/efo/EFO_0000408, preferredName: disease, definitions: [], synonyms: [], semanticTypes: [[id: 23624581, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: disease, localConceptId: 45659/efo:EFO_0000408, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 45220/ID:0000079, conceptId: 19938278, localOntologyId: 45220, isTopLevel: 0, fullId: http://purl.bioontology.org/ontology/ExO/ID_0000079, preferredName: Disease, definitions: [], synonyms: [], semanticTypes: [[id: 23796856, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: Disease, localConceptId: 45220/ID:0000079, isPreferred: true],]
], AnnotationBean [
score = 10
concept = [localConceptId: 44776/D004194, conceptId: 19661052, localOntologyId: 44776, isTopLevel: 1, fullId: http://purl.bioontology.org/ontology/MSH/D004194, preferredName: Disease, definitions: [], synonyms: [Diseases, DIS], semanticTypes: [[id: 23510423, semanticType: T047, description: Disease or Syndrome]]]
context = [MGREP(true), from = 1, to = 7, [name: Disease, localConceptId: 44776/D004194, isPreferred: true],]
], AnnotationBean [
score = 10
concept = [localConceptId: 29684/birnlex_11013, conceptId: 20949125, localOntologyId: 29684, isTopLevel: 0, fullId: http://bioontology.org/projects/ontology's/birnlex#birnlex_11013, preferredName: Disease, definitions: [], synonyms: [], semanticTypes: [[id: 24836203, semanticType: T999, description: NCBO BioPortal concept]]]
context = [MGREP(true), from = 1, to = 7, [name: Disease, localConceptId: 29684/birnlex_11013, isPreferred: true],]
], AnnotationBean [
score = 10
concept = [localConceptId: 45355/p9:birnlex_11013, conceptId: 20248734, localOntologyId: 45355, isTopLevel: 0, fullId: http://ontology.neuinfo.org/NIF/Backend/BIRNLex-OBI-proxy.owl#birnlex_11013, preferredName: Disease, definitions: [], synonyms: [], semanticTypes: [[id: 24124693, semanticType: T999, description: NCBO BioPortal concept]]]
context = [MGREP(true), from = 1, to = 7, [name: Disease, localConceptId: 45355/p9:birnlex_11013, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 46116/64572001, conceptId: 21867281, localOntologyId: 46116, isTopLevel: 0, fullId: http://purl.bioontology.org/ontology/SNOMEDCT/64572001, preferredName: Disease, definitions: [], synonyms: [Disorders, Clinical disease AND/OR syndrome present, Diseases, Syndrome, Disorder, Clinical disease or syndrome present, NOS, Disorder, NOS, Clinical disease or syndrome, NOS, Syndrome, NOS, Clinical disease AND/OR syndrome, Disease or syndrome present, NOS, Disease (disorder), Disease, NOS, Disease AND/OR syndrome present], semanticTypes: [[id: 25829027, semanticType: T047, description: Disease or Syndrome]]]

context = [MGREP(true), from = 1, to = 7, [name: Disease, localConceptId: 46116/64572001, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 46205/obo:OGMS_0000031, conceptId: 21903703, localOntologyId: 46205, isTopLevel: 0, fullId: http://purl.obolibrary.org/obo/OGMS_0000031, preferredName: disease, definitions: [], synonyms: [], semanticTypes: [[id: 25870640, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: disease, localConceptId: 46205/obo:OGMS_0000031, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 45713/obi:OBI_1110055, conceptId: 19753424, localOntologyId: 45713, isTopLevel: 0, fullId: http://purl.obolibrary.org/obo/OBI_1110055, preferredName: disease,

definitions: [], synonyms: [], semanticTypes: [[id: 23611723, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: disease, localConceptId: 45713/obi:OBI_1110055, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 45158/EDAM:0000634, conceptId: 19927934, localOntologyId: 45158, isTopLevel: 0, fullId: http://purl.bioontology.org/ontology/EDAM/EDAM_0000634, preferredName: Disease, definitions: [], synonyms: [], semanticTypes: [[id: 23786512, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: Disease, localConceptId: 45158/EDAM:0000634, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 45715/DOID:DOID_4, conceptId: 19957511, localOntologyId: 45715, isTopLevel: 0, fullId: http://purl.org/obo/owl/DOID#DOID_4, preferredName: disease, definitions: [], synonyms: [], semanticTypes: [[id: 23816089, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: disease, localConceptId: 45715/DOID:DOID_4, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 45230/http://www.semanticweb.org/ontology’s/2010/5/22/Ontology1277229984000.owl#HPI:0000026, conceptId: 19936500, localOntologyId: 45230, isTopLevel: 0, fullId: http://www.semanticweb.org/ontology’s/2010/5/22/Ontology1277229984000.owl#HPI

