SEMANTIC RETRIEVAL AND DISTRIBUTION OF RELEVANT MEDICAL KNOWLEDGE

by

ASMITA RAHMAN

(Under the Direction of Ismailcem Budak Arpinar)

ABSTRACT

In the fast growing world of information, the amount of medical knowledge is growing at an exponential level. It has now become a very difficult task for a regular person to keep up with all the new discoveries and updates in this domain. This thesis describes an approach to semantically retrieve and distribute the medical data/information to the respective health records (people). This system comprises of sample health records and health publications from PubMed. The system performs a semantic matchmaking algorithm to find the relevant publications in PubMed for any particular health record (profile) using BioPortal Ontologies and UMLS. It then assigns a rank based on a semantic ranking algorithm and displays the results to the user. Our system empowers the users and enables them to discover hidden but relevant information. The result of the evaluation clearly proves that our system retrieves all the relevant information better than syntactic searches.

INDEX WORDS: Semantic Matchmaking, Matchmaking algorithm, Semantic Ranking, Knowledge Discovery, Electronic Health Records, Health Publications, Ontology.

SEMANTIC RETRIEVAL AND DISTRIBUTION OF RELEVANT MEDICAL KNOWLEDGE

by

ASMITA RAHMAN

BS, California State University Fullerton, 2009

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

2011

SEMANTIC RETRIEVAL AND DISTRIBUTION OF RELEVANT MEDICAL KNOWLEDGE

by

ASMITA RAHMAN

Major Professor: Ismailcem Budak Arpinar

Committee: Khaled Rasheed
Thaib Taha

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2011

DEDICATION

To my parents.

ACKNOWLEDGEMENTS

First and the foremost I would like to thank God for his blessings and for making it possible for me to work on a project that I truly enjoyed. Then, I would like to thank my major professor Dr Arpinar for all his guidance and support throughout the project. He has been a huge inspiration and provided great help and key suggestions throughout for the success of this project. I would also like to thank Dr. Ramaswamy, my team partner in this project Priya Wadhwa and my committee members Dr. Taha and Dr. Rasheed for their time and support.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

CHAPTER 1

OVERVIEW

We all know that today the knowledge in the medical domain is growing at a very fast pace. It is becoming harder and almost impossible for a normal person to keep up with all the updates in this field. Every day there are several new drugs coming to market, several new treatment options being introduced, many old medications being replaced, several discoveries being made etc. In this fast moving world, there is barely any time left for a normal person to read and research about the new updates in the medical industry. Our research is going to contribute in this field by making relevant information easily available.

1.1 Motivation:

1.1.1 Motivating Scenario # 1:

Consider a woman Martha who is 65 years old. She is retired and lives alone. She is suffering from mild Asthma and thus takes regular medication and uses inhaler to stay healthy. Her condition is quiet stable and therefore she visits the doctor every six months to a year for yearly check-ups. She uses the inhaler on a daily basis and the only way for her to know about a change in it or about a new discovery is through her doctor, which is once a year. Now, let's say that she was on Primatene Mist Asthma Inhaler. Recently,

there is a change in this inhaler and it has been announced to be taken away from the shelves. However, since she has enough stock at home, she will continue to use the inhaler until she visits the store for more of the same inhaler or until she hears something from her doctor.

This may be very dangerous and unsafe as she continues to use the inhaler which has been removed. In addition, she is unknowingly avoiding taking a better inhaler because of her lack of knowledge. In this case, the affects of ignorance can be severe. However, our system makes this knowledge discovery easier for her. With the help of our system any update related to her disease, drugs, medications, symptoms etc. will be sent to her. She does not need to read through several hundred new publications at PubMed or search through hundreds of pages on the internet to find the most relevant information. In our system, all this related information is provided via semantic matchmaking and the results would be relevant and accurate.

1.1.2 Motivating Scenario # 2:

Mr. Burton is a patient of Dr. Brown. Mr. Burton has had a heart attack in 2005. Dr. Brown has prescribed the drug Plavix to reduce the risk of future heart attacks. As Plavix leads to acid reflux, the doctor has also prescribed the drug Prilosec to lower acidity. Note that until recently this has been the standard treatment regimen for patients with heart attack histories. In March 2009, a study appeared in the Journal of American Medical Association, which indicated that combination of drugs Clopidogrel (Plavix is

the brand name of Clopidogrel) and proton pump inhibitor (PPI) Prilosec is one of the PPIs) in patients with previous histories of heart attacks can actually double the risk of second heart attack. This research finding has direct implication on the treatment regimen of Mr. Burton as it puts him in high-risk category for a second heart attack. Currently, there are a few ways in which Dr. Brown can learn about the discovery: (a) searching and browsing relevant web sites (e.g., PubMed); (b) attending a conference/ professional meeting where recent research findings are discussed; or (c) through colleagues who may have knowledge about the new discoveries. However, in all of these methods, there could be significant delays between publishing of new information and Dr. Brown becoming aware of the information. Even after Dr. Brown becomes aware of the study, his staff has to search through patients' medical records to identify the patients who are on Plavix and Prilosec simultaneously which can be difficult process [26].

Since the matchmaking in our system is done on the semantics rather than syntax, the knowledge discovery enables the system to find such relevant publications and provide the results to the patient.

Today, the healthcare industry is heavily adopting the trend of information technology in the form of creating electronic personal health records (PHRs). This enables the patient to have more control on their health. This allows them to be pro-active and informed about the decisions they make. The Markle foundations connecting for health collaborative, has defined a PHR as an "electronic application through which individuals can access, manage and share their health information and of others for whom

they are authorized, in a private, secure and confidential environment." With the trend of having electronic personal health records, several new applications are now being built on top of it. These applications are third party applications that build on top of the personal health record use the information in the personal health records provide various new services.

Our system is also a similar kind of application that takes the information stored in the personal health records, processes that information, performs the matchmaking and ranking and then displays the results. This matchmaking is done between the personal health records and pubMed publications enabling the system to display the most relevant publication to a particular health record. This empowers the users and doctors to take control of their health decisions and to easily stay informed about the new discoveries related to their disease, symptoms or medications. The same system would also be useful for people who have certain symptoms but are un-aware of the disease that they may be suffering from. This would also be very beneficial to people suffering from incurable diseases such as cancer, where every new research and discovery brings a new hope for life.

1.2 Outline:

The Chapter 2 talks about the related works in this field. Chapter 3 discusses our approach to the solution with an overview of our system's functionality. Chapter 4 talks about the building blocks of the system as well as the technologies used. Chapter 5 discusses the semantic matchmaking algorithm. Chapter 6 discusses the testing of the

matchmaking algorithm with a comparison to syntactic matchmaking. Chapter 7 talks about the semantic ranking algorithm. Chapter 8 gives a complete example workflow of the system. Chapter 9 gives a summary of the work with proposed future works.

CHAPTER 2

RELATED WORKS

2.1 TrialX

TrialX is a system, a third party tool that is built for recruiting related heard records for clinical trials. As one must realize that before any medication becomes available to the market, there are clinical trials performed to measure the efficiency and side effects of the same. However, this process of clinical trials currently takes over years due to the fact that finding appropriate people for testing the drug is a laborious process. However, TrialX makes it easier for the people to find the clinical trials related to their health record. It performs a matchmaking algorithm and finds the related clinical trials for any particular health record. Here is the overview the information flow of the system:

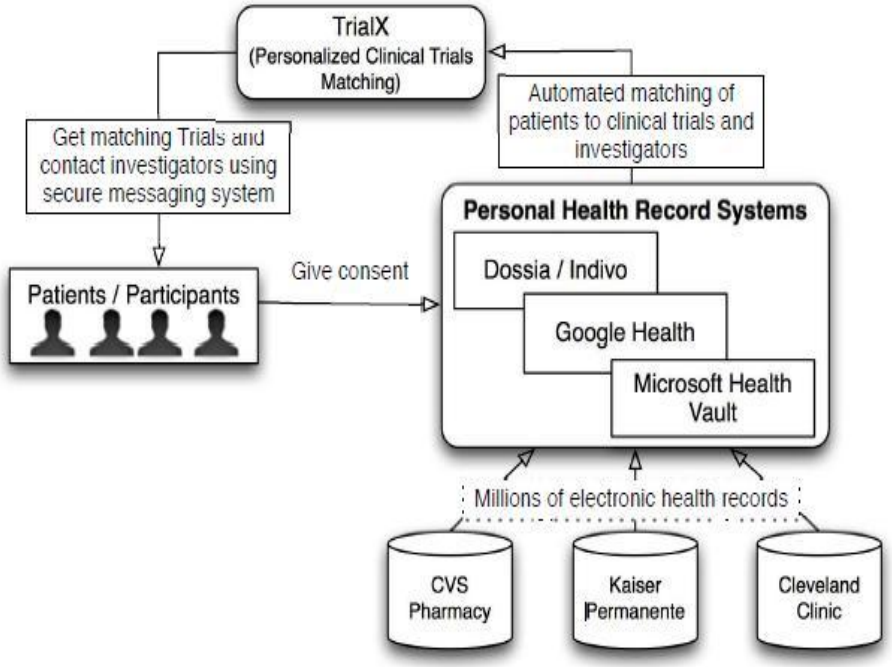


Figure 1: Overview of TrialX [15]

The Figure 1 shows the information flow in the TrialX System. The system is built on top of the Personal health records and provides the results of matching clinical trials. Here are the core components [15] of the trialX system:

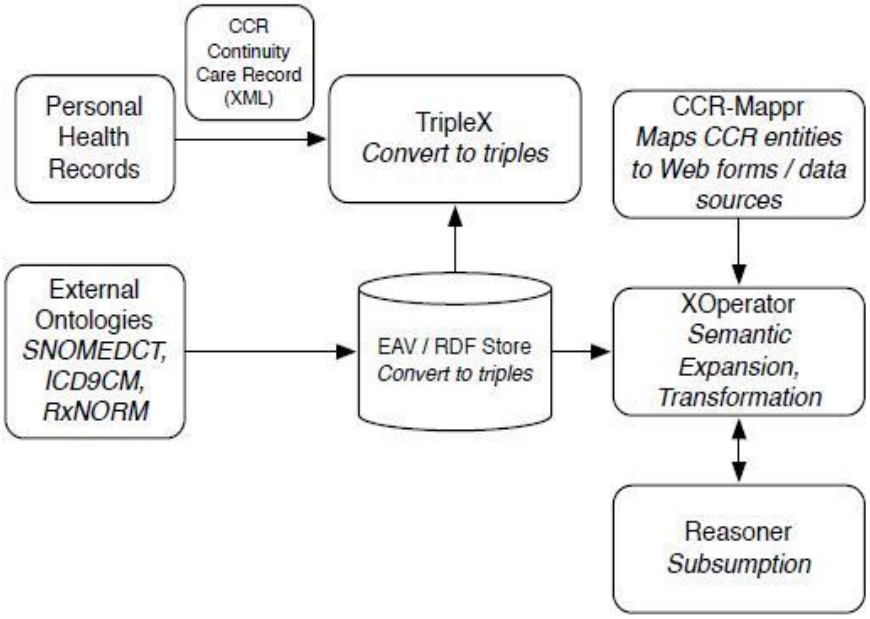


Figure 2: Core Components of TrialX [15]

Figure 2 shows the core components. The health records as well as the concepts are stored in the form of RDF triples. Once the data set is in the form of triples, matchmaking is performed and the results are obtained.

For the purpose of matching, Columbus matching algorithm is used. This algorithm uses Natural Language processing techniques with UMLS (Unified Medical Language System) to obtain the results. Another component of the system is the form interface. This form interface allows the health record owner to add and remove clinical trials from their profile. It also allows them to manage new and upcoming clinical trials.

2.2 Google Health:

The idea behind Google Health is their motto “Better Health comes from Better Information” [21]. It allows people to manage, track, organize and act on their health information. It offers a single place to store and share all the information. As a user, one can decide how much information one would like to share and how much information should be kept private.