:0000026, preferredName: disease, definitions: [], synonyms: [], semanticTypes: [[id: 23795078, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: disease, localConceptId: 45230/http://www.semanticweb.org/ontology’s/2010/5/22/Ontology1277229984000.owl#HPI:0000026, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 45127/GRO:Disease, conceptId: 15488032, localOntologyId: 45127, isTopLevel: 0, fullId: http://www.bootstrep.eu/ontology/GRO#Disease, preferredName: disease, definitions: [], synonyms: [], semanticTypes: [[id: 18960398, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: disease, localConceptId: 45127/GRO:Disease, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 45369/transmed:TMO_0047, conceptId: 19926643, localOntologyId: 45369, isTopLevel: 0, fullId: http://www.w3.org/2001/sw/hcls/ns/transmed/TMO_0047, preferredName: disease, definitions: [], synonyms: [], semanticTypes: [[id: 23785221, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: disease, localConceptId: 45369/transmed:TMO_0047, isPreferred: true],]

], AnnotationBean [

score = 10

concept = [localConceptId: 44774/LP21006-9, conceptId: 13978091, localOntologyId: 44774, isTopLevel: 1, fullId:

http://purl.bioontology.org/ontology/LNC/LP21006-9, preferredName: Disease, definitions: [], synonyms: [], semanticTypes: [[id: 17387138, semanticType: T047, description: Disease or Syndrome]]]

context = [MGREP(true), from = 1, to = 7, [name: Disease, localConceptId: 44774/LP21006-9, isPreferred: true],]

], AnnotationBean [

score = 8

concept = [localConceptId: 45400/Diseases_and_Disorders, conceptId: 20334066, localOntologyId: 45400, isTopLevel: 0, fullId: http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#Diseases_and_Disorders, preferredName: Disease or Disorder, definitions: [], synonyms: [condition, Disorders, Disorder, Diseases, Diseases and Disorders, Disease], semanticTypes: [[id: 24210025, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: Disease, localConceptId: 45400/Diseases_and_Disorders, isPreferred: false],]

], AnnotationBean [

score = 8

concept = [localConceptId: 45138/phare:Disease, conceptId: 19935796, localOntologyId: 45138, isTopLevel: 0, fullId: http://www.stanford.edu/~coulet/phare.owl#Disease, preferredName: phare:Disease, definitions: [], synonyms: [formation, type, presence, patophysiology, injury, episode, observation, admission, maintenance, experience, pathology, event, diagnosis, model, sequela, onset, diseases, incidence, disease, lesion, syndrome, occurence, appearance, disorder, form, period, impairment, pathogenesis], semanticTypes: [[id: 23794374, semanticType: T999, description: NCBO BioPortal concept]]]

context = [MGREP(true), from = 1, to = 7, [name: disease, localConceptId: 45138/phare:Disease, isPreferred: false],]

]]

]

Hence as it is clear from the example how top level concept is used for calculating the strength of the paper lets now see how the ranking algorithm works.

The rank for the papers is calculated as follows:

There are two scenarios one is for strength as 0.5 and the other is for strength as 1.

The strength 0.5 adds +1 to the overall rank of the paper whereas the strength 1 adds +2 to the overall rank of the paper.

If the paper has any of the terms present in patient's profile in the title then this gives a score 2 to the overall rank.

If the paper has any of the terms in the patient profile only in the abstract of the paper this adds the score based on which of the terms are there in the abstract to the overall rank.

If the paper has medication in it this gives a score 2 to the overall rank. But suppose the paper does not have the exact name of the medication in it but has the synonym of the medication then also that paper is given a score 2 as it is still talking about the medication that a patient is prescribed. For example a patient suffering from Asthma is prescribed a medicine Alvesco but there is a paper which talks about Flunisolide then that paper is given a score 2 for the patient profile with Asthma as Flunisolide is obtained from the annotations of the term of Alvesco which says that it's a synonym of Alvesco.