It enables the users to have an electronic health record and enables them to take advantage of third party tools available for the Google Health users. In addition, one can set-up personal goals and use a timeline to track them. However, Google would be discontinuing this service in Jan 2012. Here is a snapshot of Google Health Homepage:

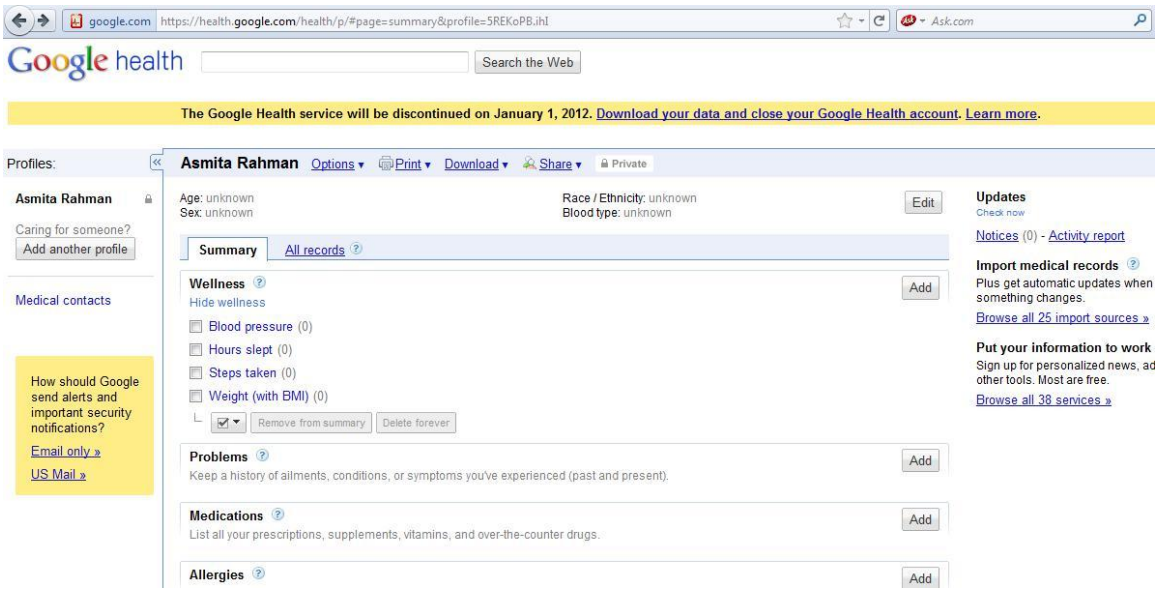


Figure 3: Google Health SnapShot

2.3 Microsoft Health Vault:

Microsoft Health Vault is an online tool enabling users to store all their health information at one location [17]. Similar to Google Health, it provides a way to manage and track the health information. It has the following additional features:

1. It helps to prepare for family emergencies
2. It helps users to access complete family's health information at one place
3. It helps users to monitor their health conditions and stay in touch with their doctor(s).
4. It helps users to track their progress
5. It provides control to the users enabling them to decide who sees how much of their health information.
6. It provides several tools built on top of it, that a user can take advantage of depending on their health preferences and conditions.

Here is a snapshot of the homepage of Microsoft Vault:

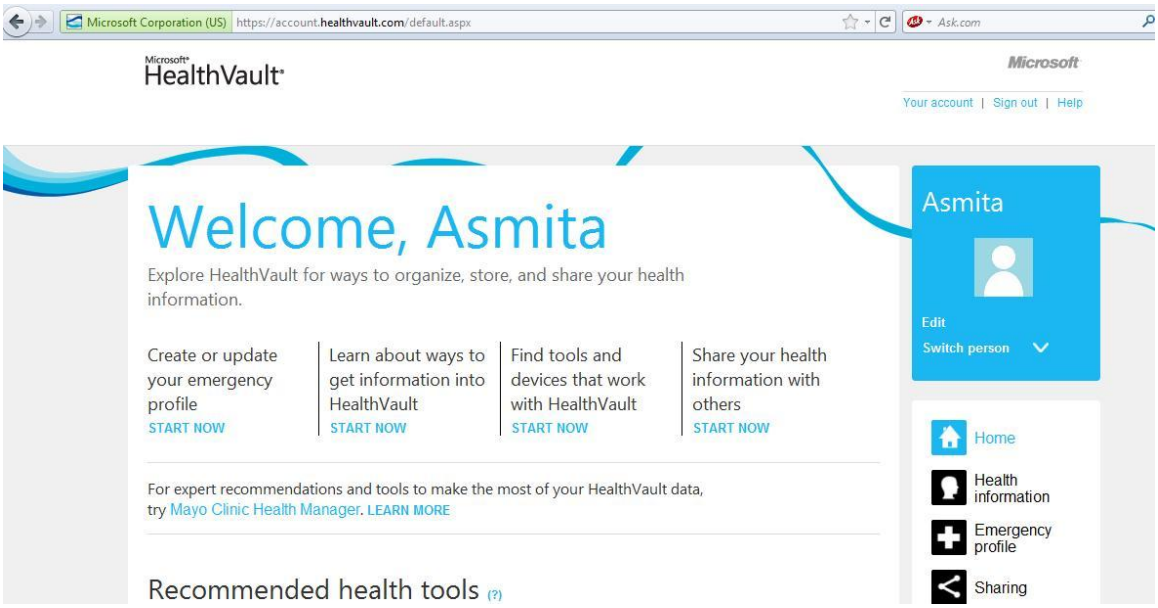


Figure 4: Microsoft Vault Snapshot

2.4 Life Record- Personal Medical record System:

My Life Record software is an alternate solution for managing the health information. It is available in the form of an IPhone app enabling easy access to its users. It has the following features [18]:

1. Keeps official and verifiable copies of your records
2. Keeps track of your medications
3. Keeps track of your laboratory information
4. Keeps track of your doctors, doctor's visits, appointments etc.
5. Portable, available as an Iphone app
6. Empowers the health record owner to make informed decisions.

This software is developed by Life Record Inc. who has been developing medical software since 1998. Here is a snapshot of this application:

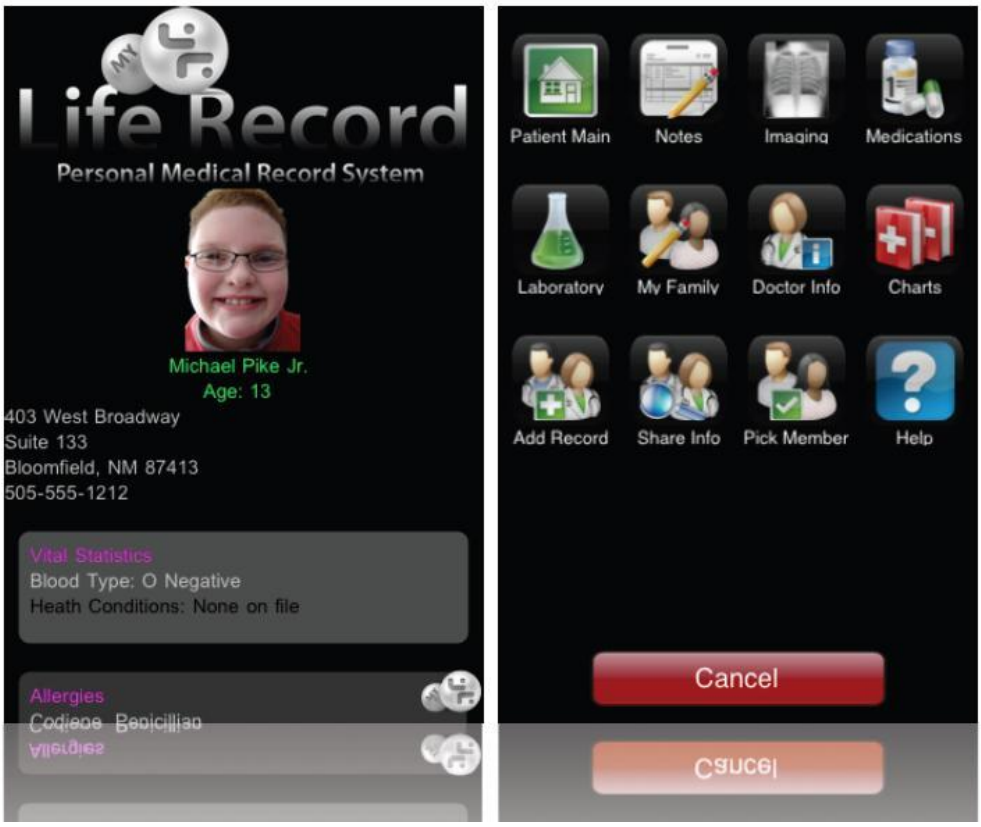


Figure 5: Life Record Snapshot [19]

2.5 Retrieval of Similar Electronic Health Records Using UMLS Concept Graphs [23]:

Physicians are often faced with a decision making challenge, in which case they can use the information available to them about the previous clinical trials. However, since the amount of information in this field is large, exhaustive search is unfeasible. This

paper proposes an approach to deal with this issue. They propose an approach for the retrieval of similar clinical cases, based on mapping the text onto UMLS concepts and representing the patient records as semantic graphs. They also did a thorough evaluation of the proposed method and the results show that their method correlates well with the expert judgments and outperforms remarkably the traditional term-vector space model.

2.6 Fighting Diabetes with Information [24]:

This research paper talks about the great potential of the interrelationship of information, people and technology for improving the health care. The research discusses several information challenges associated with diabetes. This allowed researchers unfamiliar with healthcare to observe the social and organizational factors in the ebb and flow of information around complex diagnoses. In addition, this paper suggested addressing a set of problems that will improve the lives of not only the patients but also their families, and friends. It will also make the provision of diabetes care more effective and cost efficient.

Similar to the systems mentioned above, our system is going to empower the users (both patients and doctors) to be able to make healthy decisions. However, in addition to the above systems, our system is going to use semantics for the purpose of matchmaking which will enable both direct and un-direct matches to be detected during the matchmaking process. This semantic matchmaking will give an edge to our system over all the other systems in this field.

CHAPTER 3

APPROACH

3.1 Overview:

This system consists of two major parts:

- Semantic Matchmaking
- Semantic Ranking

The matchmaking performs all the core operations of finding the relevant results for any particular health record. Once the results are found the Semantic Ranking provides us a way of calculating the relevance to a particular record.

The matchmaking and the ranking process would be is performed semantically which will enable the system to use ontology mapping, synonyms calculation and hierarchy verification for better results.

3.2 How it works:

Here is a diagram showing the overview of the functionality of our system:

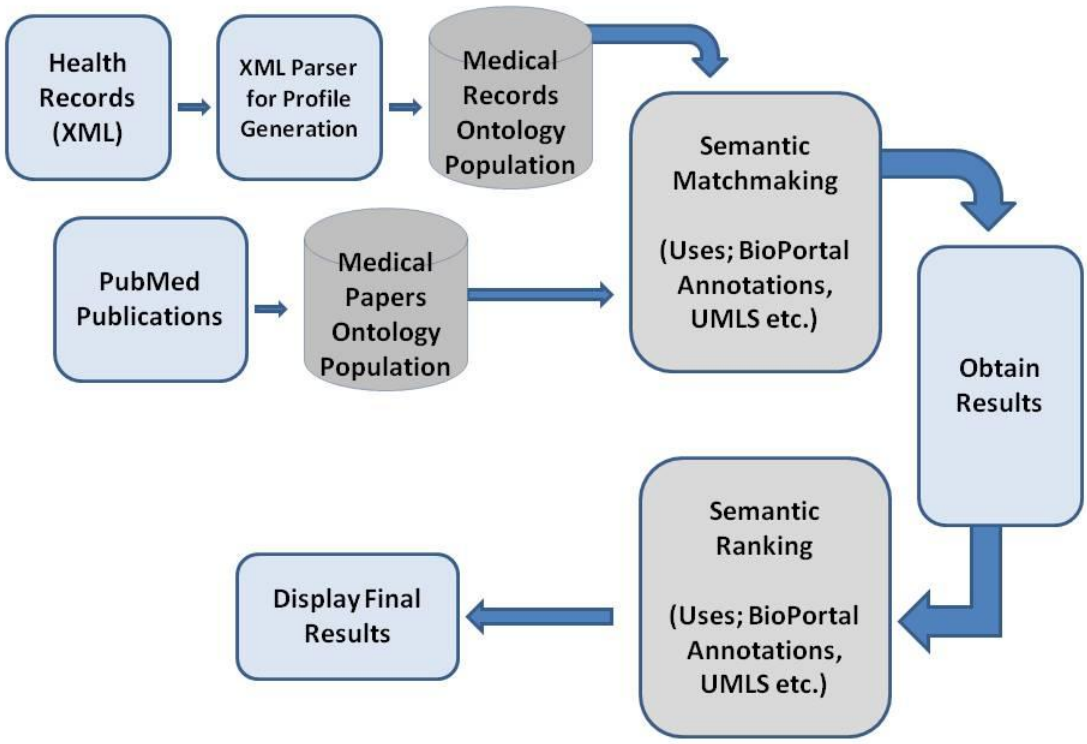


Figure 6: Overview of the System

Our health record consists of the following personal information. Here is a sample record in XML format:

<Patient>

<Name>John Smith</Name>

<Address>123 main st</Address>

<City>Duluth</City>

<State>Georgia</State>

<Zip>30098</Zip>

<Country>United States</Country>

<Id>1234</Id>

<Age>34</Age>

<KnownDisease>Anemia</KnownDisease>

<Medications>Feosol, Slow-Fe</Medications>

<Gender>Male</Gender>

<symptoms>Headache</symptoms>

<PrimaryPhysician>Dr Smith</ PrimaryPhysician>

<PhysicianId>dc1245</PhysicianId>

<PrimaryPharmacy>CVS pharmacy</PrimaryPharmacy>

<PrimaryPharmacyId>235Phar</PrimaryPharmacyId>

</Patient>

The above template was used for generating test health records for our system. Since there was no standard found for generating health records, we used Google health's format as the reference. This health record information is then parsed to create a patient profile with all the pertinent information. Once the profiles have been generated, all the data is populated into an ontology for semantic matchmaking. On the other hand, the PubMed publications were downloaded. Another ontology was populated consisting of all the information about the medical papers.

Once both the ontologies have been populated with health records and medical publications information respectively, the system can begin the matchmaking procedure.

One of the most important parts of matchmaking is to be able to annotate the unstructured text. The publications of the PubMed including their title and abstract would be supplied and would need to generate annotations for matchmaking purposes. Also, the same would be used to annotate the information received from the health records. Once both the annotations are received, one can continue in the matchmaking procedure.

CHAPTER 4

BUILDING BLOCKS

The following components played an important role in the implementation of the system:

- Google Health Records
- PubMed
- UMLS
- NCBO BioPortal
- MetaMap

4.1 Google Health Records:

In order to be able to test the system, one must realize the need of health records. However, due to the sensitivity of health records and the information within, it is nearly impossible to be able to work with real records. In order to deal with this shortcoming, I created sample health records for testing purposes. These records were created in the same format as the format provided by Google Health. Here is a sample of the Google Health Record, available on the web to download [9]:

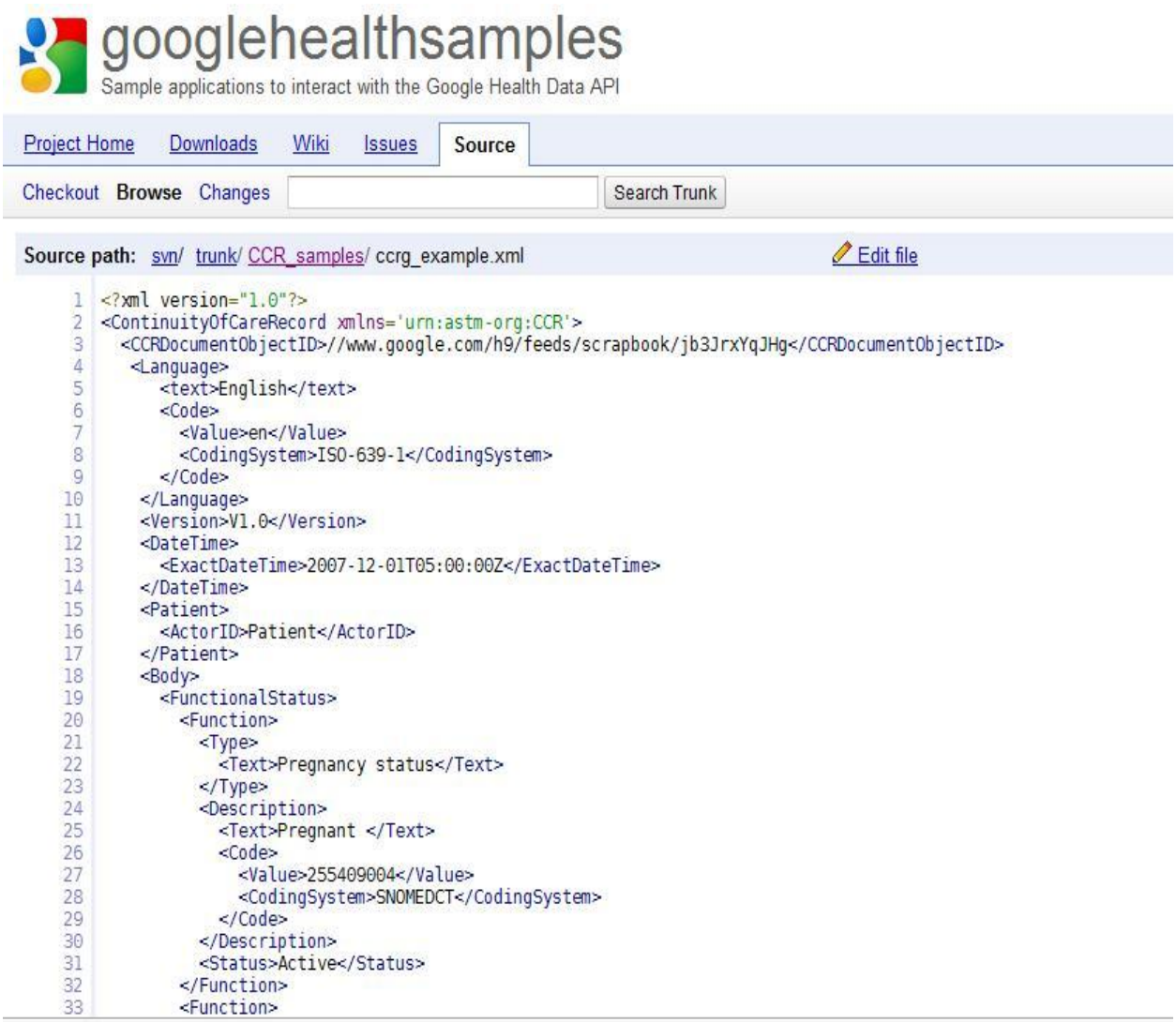


Figure 7: Google Health Samples [9]

We generated sample health records (75) in XML format, similar to Google Health records for testing purposes. Attempts were made to obtain real health records for testing, however, due to the privacy issue and the sensitivity of the information we were not able to obtain real personal records. However, our sample health records would enable the application to work properly when fed with real health records. Here are a few samples of the health records that were generated:

```
  <Patient>
   <Name>Robin Woods</Name>
   <Address>1563 South Milton st</Address>
   <City>Tuscon</City>
   <State>AZ</State>
   <Zip>92009</Zip>
   <Country>United States</Country>
   <Id>1235</Id>
   <Age>25</Age>
   <KnownDisease>Asthma</KnownDisease>
   <Medications>Aerobid, Alvesco</Medications>
   <Gender>Male</Gender>
   <symptoms>Vomiting</symptoms>
   <PrimaryPhysician>Dr Smith</ PrimaryPhysician>
   <PhysicianId>dc1247</PhysicianId>
   <PrimaryPharmacy>Walgreens</PrimaryPharmacy>
   <PrimaryPharmacyId>247Phar</PrimaryPharmacyId>
</Patient>
<Patient>
       <Name>John Childs</Name>
       <Address>6776 South Northridge Ave</Address>
       <City>West Covina</City>
       <State>CA</State>
       <Zip>91790</Zip>
       <Country>United States</Country>
       <Id>1236</Id>
       <Age>68</Age>
       <KnownDisease>High Blood Pressure</KnownDisease>
```

```
        <Medications>chlorthalidone, Hygroton, Diuril, chlorothiazide
</Medications>
        <Gender>Male</Gender>
        <symptoms>Loss of breadth</symptoms>
        <PrimaryPhysician>Dr George</PrimaryPhysician>
        <PhysicianId>dc1248</PhysicianId>
        <PrimaryPharmacy>CVS pharmacy</PrimaryPharmacy>
        <PrimaryPharmacyId>235Phar</PrimaryPharmacyId>

</Patient>

<Patient>
        <Name>Andrew Smith</Name>
        <Address>10 Miledge Ave</Address>
        <City>Athens</City>
        <State>GA</State>
        <Zip>30606</Zip>
        <Country>United States</Country>
        <Id>1237</Id>
        <Age>88</Age>
        <KnownDisease>Anemia</KnownDisease>
        <Medications>Feosol, Slow-Fe </Medications>
        <Gender>Male</Gender>
        <symptoms>Loss of energy</symptoms>
        <PrimaryPhysician>Dr Keith</PrimaryPhysician>
        <PhysicianId>dc1241</PhysicianId>
        <PrimaryPharmacy>Kroger</PrimaryPharmacy>
        <PrimaryPharmacyId>239Phar</PrimaryPharmacyId>
 </Patient>
```

The Name, Address, City, State, Zip, Country, Age and Gender contain the information of a particular patient. The ID is generated by the system and is a unique identifier. In addition, the nodes “KnownDisease” contain the information of any disease or diseases that the particular patient might be suffering from. This node may be empty if the patient does not have any current known disease. The nodes “Medications” contain the names of current medications that the patient might be on. Similarly, the node “Symptoms” contains the current symptoms. They may or may not be empty. The PrimaryPhysician is the name of patient’s primary physician with its ID in PhysicianId. In addition, the Pateient’s primary pharmacy name and ID is stored in PrimaryPharmacy and PrimaryPharmacyId respectively.

The creation of the sample records allowed us to test the application in various scenarios and analyze the efficiency of the system.

4.2 PubMed:

PubMed comprises more than 21 million citations for biomedical literature. The sources of these citations are MEDLINE, life science journals, and online books. These citations are a combination both links to full-text content from PubMed Central and from publisher web sites [10]. PubMed is maintained by the National Center for Biotechnology Information(NCBI), at the U.S. National Library of Medicine[11].

Pubmed is a free resource and it provides an easy to use search interface to search the publications via the title, journal name, names of authors, specific citations, keywords etc. It then displays the related results in the following format:

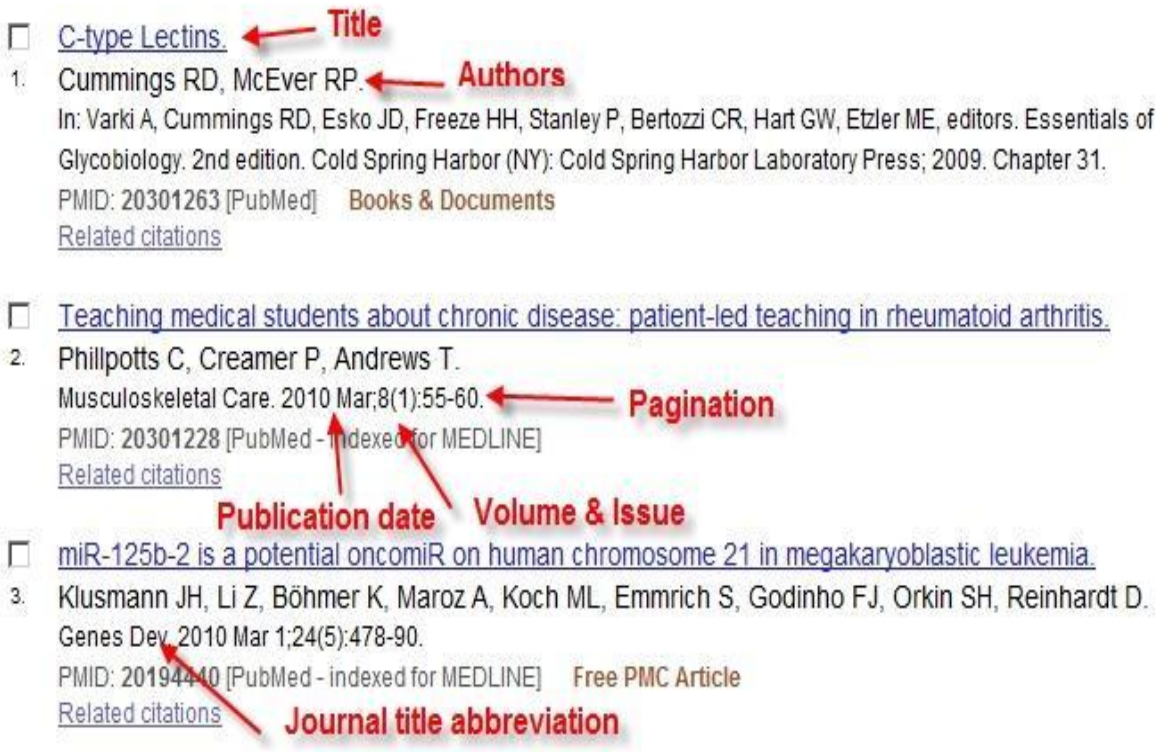


Figure 8: PubMed Sample Output[12]

We have used PubMed as the knowledge resource in this research. About a couple hundred research publications (Abstracts) were downloaded, annotated and then the knowledgebase (Ontology) is populated. This would allow the system to be able to do proper matchmaking and display relevant results.

4.3 UMLS:

UMLS stands for Unified Medical Language System and it is a system that brings together health vocabularies, biomedical terms and standards. It enables to enhance and develop applications with use of such information and promotes interoperability. It is a source of a large number of national and international vocabularies and classifications (over 100) and provides a mapping structure between them [13].

The UMLS can be used to design information retrieval for patient record systems, to facilitate the communication between different systems, or to develop systems that parse the biomedical literature. UMLS consists of three knowledge sources [14]:

- Metathesaurus,
- Semantic Network,
- SPECIALIST Lexicon and Lexical Tools.

4.3.1 Metathesaurus:

This serves as the base of the UMLS. It contains over 1 million biomedical concepts and 5 million concept names. The Metathesaurus is organized by concept and each concept has specific attributes defining its meaning. Each concept is also linked to the corresponding concept names in the various source vocabularies. There are several

relationships established between the concepts such as : is a, is part of, is caused by etc. In addition, all hierarchical information is retained in the Metathesaurus.

4.3.2 Semantic Network:

As we know now that Metathesaurus consists on concepts, each concept in Metathesaurus is assigned to a semantic type. These types are then related to each other via semantic relationships. Semantic Network comprises of all such semantic types and relationship. Currently there is a total of 135 semantic types and 54 relationships.

1. Semantic Types: Semantic types consist of the following:

    - Organisms
    - Anatomical structures
    - Biologic function
    - Chemicals
    - Events
    - Physical objects
    - Etc.