If the paper has symptom of a particular patient in it this gives a score 2 to the overall rank. Again the synonymy relationship is considered into account while providing the score to the paper so the paper is searched for the symptom as well as its synonym to give it a score.

If the paper has known disease in it this gives a score 3 to the overall rank. Here again the synonymy relationship is exploited.

If the publication date is between 2009 and 2011 then this adds to +1 to the overall rank otherwise 0.5 is added.

Based on these values the overall rank of the paper is calculated and assigned to it by adding the values and hence the minimum rank for a paper is 2 and the maximum is 12.

The various rank values assigned to a paper can be better explained with the following table where we have considered the possible combinations of occurrence of various terms in any of the papers related to the patient's profile. The ranking shown in this table is for strength =1 and publications between year 2009 and 2011 so the strength and publication date will give a score 3 to the overall ranking of the paper. Similarly there can be other cases too for the papers with strength 0.5 and publication date something else other than the past 3 years. This table is only an example to explain the ranking in a better way and to give an idea how ranking would work in different cases.

Table 2: Semantic Rank Example

Title	yes	no	no	yes	Yes	yes	no	no	no	yes	No	yes	no	Yes	no
Disease	yes	no	yes	no	Yes	yes	yes	no	no	yes	No	no	yes	No	yes
Medication	yes	no	yes	yes	No	yes	no	yes	no	no	Yes	yes	no	No	yes
Symptoms	yes	no	yes	yes	Yes	no	no	no	yes	no	Yes	no	yes	Yes	no

Rank	12	3	10	9	10	10	6	5	5	8	7	7	8	7	8

The first column describes a case where a paper has the any of the terms(whether disease, medication or symptom or their synonyms in the title) as well as all of these terms in the paper which can be the best case for any scenario then the rank would be as 12 which is calculated as:

Rank= score of title+ score of disease+ score of medication+ score of symptom+ strength score+ publication date score

So Rank=2+3+2+2+2+1=12

In the same the rank for other cases is calculated.

The minimum rank which is 3 is never assigned to a paper because that paper won't be a match for any profile as nothing is matched against the patient's profile.

CHAPTER 7

TEST CASES AND RESULTS

The query for the patient id 1235 was run and following results were obtained:

Name:RobinHood

Patient Record:

Number:1235

Disease: Asthma

Rank:8

Preview: Current approaches to the diagnosis and management of asthma are based on guideline recommendations, which have provided a framework for the efforts. Asthma, however, is emerging as a heterogeneous disease, and these features need to be considered in both the diagnosis and management of this disease in individual patients. These diverse or phenotypic features add complexity to the diagnosis of asthma, as well as attempts to achieve control with treatment. Although the diagnosis of asthma is often based on clinical information, it is important to pursue objective criteria as well, including an evaluation for reversibility of airflow obstruction and bronchial hyperresponsiveness, an area with new diagnostic approaches. Furthermore, there exist a number of treatment gaps (ie, exacerbations, step-down care, use of antibiotics, and severe disease) in which new direction is needed to improve care.

Title: Yes Medication: No Symptoms: No Disease: Yes Strength: 1 Publication year: 2011

The above paper has disease name Asthma in it which also appears in the title of the paper, has no medication or its synonyms and no symptoms or its synonyms, the strength of the paper is 1 and publication year is 2011 so according to the formula for rank the rank is calculated as:

Rank= Score of title+ Score of medication+ Score of symptom+ Score of Disease+ Score of strength+ Score of Publication Year

Hence, Rank =2+0+0+3+2+1=8

Rank:7

Link:http://chestjournal.chestpubs.org/content/139/2/311.long

Preview: BACKGROUND:Patients with mild persistent Asthma constitute about 70% of the asthma population; thus, it is important to know which first-line treatment is best for the management of mild asthma. We compared benefits of first-line treatment with ciclesonide and a combination of fluticasone and salmeterol in patients with mild asthma.METHODS:Patients aged 12 to 75 years with mild persistent asthma were enrolled in a randomized, double-blind, placebo-controlled study. After run-in, patients were randomized to ciclesonide 160 micro g once daily (CIC160), fluticasone propionate/salmeterol 100/50 micro g bid (FP200/S100), or placebo for 52 weeks. The primary variable was time to first severe asthma exacerbation; the coprimary variable was the percentage of poorly controlled asthma days. Patients recorded asthma symptoms and salbutamol use in electronic diaries and completed a standardized version of the Asthma Quality of Life Questionnaire.RESULTS:Compared with placebo,