Here is a portion of the UMLS Semantic Network, showing the "Biologic Function" Hierarchy:

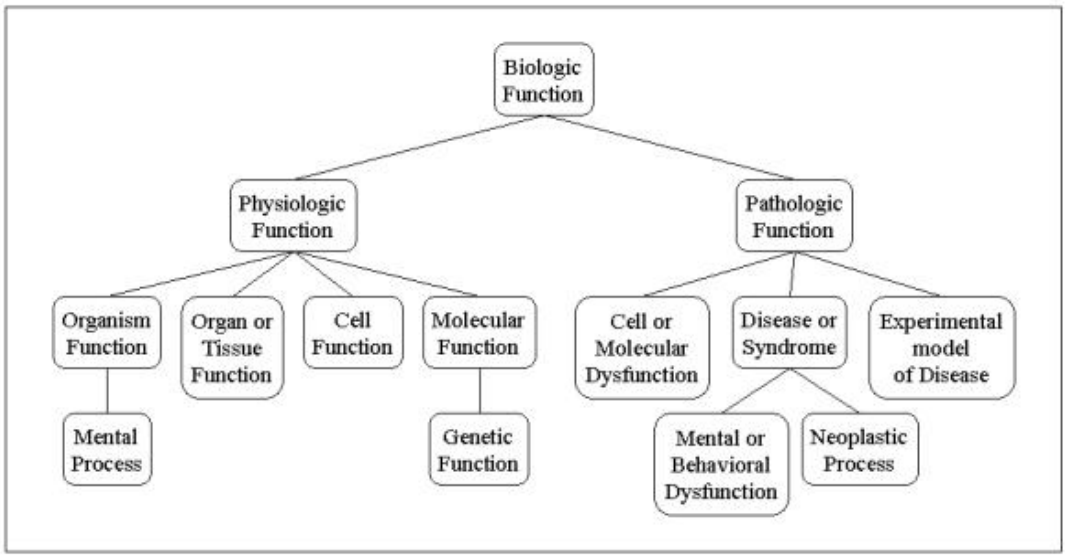


Figure 9: A Portion of the UMLS Semantic Network: "Biologic Function" Hierarchy [25]

2. Semantic Relationships: The primary relationship is an "isa" relationship, which identifies a hierarchy of types. The network has another five (5) major categories of non-hierarchical relationships; these are:
    - "physically related to"
    - "spatially related to"
    - "temporally related to"
    - "functionally related to"
    - "conceptually related to"

The information about a semantic type includes: identifier, definition, examples, hierarchical information about the encompassing semantic type(s), associative relationships (known as weak relationship).

4.3.3 SPECIALIST Lexicon:

This contains information about English language, biomedical terms, terms in Metathesaurus and terms in MEDLINE. It contains the syntactic information, morphological information as well as Orthographic information:

- Syntactic Information: This contains the information on how the words can be put together to generate meaning, syntax etc.
- Morphological Information: This contains information about the structuring and forms.
- Orthographic Information: This contains information about the spellings.

UMLS plays a vital role in our system as we use it for the purpose of obtaining annotations. The annotations in turn play a vital role in the matchmaking process. These annotations are obtained in two steps. First is the direct annotations are obtained by matching the raw text with the preferred name and then expended annotations by considering the UMLS ontology mappings and hierarchy.

4.4 NCBO BioPortal

NCBO (National Center for Biomedical Ontology) offers a BioPortal, which can be used to access and share ontologies that are actively used on the biomedical community. By using the BioPortal, one can search the ontologies, search

biomedical resources, obtain relationship between terms in different ontologies, obtain ontology based annotations of the text etc. Bio portal is a web- based application [4]. It can be used for the following:

- Browse, find, and filter ontologies in BioPortal library
- Search all ontologies in the BioPortal library with specified terms
- Submit a new ontology to BioPortal library
- Views on large ontologies
- Explore mappings between ontologies

4.4.1 BioPortal's Implementation and functionality:

BioPortal provides access to one of the largest repositories of biomedical ontologies. We can access these by one of the following ways:

- Web Browsers
- Web Services (RESTful services)

The BioPortal library consists of the following:

- Total number of ontologies: 173
- Number of classes/types: 1,438,792

These ontologies provide us a basis of the domain knowledge which can be used for data integration, information retrieval etc. Here is a snapshot of the NCBO BioPortal (via web browser):

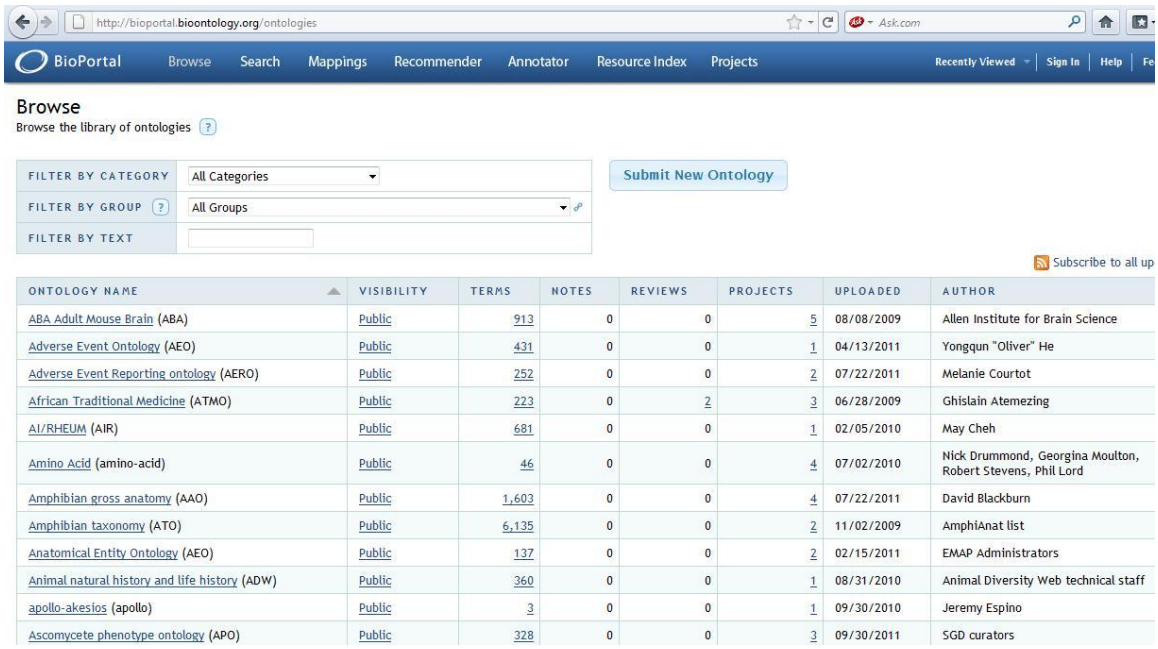


Figure 10: BioPortal Snapshot

4.4.3 NCBO Annotator:

The NCBO annotator provides us with a web service that we can use to process text to recognize relevant biomedical ontology terms. The NCBO Annotator annotates or "tags" free-text data with terms from BioPortal and UMLS ontologies. It can be accessed via the browser or via the web service. The web service is flexible enough to allow for customizations particular to any application[5]. For example we can limit results to a particular ontology (e.g. Anatomical entity Ontology) or to a certain UMLS semantic type (e.g. T017 for 'Anatomical Structure').

The concept recognition engine is called MGREP. It was developed by the National Center for Integrative Bioinformatics and is combined with BioPortal Ontology Web services to create the NCBO Annotator service to make the task of

creating ontology-based annotations accessible for any biomedical researcher. Here is a snapshot of the Annotator via the browser:

Figure 11: Annotator Snapshot

The annotations are performed in two steps; first is the direct annotations by matching the raw text with the preferred name and then expanding the annotations by considering the ontology mappings and hierarchy. Here is the workflow of the annotator web service:

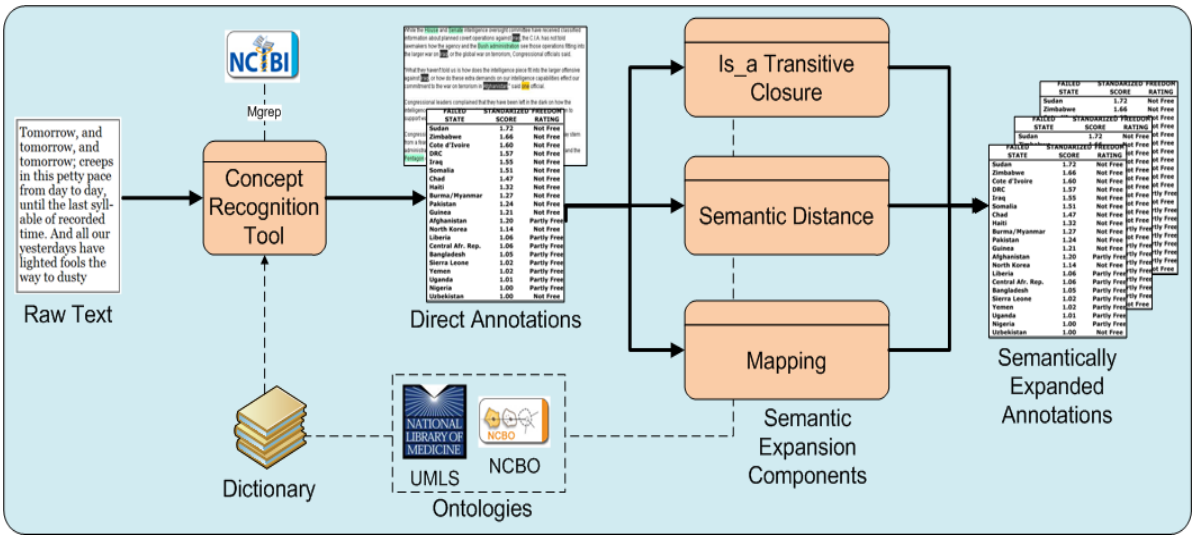


Figure 12: Annotator Workflow [22]

The diagram shows the information flow in the web service of annotator. The process starts from raw text, from which direct annotations are obtained based on syntactic concept recognition. These annotations are obtained using the concept recognition tool provided by NCBI, which gives direct annotations from the given text. Once we have the direct annotations, these annotations are then expanded with the help of introducing semantics. These semantics are introduced by the ontologies (UMLS and others) and expanded annotations are obtained.

4.5 MetaMap:

Another alternative solution for obtaining annotations of free text medical documents is MetaMap. Meta Map is a highly customizable program developed by National Library of Medicine (NLM) to annotate and map biomedical text.

MetaMap can be used for mapping the text into the concepts from UMLS Metathesaurus. This is achieved by processing the text, taking it through a series of procedures and finally breaking it down into the components that include sentences, phrases, lexical elements and tokens [6]. On the other hand tentative concepts from UMLS Metathesaurus are retrieved and evaluated against the results obtained. Finally, the final mapping is obtained which best describes the text originally entered.

Meta map provides a link between the text of biomedical literature and the knowledge, including synonymy relations, embedded in the Metathesaurus. It arose in

the context of an effort to improve biomedical text retrieval, specifically retrieval of MEDLINE/PubMed citations. Here are some of the common uses of MetaMap [7]:

- Information extraction
- Classification/categorization
- Text summarization
- Question answering
- Data-mining
- Literature-based discovery
- Text understanding
- UMLS concept-based indexing and retrieval
- Natural-language analysis of biomedical literature and clinical text

4.5.1 System requirements for MetaMap:

In order to use MetaMap, we must deploy it locally and then operate. The system requirements include: 7GB of disk space and 1GB of Memory.

4.5.2 MetaMap's Functionality:

The text goes through several modules throughout the process. Here is a diagram indicating the functionality and workflow of Meta map in detail:

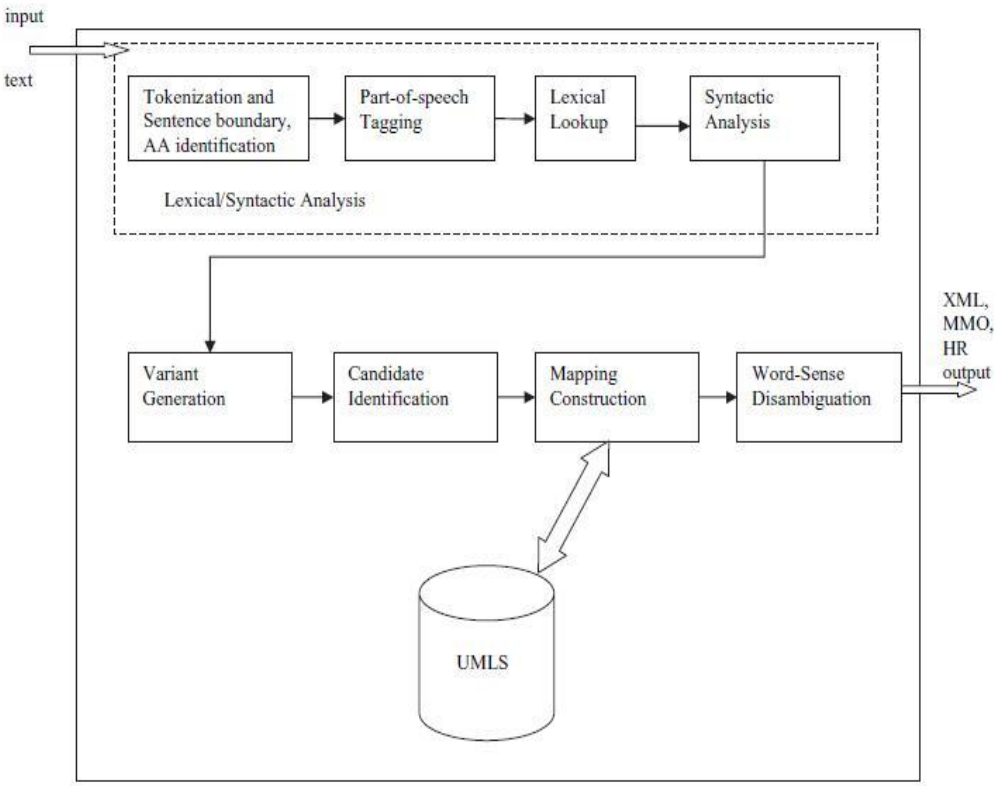


Figure 13: MetaMap Functionality [8]

The process starts from raw text which then processes through several modules including tokenization, part of speech tagging, lexical lookup, syntactic analysis, variant generation, candidate identification, mapping construction (UMLS), word- sense disambiguation. Finally the output is received in XML.

4.6 NCBO vs. MetaMap:

As we can see from the above details both NCBO annotator and MetaMap provide similar services. For the purpose of our research, we may use either one of these suitable solutions. However, there are some subtle differences which may give a preference to one solution over another.

We must consider the fact that BioPortal's solution can be used via the browser or via a web service. We can easily customize the web service and thus will not need to download everything on our local machine. On the contrary, MetaMap would require us to have it implemented locally, consuming a lot of disk space and memory.

Also, the speed of execution is another point to consider. Based on a research done in 2010 on concept recognition by both BioPortal and MetaMap, it was evident that BioPortal annotator was faster than MetaMap [3]. While our current test set of health records is relatively small, once this solution is built on a larger scale which would deal with millions of records, this time efficiency would become very significant.

Another deciding factor that gives BioPortal an edge over the Metamap is the fact that BioPortal is using 173 Ontologies in addition to UMLS for obtaining the mappings and hierarchy. Although UMLS plays the vital role in the annotations, the use of more ontologies increases recall. From the above, it can be seen that BioPortal

and MetaMap are very comparable. They have different functional workflows for obtaining the annotation of unstructured text. However, both have some pros and cons.

In light of our research; BioPortal Annotator is a better fit to our project. It will provide an easy to access web service that can be used to obtain the annotations. In addition, the web service can also be customized to fit our needs. Also, the usage of UMLS with another 170 ontologies would give better results. Here is a sample output of annotations obtained for disease “Anemia”:

ObaResultBean [
ResultBean [
resultID = OBA_RESULT_c925
statistics = [(MGREP, 26) , (CLOSURE, 0) , (MAPPING, 0)]
parameters = [longestOnly = false, wholeWordOnly = true, filterNumber = true, withSynonyms = true, withContext = true, ontologiesToExpand = [], ontologiesToKeepInResult = [], isVirtualOntologyId = false, semanticTypes = [T020, T052, T100, T003, T087, T116, T011, T190, T017, T008, T195, T194, T007, T053, T038, T123, T091, T122, T012, T029, T023, T030, T031, T022, T118, T088, T025, T026, T043, T049, T103, T120, T104, T185, T201, T200, T077, T019, T056, T060, T047, T203, T065, T111, T196, T018, T071, T069, T126, T204, T051, T050, T099, T033, T013, T168, T021, T169, T004, T028, T045, T083, T064, T096, T102, T131, T058, T093, T125, T016, T068, T078, T129, T130, T055, T037, T197, T170, T009, T998, T034, T059, T171, T119, T066, T015, T073, T074, T048, T041, T063, T044, T085, T070, T999, T191, T124, T114, T086, T090, T057, T042, T109, T001, T032, T040, T092, T115, T046, T101, T121, T067, T072, T039, T002, T098, T097, T094, T080, T081, T192, T089, T014, T062, T075, T006, T095, T184, T054, T082, T110, T167, T079, T061, T024, T000, T010, T005, T127], levelMax = 0, mappingTypes = [null], stopWords = [], withDefaultStopWords = true, isStopWordsCaseSenstive = false, text to annotate = Anemia]
]
ontologies = [[ICPC-2 PLUS, nbAnnotation: 2, score: 32, (42297, 2005, 1429)], [SNOMED Clinical Terms, nbAnnotation: 3, score: 30, (46116, 2010_07_31, 1353)], [Logical Observation Identifier Names and Codes, nbAnnotation: 2, score: 20, (44774, 232, 1350)], [MedDRA, nbAnnotation: 1, score: 10, (42280, 12.0, 1422)], [National Drug File, nbAnnotation: 1, score: 10, (40402, 2008_03_11, 1352)], [Human Phenotype Ontology, nbAnnotation: 1, score: 10, (45774, unknown, 1125)], [Mammalian phenotype, nbAnnotation: 1, score: 10, (45771, 1.419, 1025)], [MedlinePlus Health Topics, nbAnnotation: 1, score: 10, (40397, 20080614, 1347)], [Medical Subject Headings, nbAnnotation: 1, score: 10, (44776, 2011_2010_08_30, 1351)], [Human disease, nbAnnotation: 1, score: 10, (45769, unknown, 1009)], [Symptom Ontology, nbAnnotation: 1, score: 10, (44749, unknown, 1224)], [CRISP Thesaurus, 2006, nbAnnotation: 1, score: 10,

(44432, 2006, 1526)], [NCI Thesaurus, nbAnnotation: 1, score: 10, (45400, 11.01e, 1032)], [Online Mendelian Inheritance in Man, nbAnnotation: 1, score: 10, (45553, 2010_04_08, 1348)], [NCBI organismal classification, nbAnnotation: 1, score: 10, (38802, 1.2, 1132)], [COSTART, nbAnnotation: 1, score: 10, (40390, 1995, 1341)], [Physician Data Query, nbAnnotation: 1, score: 10, (45074, 2010_08_10, 1349)], [Common Terminology Criteria for Adverse Events, nbAnnotation: 1, score: 10, (40984, 4.02, 1415)], [Suggested Ontology for Pharmacogenomics, nbAnnotation: 1, score: 10, (39343, 2.1.2, 1061)], [Bone Dysplasia Ontology, nbAnnotation: 1, score: 10, (46301, 1.0, 1613)], [MDSS Mo, nbAnnotation: 1, score: 8, (40649, 1.0, 1395)], [Malaria Ontology, nbAnnotation: 1, score: 8, (44686, 1.22, 1311)]]

annotations = [AnnotationBean [

score = 16

concept = [localConceptId: 42297/B82005, conceptId: 16265242, localOntologyId: 42297, isTopLevel: 1, fullId: http://purl.bioontology.org/ontology/ICPC2P/B82005, preferredName: anaemia, definitions: [], synonyms: [Anemia, Anaemia, anemia], semanticTypes: [[id: 19756755, semanticType: T047, description: Disease or Syndrome]]]

context = [MGREP(true), from = 1, to = 6, [name: anemia, localConceptId: 42297/B82005, isPreferred: false],]

]

The annotation gives us the information about the ontology, local ID of the concept, a score to that term (local to an ontology), definition, synonyms and information about the semantic types.

CHAPTER 5

SEMANTIC MATCHMAKING

5.1 Introduction to Matchmaking:

Matchmaking is a process by which we calculate or compute the related results with respect to a certain entity. For example, if the entity in question was entity A, by applying a matchmaking algorithm, we would search and obtain all the entities and resources related to entity A. This list of results should be calculated based on the semantics of the entity A as well as the semantic annotations of the resulting resources. With respect to our domain, our purpose of matchmaking in this thesis is to obtain relevant publications to a particular patient (health record). We perform the matchmaking between the health record and paper publications to obtain relevant results.

In our research, we are dealing with not just matchmaking but semantic matchmaking. Semantic matchmaking is different from any other matchmaking in a way that in semantic matchmaking the results are obtained in light of a shared conceptualization for the knowledge domain at hand, which we call ontology.

The main goal of semantic matchmaking is to obtain relevant results. In order to obtain relevant results, we must ensure that the semantic annotations are proper and accurate. Also, the underlying ontology used should be appropriate, relevant and should provide us with all the possible outputs. One can also use more than a single Ontology to

obtain better results. In our matchmaking process we are using UMLS for obtaining the annotations. In addition to obtaining the results, another aspect of matchmaking is the ordering of the results. Once the matchmaking algorithm has been performed all the results obtained should have a ranking mechanism differentiating the most relevant information from the least relevant information.

5.2 Health Records Ontology:

This ontology contains all the patients information with all the results obtained after the annotation process. It consists of the following:

1. Name
2. ID (unique)
3. Age
4. Gender
5. Known disease
6. Medications
7. Symptoms
8. Annotations results for known disease (including synonyms)
9. Annotations results for medications (including synonyms)
10. Annotations results for symptoms (including synonyms)

Here is a sample of the health record ontology populated with annotations:

```
<!--
http://www.semanticweb.org/ontologies/2011/8/MedicalInoHealthRecords.ow
l#patient2 -->

    <owl:Thing rdf:about="#patient2">

        <Name>Robin Woods</Name>

        <Id>1235</Id>

        <Age>25</Age>

        <KnownDisease>Asthma</KnownDisease>

        <Medications>Aerobid</Medications><Medications>
Alvesco</Medications>

<symptoms>vomiting</symptoms>

<MedicationsSynonyms>
flunisolide</MedicationsSynonyms><MedicationsSynonyms>
Syntaris</MedicationsSynonyms><MedicationsSynonyms>
Flunisolide</MedicationsSynonyms><MedicationsSynonyms> Apo-
Flunisolide</MedicationsSynonyms><MedicationsSynonyms>
Flunisolide</MedicationsSynonyms><MedicationsSynonyms>
Rhinalar</MedicationsSynonyms><MedicationsSynonyms>
Nasarel</MedicationsSynonyms><MedicationsSynonyms> ratio-
Flunisolide</MedicationsSynonyms><MedicationsSynonyms> flunisolide
hemihydrate</MedicationsSynonyms><MedicationsSynonyms>
(6alpha</MedicationsSynonyms><MedicationsSynonyms>11beta</MedicationsSy
nonyms><MedicationsSynonyms>16alpha)-
isomer</MedicationsSynonyms><MedicationsSynonyms>
Nasalide</MedicationsSynonyms><MedicationsSynonyms> Apotex Brand of
Flunisolide</MedicationsSynonyms><MedicationsSynonyms> Elan Brand 1 of
Flunisolide</MedicationsSynonyms><MedicationsSynonyms> Roche Brand of
Flunisolide</MedicationsSynonyms><MedicationsSynonyms> Forest Brand of
Flunisolide</MedicationsSynonyms><MedicationsSynonyms> Ivax Brand of
Flunisolide</MedicationsSynonyms><MedicationsSynonyms> Dermapharm Brand
of Flunisolide</MedicationsSynonyms><MedicationsSynonyms>
flunisolide</MedicationsSynonyms><MedicationsSynonyms>
(6beta</MedicationsSynonyms><MedicationsSynonyms>11beta</MedicationsSyn
onyms><MedicationsSynonyms>16alpha)-
isomer</MedicationsSynonyms><MedicationsSynonyms>
Inhacort</MedicationsSynonyms><MedicationsSynonyms>
AeroBid</MedicationsSynonyms><MedicationsSynonyms> flunisolide
HFA</MedicationsSynonyms><MedicationsSynonyms>
flunisolide</MedicationsSynonyms><MedicationsSynonyms> 6 alpha-fluoro-
11 beta</MedicationsSynonyms><MedicationsSynonyms>16
alpha</MedicationsSynonyms><MedicationsSynonyms>21- tetrahydroxypregna-
1</MedicationsSynonyms><MedicationsSynonyms>4-diene-
3</MedicationsSynonyms><MedicationsSynonyms>20-dione cyclic
16</MedicationsSynonyms><MedicationsSynonyms> 17-acetal with
acetone</MedicationsSynonyms><MedicationsSynonyms> RS-
```

```
3999</MedicationsSynonyms><MedicationsSynonyms> 6 alpha-
fluorodihydroxy-16 alpha</MedicationsSynonyms><MedicationsSynonyms>17
alpha-isopropylidenedioxy-
1</MedicationsSynonyms><MedicationsSynonyms>4-pregnadiene-
3</MedicationsSynonyms><MedicationsSynonyms>20-
dione</MedicationsSynonyms><MedicationsSynonyms>
Alvesco</MedicationsSynonyms><MedicationsSynonyms> (R)-
11beta</MedicationsSynonyms><MedicationsSynonyms>16alpha</MedicationsSy
nonyms><MedicationsSynonyms>21-tetrahydroxypregna-
1</MedicationsSynonyms><MedicationsSynonyms>4-diene-
3</MedicationsSynonyms><MedicationsSynonyms>20-dione cyclic
16</MedicationsSynonyms><MedicationsSynonyms>17-acetal with
cyclohexanecarboxaldehyde</MedicationsSynonyms><MedicationsSynonyms>
21-isobutyrate</MedicationsSynonyms><MedicationsSynonyms>
Omnaris</MedicationsSynonyms><MedicationsSynonyms>Alvesco</MedicationsS
ynonyms><MedicationsSynonyms>Omnaris</MedicationsSynonyms>
<Synonyms>Bronchial hypersensitivity</Synonyms> <Synonyms>BHR -
Bronchial hyperreactivity</Synonyms>      <Synonyms>Airway
hyperreactivity</Synonyms><Synonyms>Bronchial
hyperreactivity</Synonyms><Synonyms>Hyperreactive airway
disease</Synonyms><Synonyms>Exercise-induced asthma</Synonyms>

<Gender>Male</Gender>

<SymptomsSynonyms>Vomiting</SymptomsSynonyms><SymptomsSynonyms>haematem
esis</SymptomsSynonyms><SymptomsSynonyms>Bilious
attack</SymptomsSynonyms><SymptomsSynonyms>throwing
up</SymptomsSynonyms>

    </owl:Thing>
```