the time to first severe asthma exacerbation was prolonged, and lung function was improved with FP200/S100 treatment (P = .0002) but not with CIC160. Both CIC160 and FP200/S100 provided significantly fewer poorly controlled asthma days than placebo (P less than .0016 for both active treatments). Moreover, both active treatments provided significantly more asthma symptom-free days (P less than.0001), rescue medication-free days (P = .0005, one-sided), and days with asthma control (P less than .0033). Overall Asthma Quality of Life Questionnaire scores were significantly higher in both active treatment groups than placebo (P less than .0017).

Title: Yes Medication: No Symptoms: No Disease: Yes Strength: 0.5 Publication Date: 2011

The above paper has disease name in the title as well abstract, no medication, no symptom, the strength is 0.5 and publication year as 2011 so the rank is given as 7 which is calculated by the formula explained in the previous chapter.

Hence Rank = 2+0+0+3+1+1=7

Rank:7

Link:http://www.sciencedirect.com/science/article/pii/S0954611111002526

Preview: BACKGROUNDFew large-scale studies have examined inhaled corticosteroid treatment in preschool children with recurrent wheeze. We assessed the effects of ciclesonide in preschool children with recurrent wheeze.METHODS:We included children 2-6 yrs with recurrent wheeze and a positive asthma predictive index or aeroallergen sensitization to, excluding patients with episodic viral wheezing. After a 2-4-week baseline period, patients with ongoing symptoms or rescue medication use were randomised to once-daily ciclesonide 40, 80, 160 micro g or placebo for 24

weeks.RESULTS:The number of wheeze exacerbations requiring systemic corticosteroids was unexpectedly low in all groups: 25 (10.2%) in placebo group, as compared to 11 (4.4%), 18 (7.3%), and 17 (6.7%) in ciclesonide 40, 80, and 160 micro g, respectively. The difference in time to first exacerbation was not significantly different between groups (p = 0.786), but the difference in exacerbation rates between placebo and the pooled ciclesonide groups was (p = 0.03). Large and significant (p less than 0.0001) improvements in symptom scores and rescue medication use occurred in all groups, including placebo. Improvements in FEV(1) and FEF(25-75) (measured in 284 4-6 yr olds) were larger in the ciclesonide than in the placebo group. No differences in safety parameters (adverse events, height growth, serum and urinary cortisol levels) between ciclesonide and placebo were observed.CONCLUSIONS:In preschool children with recurrent wheeze and a positive Asthma predictive index, ciclesonide modestly reduces wheeze exacerbation rates and improves lung function

Title: Yes Medication: No Symptoms: No Disease: Yes Strength:0.5 Publication Date: 2011

Hence:Rank=2+0+0+3+1+1=7

Rank:8

Link:http://www.sciencedirect.com/science/article/pii/S1081120611004273

Preview: The safety of long-acting beta-2-adrenergic agonists is increasingly questioned by physicians. Although formoterol is frequently used in childhood, its effects on the autonomic cardiovascular system have not been studied.OBJECTIVE:To investigate the effects of inhaled formoterol on autonomic nervous system using heart rate variability in

adolescents with persistent Asthma.METHODS:Electrocardiography of 20 asthmatic adolescents (12-20 years) was monitored for 5 specific days. The first day served as basal measurement, and the 2nd and 3rd days reflected the effects of a single and 2 doses of formoterol, respectively. From days 4 to 29, patients received regular treatment with formoterol/budesonide and were monitored on days 30 and 31 to evaluate the development of cardiac and respiratory tolerance after single-dose and 2 doses of formoterol, respectively. Electrocardiographs were analyzed for heart rate, heart rate variability (both time and frequency domain parameters), and spirometry tests were performed.RESULTS:Inhalation of single-dose formoterol increased heart rate and decreased heart rate variability parameters (ratio of the normal-to-normal [NN] interals changing in excess of 50 ms to total of NN intervals [pNN50], total power [TP][ms], TP[ln]) compared with the corresponding baseline values during the first 12 hours of the day. The heart rate variability parameters (pNN50, TP[ms], TP[ln], root mean square of differences between adjacent NN intervals) during the first 12 hours were increased on the 30th day compared with the 2nd day and decreased on the 31st day compared Title: Yes Medication: No Symptoms: No Disease: Yes Strength:1 Publication year:2010