5.3 Paper Publication Ontology:

This ontology contains all the paper publications information. A couple hundred (150) of the publications abstracts were downloaded from PubMed for testing purposes. Since the entire paper consists of figures, images, calculations etc. which results in excessive and/or unnecessary annotations, we choose to use only the abstracts for the annotations. This enabled us to get precise annotations and thus better results. Similar to

the health records; annotations were obtained to supply better results for the matchmaking. This ontology contains the following information:

1. Title
2. Abstract
3. Publication date
4. Authors names
5. Annotations for title
6. Annotations for abstract
7. Strength of the paper*

The Strength of the paper*: is calculated by processing the results of the annotations obtained. Considering the number of Top Level concepts found in the title and the abstract of any particular paper; the strength of that paper is calculated. The Top Level concept indicates the hierarchy of a particular concept in the paper in any given Ontology (UMLS, etc). Top Level indicates that a particular concept is in the Top Level; meaning it is a root in the ontology and not the leaves. If a paper has more Top Level concepts it indicates the greater strength of the paper compared to a one with no or lesser Top Level Concepts. For example, a word like “disease” appears in many ontologies, however, it is not the Top Level concept in most of them. On the other hand, specific medication like “Aerobid” is a Top Level concept in all the ontologies that it appears. A

paper with more Top Level concepts thus has a better strength than a paper with lesser Top Level concepts. The Formula for calculating the Strength of the paper is:

Strength of the Paper= (Number of Top Level Concepts/Total Number of Concepts)

The Strength of the paper is between zero (0) and one (1); where one (1) is the highest and zero(0) is the lowest. Here are a few examples showing the functionality of Strength Of Paper:

- **Paper 1:**

Top Level Concepts in Title and Abstract: 4

Total Concepts in Title and Abstract: 8

Strength of the paper: (Number of Top Level Concepts/Total Number of Concepts)

Strength of the paper: 4/8 = 0.5

- **Paper 2:**

Top Level Concepts in Title and Abstract: 10

Total Concepts in Title and Abstract: 10

Strength of the paper: (Number of Top Level Concepts/Total Number of Concepts)

Strength of the paper: 10/10= 1

5.4 Matchmaking Algorithm:

As seen in the Figure 13, the matchmaking algorithm starts from the two ontologies. One is for health records and the other one for PubMed Publications. Once the ontologies are populated, matchmaking is performed based on the data and annotations obtained. Here is the workflow indicating the flow of information and the matchmaking process:

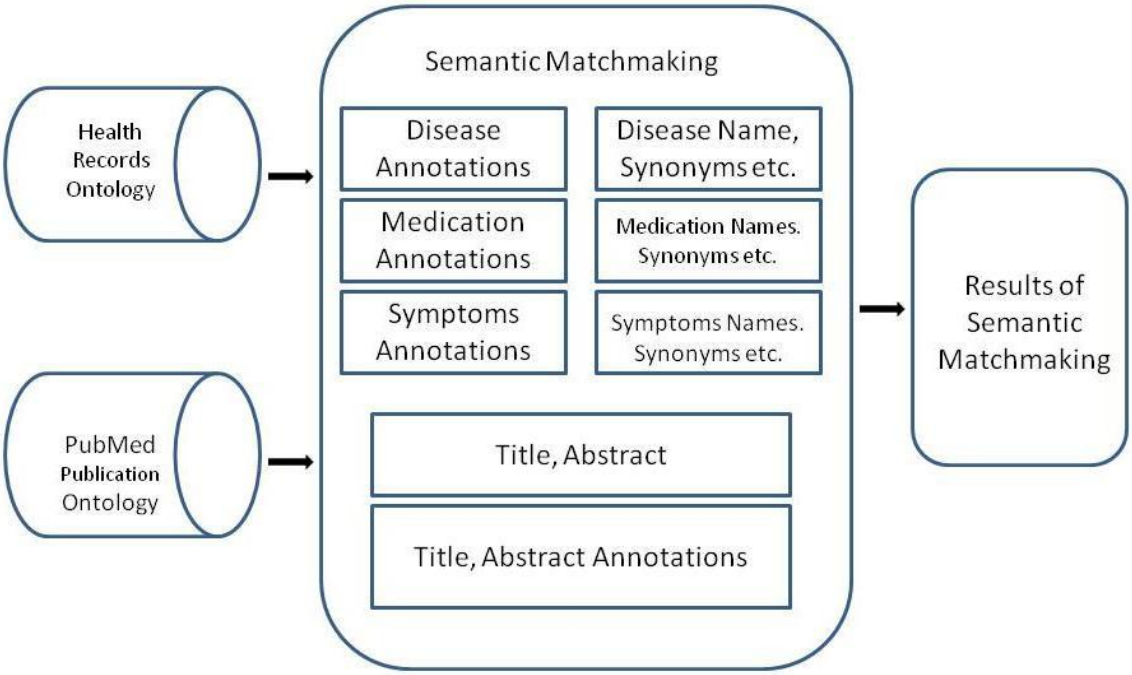


Figure 14: Matchmaking Workflow

The system performs matchmaking of the health records and publications based on the following information:

For the Heath Records:

1. Disease name
2. Annotations and synonyms of the disease names.
3. Medications
4. Annotations and synonyms of the medication names.
5. Symptoms
6. Annotations and synonyms of the medication names.

For the Publications:

1. Title of the paper
2. Abstract of the paper
3. Annotations of the title (Considering semantic hierarchy. i.e. strength of the concepts)
4. Annotations of the abstract (Considering semantic hierarchy i.e. strength of the concepts)

Here is a sample record and the resulting annotations which are used for matchmaking purposes:

Health Record:

1. Disease name: Asthma

   Annotations/ Synonyms:

   Bronchial hypersensitivity

   BHR - Bronchial hyperreactivity

   Airway hyperreactivity

   Bronchial hyperreactivity

   Hyperreactive airway disease

   Exercise-induced asthma

2. Symptoms: Vomiting

   Annotations/ Synonyms:

   Vomiting

   Haematemesis

   Bilious attack

   Throwing up

Our system now performs the matchmaking and provides the results accordingly. In the matchmaking process, the system not only performs the keyword matching, but also takes into consideration the semantic hierarchy, synonyms, annotations etc.

Once the matchmaking is done semantically, it goes above and beyond the keyword matches. This enables the user to get the relevant results regardless of the "word" or the "term" they enter. For example, a person has a symptom of vomiting, however, is unaware of the disease. Suppose that there is a new discovery about people having symptoms of Bilious attack and this discovery is found in one of the new research publications. However, if that person were to search a normal keyword search from their symptoms they would not be able to locate the paper, which discusses about the new discovery with symptoms of Bilious attack. However, with this system and with the underlying ontologies that person will get the results of this new discovery even if the paper does not have the word "vomiting" in it.

Here is pseudo-code of the matchmaking algorithm:

For a particular health record:

{

While there is a publication (do the following for title, abstract and their annotations):

(check the disease name, synonyms and annotations)

{

(If match found)

store in results

}

(check the symptoms, synonyms and annotations)

{

(If match found)

store in results

}

(check the medication, synonyms and annotations)

{

(If match found)

store in results

}

Sort results and assign scores based on where the match was found

Apply ranking algorithm

Print results

}

The system can be run in the following various ways to obtain the relevant information:

1. For a particular health record and obtaining all the results relevant to that particular record
2. For a cluster (more than one health record) of health records and obtaining all the results relevant to that particular cluster
3. For a particular disease and obtaining all the results relevant to that particular disease

4. For a cluster of disease names and obtaining all the results relevant to that cluster

5. For a particular medication and obtaining all the results relevant to that particular medication.

6. For a cluster of medications and obtaining all the results relevant to that cluster

7. For a particular symptom and obtaining all the results relevant to that particular symptom.

8. For a cluster of symptoms and obtaining all the results relevant to that cluster

CHAPTER 6

TESTING THE MATCHMAKING ALGORITHM

6.1 Test Case # 1:

Let's consider the motivating scenario # 1, which was mentioned in section 1.1.1. Our system gives the following results to Martha giving her links to all the related publications in the system. This enables her to view the update about the inhaler (medication) she is currently on and be able to switch to correct medication without losing any time, of course in consultation with her doctor.

Results:

```
Here are the results related to: Martha Jackson

Patient Record Number: 1288

Disease: Asthma


Here are the results related to : Martha Jackson
Patient Record Number:1288
Disease:Asthma

Rank is:7
Link is:
http://www.sciencedirect.com/science/article/pii/S0954611111002526

Rank is:7
Link is:
http://www.sciencedirect.com/science/article/pii/S1081120611004261

Rank is:8
Link is:
http://www.sciencedirect.com/science/article/pii/S1081120611004273

Rank is:7
```

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

Rank is:8
Link is:
http://www.sciencedirect.com/science/article/pii/S0091674911013145

Rank is:7
Link is: http://www.ncbi.nlm.nih.gov/pubmed/21720220

Rank is:7
Link is: http://chestjournal.chestpubs.org/content/139/2/311.long

Rank is:7
Link is: http://www.ncbi.nlm.nih.gov/pubmed/21658314

Rank is:10
Link is: http://www.ncbi.nlm.nih.gov/pubmed/19505390

Rank is:6
Link is: http://www.ncbi.nlm.nih.gov/pubmed/17509852

Rank is:6
Link is: http://www.ncbi.nlm.nih.gov/pubmed/21359665

Rank is:6
Link is: http://www.ncbi.nlm.nih.gov/pubmed/21573267

Rank is:10
Link is: http://www.aafp.org/online/en/home/publications/news/news-
now/health-of-the-public/20111010primatenemist.html
```

*(Ranking of the results will be described in more detail in the MS defence by Priya Wadhwa)

```
Here is a Snapshot of the User Interface results:
```

Name: Martha Jackson
Patient Record Number: 1288
Disease: Asthma

Here are all the related paper, please click the corresponding links:

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 presc
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
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.RESUL
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 ciclesor
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 t
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. Improvem
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
between ciclesonide and placebo were observed.CONCLUSIONS:In preschool children with recurrent wheeze and a positive Asthma predictive index, ciclesonide modestly reduces wheeze
rates and improves lung function. A large placebo response and unexpected selection of patients with mild disease may have affected outcomes, highlighting the heterogeneity of preschool whe

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

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 o

Figure 15: Snapshot of results

6.2 Test Case # 2:

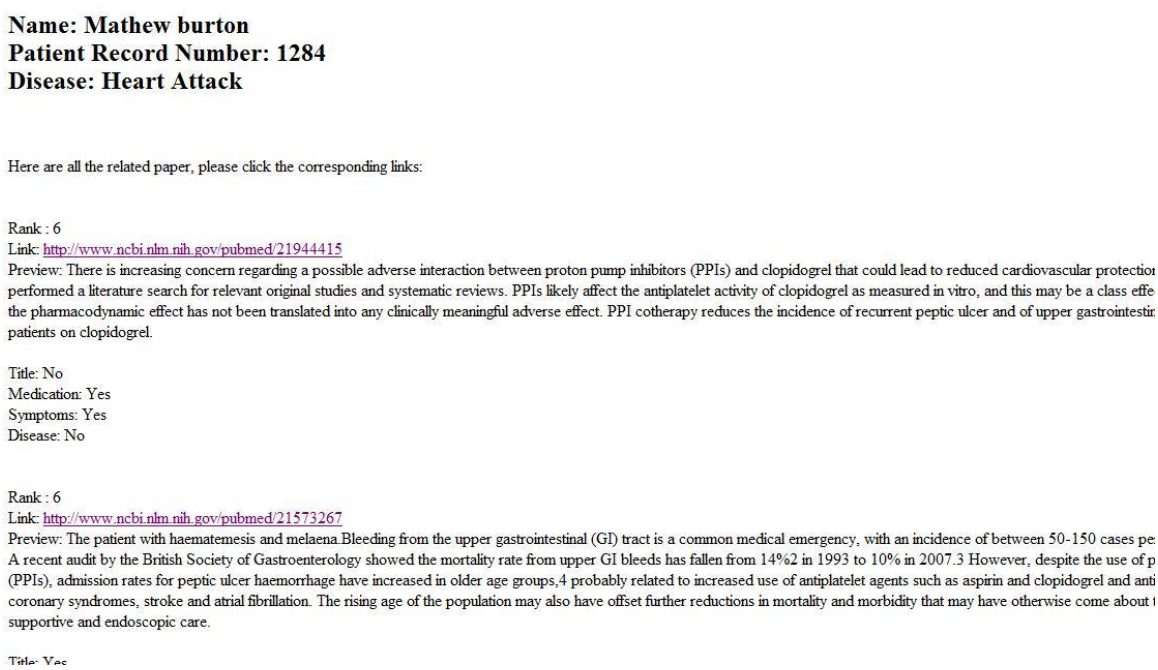
Let's consider motivating scenario # 2 mentioned in section 1.1.2. Our system performs the semantic matchmaking and thus provides the following results. It clearly identifies the semantic relationship between the two drugs and thus shows the paper indicating the affects of both drugs when taken together.

Results:

```
Here are the results related to: Mathew burton
Patient Record Number: 1284
Disease: Heart Attack

Rank is:6
Link is: http://www.ncbi.nlm.nih.gov/pubmed/21944415

Rank is:6
Link is: http://www.ncbi.nlm.nih.gov/pubmed/21573267

Rank is:5
Link is: http://www.ncbi.nlm.nih.gov/pubmed/21884023
```

```
Rank is:4
Link is: http://www.ncbi.nlm.nih.gov/pubmed/20729752

Rank is:7
Link is: http://www.ncbi.nlm.nih.gov/pubmed/22053219

Rank is:8
Link is: http://www.ncbi.nlm.nih.gov/pubmed/22053225
```

Here is a Snapshot of the User Interface results:



Name: Mathew burton
Patient Record Number: 1284
Disease: Heart Attack

Here are all the related paper, please click the corresponding links:

Rank : 6
Link: http://www.ncbi.nlm.nih.gov/pubmed/21944415
Preview: There is increasing concern regarding a possible adverse interaction between proton pump inhibitors (PPIs) and clopidogrel that could lead to reduced cardiovascular protectio performed a literature search for relevant original studies and systematic reviews. PPIs likely affect the antiplatelet activity of clopidogrel as measured in vitro, and this may be a class effe the pharmacodynamic effect has not been translated into any clinically meaningful adverse effect. PPI cotherapy reduces the incidence of recurrent peptic ulcer and of upper gastrointestir patients on clopidogrel.

Title: No
Medication: Yes
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 pe 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 p (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 anti 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 supportive and endoscopic care.

Title: Yes

Figure 16: Snapshot of results of test case 2

6.2 Comparison with Syntactic Matchmaking (PubMed):

6.2.1 Advance Ontological Search:

The semantic matchmaking enables the system to perform advance search based on the ontology concepts and hierarchy, which is not possible by a syntactic

matchmaking process. This enables the user to be able to discover and retrieve results that would not be found by a simple keyword search. This is an efficient way to discover hidden but important information.

6.2.2 Discovery of Medication side Effects:

Our system enables a user to not only get the related publications based on the disease they are suffering from, but also enables them to discover any side effects of the medications and drugs they are taking. Since the system is using UMLS with 173 ontologies for matchmaking, the results are not just limited to the disease's name. For example if a person is on some medication for a long time and if that drug or medication has some long term side effects; such publications should be displayed to the user. Our system does the same. It gives the users publications related to the effects of the drugs or medications they are on. For example; a query that was ran on a record suffering from breast cancer the following result was not only retrieved but also given a good rank:

- Rank is:8
- Link is: http://www.ncbi.nlm.nih.gov/pubmed/21993405

Title: Second cancer after radiotherapy 1981-2007

Our system allows the side effects of drugs to be discovered whether they appear directly or not in the paper since it checks the annotations, synonyms etc. This is something that cannot be achieved by syntactic matchmaking.

6.2.3 Extended Search via Profile:

Our system enables the user to retrieve publications that are not only related to his current disease or medications but also papers who may have some synonyms of the current medications or the papers which may have synonyms of the current symptoms. This search goes beyond the keyword search and retrieves the papers semantically related. For example, if we like to conduct matchmaking for someone with the symptoms of vomiting. Also, let's suppose that the patient does not suffer from any disease currently. In a syntactic search we will be able to receive all the results related to vomiting. However, with the help of semantic matchmaking the user will get results pertaining to vomiting including Haematemesis, Bilious attack and Throwing up etc. This enables the user to retrieve complete results regardless of the search term. When searched for symptoms "vomiting" we get the following results:

```
Rank is:6
Link is: http://www.ncbi.nlm.nih.gov/pubmed/12207199
Title: Vomiting

Rank is:6
Link is: http://www.ncbi.nlm.nih.gov/pubmed/21573267
Title: The patient with haematemesis and melaena

Rank is:6
Link is: http://www.ncbi.nlm.nih.gov/pubmed/21359665
```

```
Title: Gastric duplication cysts as a rare cause of
haematemesis
```

6.2.4 Knowledge Discovery without Specific Input :

Our system allows a user to discover the papers related them without having particular information about the disease they might be suffering from. Since the search can be done with either one of the parameters, the complete information is not mandatory. A person might search based on its symptoms without knowing the name of the disease or a person might just search without having any symptoms but on some particular medication. This enables them to retrieve and discover hidden knowledge. For example with our test case scenario number 2, the two drugs together had side effects which we were able to detect since we took the semantic relationship of both the drugs into consideration as indicated in the following diagram:

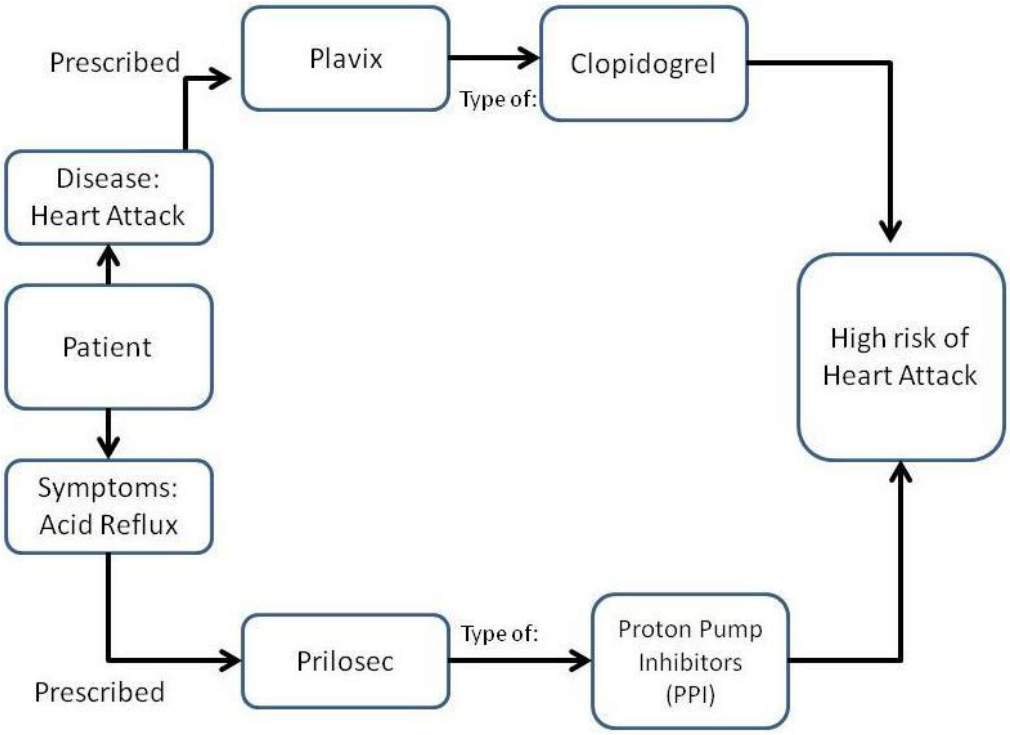


Figure 17: Test case scenario diagram

CHAPTER 7

SEMANTIC RANKING

7.1 Introduction to Ranking:

A ranking algorithm is defined to be a mechanism that calculates the relevance of all the elements in the result set and displays them to the user accordingly. There may be several ways of ranking; however, the ranking algorithm highly depends on the results and the query of the user.

There are several ranking algorithms currently being used by different systems. Various search engines use various algorithms to rank the documents/pages. One of the most popular search engines in today's world is Google. It uses a page rank algorithm which exploit's the link structure of the web to assign a rank to each page indicating that page's popularity on the current web.

7.2 Ranking Algorithm:

Once the results of matchmaking are obtained, there is a need of a mechanism of displaying these results to the user with a measure of relevance. The ranking algorithm proposed is a function of the following:

- ***Syntactic Measures:***

- **Publication Date Score (PDS):**

  A particular score is assigned based on the date of the publication of the paper. A higher score is assigned to recent publications.

- **Match on Disease Name (MDS):**

  A particular score is assigned if the name of the disease is present in the Title. Similarly another score is assigned if the name of the disease is present in the abstract etc.

- **Match on Medication Names (MMS):**

  A particular score is assigned if the name of the Medication is present in the Title. Similarly another score is assigned if the name of the Medication is present in the abstract etc.

- **Match on Symptoms Names (MSS):**

  A particular score is assigned if the name of the Symptoms is present in the Title. Similarly another score is assigned if the name of the Symptoms is present in the abstract etc.

- ***Semantic Measures:***

- **Match on Disease Annotations (MDA):**

  Beyond the keyword match in the syntactic measures part, a score is assigned if the match was found on the annotations of the disease

- **Match on Medications Annotations (MMA):**

  After the keyword match in the syntactic measures part, a score is assigned if the match was found on the annotations of the Medications

- **Match on Symptoms Annotations (MSA):**

  Beyond the keyword match in the syntactic measures part, a score is assigned if the match was found on the annotations of the symptoms

- **Strength of the Paper (SPS):**

  As described in the above section, the strength of the paper is calculated based on the top Level Concepts found in the paper. This is a semantic measure that helps us assign a better score to the result. Please refer to section (above) for detailed functioning of strength of paper.

- ***Calculating the Overall Rank:***

  Here is the formula used to calculate the overall rank:

  Syntactic Score: (PDS)+ (MDS)+ (MMS)+ (MSS)

  Semantic Score: (MDA)+ (MMA)+ (MSA)+ (SPS)

  Overall Rank: Syntactic Score + Semantic Score

The range of the score is between 3 and 12, where 3 is the lowest possible score and 12 is the highest possible score.

CHAPTER 8

SYSTEM WORKFLOW EXAMPLE

This workflow example illustrates the complete lifecycle of a record in our system. It shows what steps are precisely taken and how the results are calculated. It elaborates on the input and output at each state.

Example: Example on a health record (patient) suffering from Asthma

Workflow:

Here are the steps performed:

1. Start with Health Record (XML)
2. Parse the health record to generate a profile
3. Populate the Ontology with the health records
4. Run the NCBO annotator to get the annotations
    a. Get the annotations for Disease name
    b. Get the annotations for Medication names
    c. Get the annotations for Symptoms
5. Update the Health records Ontology with the annotations

For the Publications Ontology:

6. We begin with publications downloaded from PubMed

7. Populate the Ontology with the Publication Information

8. Run the NCBO annotator to get the annotations

9. Update the Publications with the annotations

10. Once both the Ontologies are populated; we can begin the matchmaking and ranking algorithm

11. Run the matchmaking and ranking algorithms and obtain the results.

Step 1: We begin with sample health records; these health records are created for testing purposes.

Sample Health Record (XML):

```
<Patient>

<Name>Robin Woods</Name>
<Address>1563 South Milton st</Address>
<City>Tuscon</City>
<State>AZ</State>
<Zip>92009</Zip>
<Country>United States</Country>
<Id>1235</Id>
<Age>25</Age>
<KnownDisease>Asthma</KnownDisease>
<Medications>Aerobid, Alvesco</Medications>
<Gender>Male</Gender>
<symptoms>vomiting</symptoms>
<PrimaryPhysician> Dr Smith</ PrimaryPhysician>
<PhysicianId>dc1247</PhysicianId>
<PrimaryPharmacy>Walgreens</PrimaryPharmacy>
<PrimaryPharmacyId>247Phar</PrimaryPharmacyId>

 </Patient>
```

Step 2: Once the health record is parsed, we get the following profile:

```
Patient Details:
      Name: Robin Woods
      symptoms: vomiting
      Id: 1235
      Age: 25
      Gender: Male
      Known Disease: Asthma
      Medications: Aerobid, Alvesco
```

Step 3: We can now populate the ontology with the health record(s):

```
<!--
http://www.semanticweb.org/ontologies/2011/8/MedicalInoHealthRecords.ow
l#patient2 -->

   <owl:Thing rdf:about="#patient2">

   <Name>Robin Woods</Name>
   <Id>1235</Id>
   <Age>25</Age>
   <KnownDisease>Asthma</KnownDisease>
   <Medications>Aerobid</Medications><Medications>Alvesco</Medications>
   <Gender>Male</Gender>
   <symptoms>vomiting</symptoms>
   </owl:Thing>
```

Step 4: Getting the annotations; here is a sample output file of the annotations results obtained for Asthma. Similarly, we get the annotations for Medication, Symptoms and the Publications as well.