Hence:Rank=2+0+0+3+2+1=8

Rank:7

Link:http://www.sciencedirect.com/science/article/pii/S1081120611004261

Preview: Development of the Asthma Control Composite outcome measure to predict omalizumab response.BACKGROUND:Previous assessments of response to omalizumab were based on diary-based data rather than standard validated instruments. A composite instrument that translates diary-based data into standard

validated asthma control measures would characterize patient response to treatment in terms of current asthma control definitions.OBJECTIVE:To develop the Asthma Control Composite (ACC) tool, using real-time diary-based data to predict treatment response in terms of asthma control.METHODS:The ACC tool was derived retrospectively using pooled data from two phase 3 studies in patients with moderate to severe allergic asthma. Patients were randomized to receive subcutaneous omalizumab or placebo for 16 weeks plus stable beclomethasone dipropionate therapy, followed by a 3-month corticosteroid reduction period and 5-month double-blind safety extension. Control was assessed as complete, good, or not controlled, based on a composite score of 4 elements: rescue medication (puffs/day), total asthma symptom score, average number of awakening nights/28 days, and activity limitation.RESULTS:The ACC was mapped to standard validated measures of patient-reported outcomes, with results consistent with clinical outcomes. The proportion of patients with baseline uncontrolled asthma achieving good or complete asthma control was 48% with omalizumab and 32% with placebo at approximately 4 months. The mean composite score also was improved with omalizumab (3.52) vs placebo (2.56) at approximately 4 months.CONCLUSIONS:The ACC tool accurately reflects asthma control in moderate to severe asthma patients

Title: Yes Medication: No Symptoms: No Disease: Yes Strength=0.5 Publication Year=2011

Hence Rank=2+0+0+3+1+1=7

Rank:7

Link:http://www.ncbi.nlm.nih.gov/pubmed/21720220

Preview: Updates on the use of inhaled corticosteroids in Asthma.PURPOSE OF

REVIEW:The purpose of this review is to compare and contrast the newer inhaled corticosteroid (ICS) ciclesonide with older ICSs in terms of pharmacodynamic and pharmacokinetic properties and how these affect comparative efficacy. In addition, clinical dosing strategies for ICSs including as-needed use will be explored.RECENT FINDINGS:Ciclesonide has demonstrated similar efficacy to that of fluticasone propionate and mometasone furoate in equipotent doses with a potentially improved therapeutic index. Once-daily administration of ICSs is generally not as effective as twice-daily. Continuous administration of ICSs does not change the natural history of asthma in either children or adults. Long-term administration of medium dose ICSs does not increase the risk of cataracts or osteopenia in children and young adults. Studies of as-needed ICSs in mild persistent asthma in adults and children have demonstrated mixed results, with some showing equal efficacy to continuous therapy and others showing superiority of continuous therapy.SUMMARY:Ciclesonide provides a newer ICS with favorable pharmacokinetics that may improve the therapeutic index, but assessment of its systemic effects such as growth await further studies. Continuous administration of ICSs in low to medium dose over many years is well tolerated. The use of as-needed ICSs in patients with mild persistent asthma is promising as a potential step-down therapy but awaits further studies.
Title: Yes Medication: No Symptoms: No Disease: Yes Strength: 0.5 Publication Date: 2011

Hence:rank=2+0+0+3+1+1=7

Rank:7

Link:http://www.sciencedirect.com/science/article/pii/S1081120611003929

Preview: Association of ozone exposure with Asthma, allergic rhinitis, and allergic sensitizationTo investigate the effects of air pollution on respiratory allergic diseases in school children.METHODS:A prospective survey of parental responses to International Study of Asthma and Allergies in Childhood questionnaires, together with allergy evaluation, was conducted in 1743 school children selected from metropolitan cities and industrial areas during a 2-year period. Individual exposure to air pollution was estimated by using a geometric information system with the 5-year mean concentration of air pollutants.RESULTS:A total of 1,340 children (male:female ratio, 51.4:48.6) with a mean (SD) age of 6.84 (0.51) years were included in the analysis. Each child underwent allergy evaluation at the time of enrollment and at a 2-year follow-up. After 2 years, the 12-month prevalence of wheezing was significantly decreased, whereas the lifetime prevalence of allergic rhinitis showed a significant increase. Ozone exposure was significantly associated with the 12-month prevalence of wheeze (odds ratio per 5 ppb, 1.372; 95% confidence interval, 1.016-1.852). Ozone was also associated with allergic rhinitis in children who reside in industrial areas. In addition, significant positive associations between ozone and the rate of newly developed sensitization to outdoor allergen were found (P for trend = .007).CONCLUSION:Exposure to ozone was associated with current wheeze and allergic rhinitis. An increased rate of newly developed sensitization to outdoor allergen by ozone may explain the association. Title: Yes Medication: No Symptoms: No Disease: Yes Strength: 0.5 Publication Date:2011

HenceRank=2+0+0+3+1+1=7

Rank:10

Link:http://www.ncbi.nlm.nih.gov/pubmed/19505390

Preview: The role of inhaled corticosteroids in exacerbation is debated. We compared high doses of nebulized budesonide versus high doses of nebulized flunisolide, in association with a short-acting beta-2-agonist, in the treatment of moderate exacerbation in preschool children. In this randomized, parallel group, simple blind study, 46 children aged between 3 and 5 years affected by an acute moderate attack were treated with nebulized flunisolide (Group 1) 40 microg/kg twice daily for 7 days and then 20 microg/kg twice daily for 14 days, or with nebulized budesonide (Group 2) 0.5 mg twice daily for 7 days then 0.25 mg twice daily for 15 days. Inhaled salbutamol (MDI+ spacer - 200 microg 4 times daily) was administered during the first 3 days of the study and then as needed. At T0, T7 and T21 days, airway resistances were evaluated with the forced oscillation technique before and after inhalation of inhaled salbutamol (200 mcg). Parents recorded symptoms and drug use on a diary card. Forty children completed the study. Airway resistances were significantly reduced at T7 (p less than 0.01 flunisolide; p less than 0.05 budesonide) and T21 (p less than 0.05 flunisolide; p less than 0.05 budesonide) versus T0 in both groups, although at T7 the reduction occurred faster in group 1 than in group 2 (p less than 0.01). During the first 7 days of treatment, symptom scores decreased in both groups; however, the decrease was greater in group 1 (p less than 0.05). High doses of inhaled flunisolide and budesonide are both effective in the management of moderate exacerbations in pre-school-age children.

Title: Yes Medication: Yes Symptoms: No Disease: Yes Strength: 1Publication Year: 2009

The paper above has the disease name in the title as well as the abstract also since the exact medication name does not appear in the paper but the synonym Flunisolide of the medication Aerobid occurs in the paper which we obtain from the annotations of the paper as well as the medical data thus the paper gets the score based on the medication score which is 2.The strength of the paper is 1 and publication year is 2009 so the rank is calculated as:

Hence Rank=2+2+0+3+2+1=10

Rank:7

Link:http://www.ncbi.nlm.nih.gov/pubmed/21658314

Preview: Non-allergic rhinitis (NAR) is a heterogeneous disease, characterized by nasal hyperreactivity and inflammation. Its treatment is still debated, intranasal corticosteroids may be an option. The present study is aimed at evaluating the effect of the use of intranasal flunisolide in patients with NAR, considering both clinical and cytological parameters. Sixty patients were treated with intranasal flunisolide (30) or saline solution (30) for 8 weeks. Symptom severity, turbinate size, and inflammatory cell counts were assessed, before and after treatment. Intranasal flunisolide induced a significant reduction of symptoms, turbinate size, and cellular infiltrate. Thus, intranasal flunisolide might be a therapeutic option for NAR

Title: Yes Medication: Yes Symptoms: No Disease: No

Hence Rank=2+2+0+0+2+1=7

The paper above has the medication synonym prescribed for a patient suffering from Asthma and the strength as 1 and publication year as 2010 so the rank is calculated based on the scores for the various semantic concepts.

Hence Rank=2+2+0+0+2+1=7

Rank:6

Link:http://www.ncbi.nlm.nih.gov/pubmed/17509852

Preview: To evaluate the effects of the inhaled flunisolide upon the strength and endurance of the respiratory and peripheral muscles of normal subjects.DESIGN:A randomized, double blind and placebo-controlled study.SETTING:A university-affiliated teaching hospital.PARTICIPANTS:Thirteen normal volunteers selected from a graduation course.INTERVENTION:Subjects were randomly allocated to receive a placebo or corticosteroid (flunisolide) to be inhaled twice a day for 4 weeks. After 2 weeks of a washout period, subjects who were receiving the placebo, received flunisolide and vise versa for another 4-week period.MEASUREMENTS AND RESULTS:Spirometry was used to define the volunteers as being normal in terms of pulmonary function. During the study, subjects performed tests of respiratory muscle function (strength and endurance), measurements of handgrip strength and endurance and anthropometric measurements. Muscle strength was measured each week while muscle endurance was measured every 2 weeks. There was no significant difference in the maximal inspiratory and expiratory pressure and handgrip strength during weeks 1-4 when the subjects used either flunisolide or placebo. However, we observed an increase in the endurance time of the respiratory and handgrip muscles in the 4th week of both flunisolide and placebo use, what may be considered due to a learning

effect.CONCLUSION:Inhalation of flunisolide by normal subjects for 1 month does not cause any acute or clinically perceived effect in the peripheral or respiratory muscles.

Title: Yes Medication: Yes Symptoms: No Disease: No Strength: 0.5 Publication year: 2009

Hence:Rank=2+2+0+0+1+1=6

Rank:6

Link:http://www.ncbi.nlm.nih.gov/pubmed/21573267

Preview: The patient with haematemesis and melaena.Bleeding from the upper gastrointestinal (GI) tract is a common medical emergency, with an incidence of between 50-150 cases per 100,000 per year.1 A recent audit by the British Society of Gastroenterology showed the mortality rate from upper GI bleeds has fallen from 14%2 in 1993 to 10% in 2007.3 However, despite the use of proton pump inhibitors (PPIs), admission rates for peptic ulcer haemorrhage have increased in older age groups,4 probably related to increased use of antiplatelet agents such as aspirin and clopidogrel and anticoagulants in acute coronary syndromes, stroke and atrial fibrillation. The rising age of the population may also have offset further reductions in mortality and morbidity that may have otherwise come about through improved supportive and endoscopic care.

Title: Yes Medication: No Symptoms: Yes Disease: No Strength: 0.5 Publication Year:2011

Hence Rank=2+0+2+0+1+1=6

Rank:6

Link:http://www.ncbi.nlm.nih.gov/pubmed/21359665

Preview: Gastric duplication cysts as a rare cause of haematemesis. Gastric duplication cysts are rare congenital alimentary tract anomalies. We describe the importance of imaging in two children with haematemesis due to gastric duplication cysts. We emphasize the necessity for a high clinical index of

Title: Yes Medication: No Symptoms: Yes Disease: No Strength: 0.5 Publication Year:2011

Hence:Rank=2+0+2+0+1+1=6

Rank:9

Link:http://www.ncbi.nlm.nih.gov/pubmed/12207199

Preview: vomiting, the culminating sign of nausea, is primarily a protective reflex occurring in a wide variety of vertebrates. Even tough nausea and vomiting are among the most basic neural reflexes, they remain poorly understood. Poorly understood are the pathogenetic mechanisms from the anatomic receptor and neuroendocrine point of view. This is the reason why drugs are useful in some types of vomiting but not in others. The aim of this paper is to summarize current knowledge about anatomy of vomiting reflex, neurotransmitter receptor subtypes, agonists and antagonists of serotonin and substance P. Particularly in the treatment of postchemotherpy and postoperative vomiting. It is pointed out that nausea an vomiting may be field of neurochemical and neuropharmacological research. Finally, in clinical research drugs for vomiting therapy may be useful in other pathologies (migraine, rheumatoid arthritis)

Title: Yes Medication: No Symptoms: Yes Disease: Yes Strength:0.5 Publication Year:2009

Hence Rank=2+0+2+3+1+1=9

CHAPTER 8

CONCLUSION AND FUTURE WORK

8.1 Conclusion

The ranking algorithm was run for a number of different test cases and the results obtained were much better than just running a keyword based search for the medical publications through PubMed as we have explained before also that PubMed takes into consideration the publication year for displaying the results of a particular query from most recent to the least and is only based on the keyword based match but our system takes into consideration the various semantics involved for match making and ranking of medical publications with the various patient's profile.

8.2 Future work

The things that we have considered for match making and ranking of documents involve various semantics but still there is a scope of adding a few more semantics for this process. Some of them are:

The age and gender of a person : We can further add the functionality where there is also a match based on the age of a particular patient along with the other semantics involved and thus provide a different rank to the paper which also has a match based on a patient's age and gender.

The country of a patient: There are some diseases that are specific to particular area/countries of the world. This can also be considered as a factor for match making and ranking of the publication based on the patient's profile location.

Medical history of his ancestors: Some diseases are hereditary and hence this can be a factor for match making if a person has some symptoms and those symptoms are of any disease which any of his ancestors might have then based on that we can provide him with the results of publications and rank it accordingly for that patient.

Involvement of a medical practitioner: We can get the expert advice to find out the relevance of our ranking results.

8.2.1 Evaluation Metrics

There are many different measures for evaluating the performance of the information retrieval systems. We can also use these metrics to check the performance of our ranking algorithm:

Precision: It is the fraction of the documents retrieved that are relevant to the user's need.

Recall: Recall is the fraction of the documents that are relevant to the query that are successfully retrieved**.**

Average precision [24]: Average precision is used as a metrics for evaluation when the system returns a ranked set of documents as we need to take into the consideration the order in which the documents appear. We can compute a precision and recall at every position in the ranked sequence of documents, abs then can plot a precision-recall curve, plotting precision $p(r)$ as a function of recall r. Average precision computes the average value of $p(r)$ over the interval from $r = 0$ to $r = 1$:

$$\mathrm{AveP} = \int_0^1 p(r)dr.$$

Then we can compute the mean average precision which is the mean of the average precision score for each query and is given by:

$$\mathrm{MAP} = \frac{\sum_{q=1}^{Q} \mathrm{AveP(q)}}{Q}$$

Where, Q is the number of queries.

These are a few evaluation metrics through which we can compute the performance of our system.

REFERENCES

1. [1] NCBI (NLM NIH) PubMed

2. [2] L. Ramaswamy, and I. B. Arpinar, "Semantics-enabled Proactive and Targeted Dissemination of New Medical Knowledge", CSHALS 2011: Conference on Semantics in Healthcare and Life Sciences, Feb 2011, Cambridge/Boston MA.

3. [3] Plavix Drug Information: Uses, Side Effects, Drug Interactions and Other Warnings

4. [4] Page Rank (Wikipedia)

5. [5] Zolt´an Gy¨ongyi Hector Garcia-Molina Jan Pedersen,Combating Web Spam with Trust Rank

6. [6] Julia Stoyanovich , William Mee , Kenneth A. Ross Semantic Ranking and Result Visualization for Life Sciences Publications

7. [7] Chris Halaschek, Boanerges Aleman-Meza, I. Budak Arpinar, Amit P. Sheth, Discovering and Ranking Semantic Associations over a Large RDF Metabase

8. [8] GoogleHealth (Wikipedia)

9. [9] Microsoft Health Vault (Wikipedia)

10. [10], [11] Semantic Annotation (www.ontotext.com)

11. [12], [13] NCBO Annotator Web Service (Wikipedia)

14. [14] NCBO Bioportal Annotator workflow (Wikipedia)

15. Clement Jonquet, Stanford University, Nigam H. Shah, Stanford University, Cherie H. Youn, NCBO Annotator: Semantic Annotation of Biomedical Data.

16. S., Staab, S., eds, Vol. 96 of Frontiers in Artificial Intelligence

and Applications IOSPress Annotation for the Semantic Web Handschuh

17. [15], [16], [17] UMLS basic tutorial (U.S.NLM)

18. [18],[21],[22] UMLS Metamap basics (NLM)

19. UMLS Quick start guide

20. [20]UMLS Lexical Tool

(www.nlm.nih.gov/research/umls/new_users/online_learning/LEX_001.htm)

21. Bodenreider, Olivier, Hole, William T., Humphreys, Betsy, Roth, Laura, A.Srinivasan, Suresh, Customizing the UMLS Metathesaurus for your Applications.

22. Laura Plaza and Alberto D´ıaz, Retrieval of Similar Electronic Health Records

Using UMLS Concept Graphs

23. [23] Lu Z, PubMed and beyond: a survey of web tools for searching biomedical literature

24. Bodenreider, Olivier, Hole, William; The Unified Medical Language System: What is it and how to use it?

25. [24] Information Retrieval Wikki