```
ObaResultBean [
ResultBean [
       resultID = OBA_RESULT_0961
       statistics = [(MAPPING, 0) , (CLOSURE, 0) , (MGREP, 35) ]
       parameters = [longestOnly = false, wholeWordOnly = true,
filterNumber = true, withSynonyms = true, withContext = true,
ontologiesToExpand = [], ontologiesToKeepInResult = [],
isVirtualOntologyId = false, semanticTypes = [], levelMax = 0,
mappingTypes = [null], stopWords = [], withDefaultStopWords = true,
isStopWordsCaseSenstive = false, text to annotate = Asthma]
]
       ontologies = [[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], ]
]]
]
```

a. Get the annotations for Disease name. The above annotation file is parsed to obtain the relevant information for a disease name.

b. Get the annotations for Medication names

   Similar to Step 4 (a), in this step we obtain and parse annotations for Medication Names

c. Get the annotations for Symptoms

   Similar to Step 4 (a), in this step we obtain and parse annotations for Symptoms Names

Step 5: Update the Health records with the annotations (it includes the name of the disease, its annotations, the symptoms and its annotations, the medications and its annotations respectively):

```
<!--
http://www.semanticweb.org/ontologies/2011/8/MedicalInoHealthRecords.ow
l#patient2 -->

    <owl:Thing rdf:about="#patient2">
            <symptoms>vomiting</symptoms>
        <Name>Robin Woods</Name>
        <Id>1235</Id>
        <Age>25</Age>
        <KnownDisease>Asthma</KnownDisease>
        <Medications>Aerobid</Medications><Medications>
Alvesco</Medications>
```

<MedicationsSynonyms>
flunisolide</MedicationsSynonyms><MedicationsSynonyms>
Syntaris</MedicationsSynonyms><MedicationsSynonyms>
Flunisolide</MedicationsSynonyms><MedicationsSynonyms> Apo-
Flunisolide</MedicationsSynonyms><MedicationsSynonyms>
Flunisolide</MedicationsSynonyms><MedicationsSynonyms>
Rhinalar</MedicationsSynonyms><MedicationsSynonyms>
Nasarel</MedicationsSynonyms><MedicationsSynonyms> ratio-
Flunisolide</MedicationsSynonyms><MedicationsSynonyms> flunisolide
hemihydrate</MedicationsSynonyms><MedicationsSynonyms>
(6alpha</MedicationsSynonyms><MedicationsSynonyms>11beta</MedicationsSy
nonyms><MedicationsSynonyms>16alpha)-
isomer</MedicationsSynonyms><MedicationsSynonyms>
Nasalide</MedicationsSynonyms><MedicationsSynonyms> Apotex Brand of
Flunisolide</MedicationsSynonyms><MedicationsSynonyms> Elan Brand 1 of
Flunisolide</MedicationsSynonyms><MedicationsSynonyms> Roche Brand of
Flunisolide</MedicationsSynonyms><MedicationsSynonyms> Forest Brand of
Flunisolide</MedicationsSynonyms><MedicationsSynonyms> Ivax Brand of
Flunisolide</MedicationsSynonyms><MedicationsSynonyms> Dermapharm Brand
of Flunisolide</MedicationsSynonyms><MedicationsSynonyms>
flunisolide</MedicationsSynonyms><MedicationsSynonyms>
(6beta</MedicationsSynonyms><MedicationsSynonyms>11beta</MedicationsSyn
onyms><MedicationsSynonyms>16alpha)-
isomer</MedicationsSynonyms><MedicationsSynonyms>
Inhacort</MedicationsSynonyms><MedicationsSynonyms>
AeroBid</MedicationsSynonyms><MedicationsSynonyms> flunisolide
HFA</MedicationsSynonyms><MedicationsSynonyms>
flunisolide</MedicationsSynonyms><MedicationsSynonyms> 6 alpha-fluoro-
11 beta</MedicationsSynonyms><MedicationsSynonyms>16
alpha</MedicationsSynonyms><MedicationsSynonyms>21- tetrahydroxypregna-
1</MedicationsSynonyms><MedicationsSynonyms>4-diene-
3</MedicationsSynonyms><MedicationsSynonyms>20-dione cyclic
16</MedicationsSynonyms><MedicationsSynonyms> 17-acetal with
acetone</MedicationsSynonyms><MedicationsSynonyms> RS-
3999</MedicationsSynonyms><MedicationsSynonyms> 6 alpha-
fluorodihydroxy-16 alpha</MedicationsSynonyms><MedicationsSynonyms>17
alpha-isopropylidenedioxy-
1</MedicationsSynonyms><MedicationsSynonyms>4-pregnadiene-
3</MedicationsSynonyms><MedicationsSynonyms>20-
dione</MedicationsSynonyms><MedicationsSynonyms>
Alvesco</MedicationsSynonyms><MedicationsSynonyms> (R)-
11beta</MedicationsSynonyms><MedicationsSynonyms>16alpha</MedicationsSy
nonyms><MedicationsSynonyms>21-tetrahydroxypregna-
1</MedicationsSynonyms><MedicationsSynonyms>4-diene-
3</MedicationsSynonyms><MedicationsSynonyms>20-dione cyclic
16</MedicationsSynonyms><MedicationsSynonyms>17-acetal with
cyclohexanecarboxaldehyde</MedicationsSynonyms><MedicationsSynonyms>
21-isobutyrate</MedicationsSynonyms><MedicationsSynonyms>
Omnaris</MedicationsSynonyms><MedicationsSynonyms>Alvesco</MedicationsS
ynonyms><MedicationsSynonyms>Omnaris</MedicationsSynonyms>
<Synonyms>Bronchial hypersensitivity</Synonyms> <Synonyms>BHR -
Bronchial hyperreactivity</Synonyms> <Synonyms>Airway
hyperreactivity</Synonyms><Synonyms>Bronchial
hyperreactivity</Synonyms><Synonyms>Hyperreactive airway
disease</Synonyms><Synonyms>Exercise-induced asthma</Synonyms>
<Gender>Male</Gender>

```
      <SymptomsSynonyms>Vomiting</SymptomsSynonyms><SymptomsSynonyms>ha
ematemesis</SymptomsSynonyms><SymptomsSynonyms>Bilious
attack</SymptomsSynonyms><SymptomsSynonyms>throwing
up</SymptomsSynonyms>

    </owl:Thing>
```

Step 6: We begin with publications downloaded from PubMed, currently 150 different publications (abstracts) were downloaded for testing purposes.

Step 7: Populate the Ontology with the Publication Information:

```
<!--
http://www.semanticweb.org/ontologies/2011/8/MedicalPapers.owl#paper11
-->

    <medicalpaper rdf:about="#paper11">
        <rdf:type rdf:resource="&owl;Thing"/>
      <title>Asthma diagnosis and treatment: Filling in the information
      gaps</title>

        <abstr> 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. A
      major morbidity in asthmatic patients occurs with exacerbations
      and in patients with severe disease. Novel approaches to
      treatment for these conditions will be an important advance to
      reduce the morbidity associated with asthma.</abstr>

      <url>http://www.sciencedirect.com/science/article/pii/S0091674911
      013145</url>
      <publishing_date>2011</publishing_date>
      <author>Busse WW.</author>
    </medicalpaper>
```

Step 8: Run the NCBO annotator to get the annotations. Here is a sample (partial) output file received from annotating the following title:

Title: Asthma diagnosis and treatment: Filling in the information gaps

Annotations obtained:


```
ObaResultBean [
ResultBean [
        resultID = OBA_RESULT_6dd6
        statistics = [(MAPPING, 0) , (CLOSURE, 0) , (MGREP, 96) ]
        parameters = [longestOnly = false, wholeWordOnly = true,
filterNumber = true, withSynonyms = true, withContext = true,
ontologiesToExpand = [], ontologiesToKeepInResult = [],
isVirtualOntologyId = false, semanticTypes = [], levelMax = 0,
mappingTypes = [null], stopWords = [], withDefaultStopWords = true,
isStopWordsCaseSenstive = false, text to annotate = Asthma diagnosis
and treatment: Filling in the information gaps]
]
        ontologies = [[NCI Thesaurus, nbAnnotation: 11, score: 96,
(45400, 11.01e, 1032)], [SNOMED Clinical Terms, nbAnnotation: 7, score:
88, (46116, 2010_07_31, 1353)], [Logical Observation Identifier Names
and Codes, nbAnnotation: 8, score: 76, (44774, 232, 1350)], [National
Drug File, nbAnnotation: 3, score: 50, (40402, 2008_03_11, 1352)],
[Medical Subject Headings, nbAnnotation: 5, score: 46, (44776,
2011_2010_08_30, 1351)], [Galen, nbAnnotation: 4, score: 40, (4525,
1.1, 1055)], [MedDRA, nbAnnotation: 2, score: 40, (42280, 12.0, 1422)],
[Health Level Seven, nbAnnotation: 4, score: 40, (42545, 0230, 1343)],
[ICPC-2 PLUS, nbAnnotation: 2, score: 36, (42297, 2005, 1429)], [PHARE,
nbAnnotation: 4, score: 32, (45138, 110114, 1550)], [RadLex,
nbAnnotation: 3, score: 30, (45589, 3.4, 1057)], [eVOC (Expressed
Sequence Annotation for Humans), nbAnnotation: 3, score: 30, (44302,
2.9, 1013)], [Ontology for General Medical Science, nbAnnotation: 2,
score: 20, (45302, 2011-02-21, 1414)], [PMA 2010, nbAnnotation: 2,
score: 20, (44666, 0.9.1, 1497)], [Family Health History Ontology,
nbAnnotation: 2, score: 20, (38631, 1.0, 1126)], [Brucellosis Ontology,
nbAnnotation: 2, score: 20, (44723, 1.0.67, 1537)], [CRISP Thesaurus,
2006, nbAnnotation: 2, score: 20, (44432, 2006, 1526)], [Experimental
Factor Ontology, nbAnnotation: 2, score: 20, (45659, 2.12.1, 1136)],
[Rat Strain Ontology, nbAnnotation: 1, score: 10, (45442, 3.0, 1150)],
[Human Phenotype Ontology, nbAnnotation: 1, score: 10, (45774, unknown,
1125)], [Neomark Oral Cancer-Centred Ontology, nbAnnotation: 1, score:
10, (42835, 3.1, 1501)], [MedlinePlus Health Topics, nbAnnotation: 1,
score: 10, (40397, 20080614, 1347)], [Event (INOH pathway ontology),
nbAnnotation: 1, score: 10, (45404, unknown, 1011)], [International
Classification of Primary Care, nbAnnotation: 1, score: 10, (40393,
1993, 1344)], [Host Pathogen Interactions Ontology, nbAnnotation: 1,
score: 10, (45230, 1.0, 1569)], [African Traditional Medicine,
nbAnnotation: 1, score: 10, (40223, 1.101, 1099)], [MGED Ontology,
nbAnnotation: 1, score: 10, (38801, 1.3.1.1, 1131)], [COSTART,
nbAnnotation: 1, score: 10, (40390, 1995, 1341)], [Read Codes, Clinical
Terms Version 3 (CTV3) , nbAnnotation: 1, score: 10, (42295, 1999,
1427)], [Ontology for Biomedical Investigations, nbAnnotation: 1,
```

```
score: 10, (45713, 2011-04-20, 1123)], [Situation-Based Access Control,
nbAnnotation: 1, score: 10, (45298, 1.3, 1237)], [Suggested Ontology
for Pharmacogenomics, nbAnnotation: 1, score: 10, (39343, 2.1.2,
1061)], [Malaria Ontology, nbAnnotation: 1, score: 10, (44686, 1.22,
1311)], [Loggerhead nesting, nbAnnotation: 1, score: 10, (44831,
unknown, 1024)], [WHO Adverse Reaction Terminology, nbAnnotation: 1,
score: 10, (40404, 1997, 1354)], [ICD10, nbAnnotation: 1, score: 10,
(44103, 1998 , 1516)], [VANDF, nbAnnotation: 1, score: 10, (44452,
2010_01_25, 1527)], [Human disease, nbAnnotation: 1, score: 10, (45769,
unknown, 1009)], [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)], [ICD10CM,
nbAnnotation: 1, score: 10, (44860, 2010_03, 1553)], [Bone Dysplasia
Ontology, nbAnnotation: 1, score: 10, (46301, 1.0, 1613)],
[Translational Medicine Ontology, nbAnnotation: 1, score: 10, (45369,
1.0, 1461)], [Molecule role (INOH Protein name/family name ontology),
nbAnnotation: 1, score: 8, (45784, unknown, 1029)], [NIFSTD,
nbAnnotation: 1, score: 8, (45355, 2.2 - December 20, 2010, 1084)],
[Physician Data Query, nbAnnotation: 1, score: 8, (45074, 2010_08_10,
1349)]]
        annotations = [AnnotationBean [
                score = 20
                concept = [localConceptId: 40402/C290507, conceptId:
14138290, localOntologyId: 40402, isTopLevel: 0, fullId:
http://purl.bioontology.org/ontology/NDFRT/C290507, preferredName:
INFORMATION, definitions: [], synonyms: [INFORMATION], semanticTypes:
[[id: 17572510, semanticType: T999, description: NCBO BioPortal
concept]]]
                context = [MGREP(true), from = 48, to = 58, [name:
INFORMATION, localConceptId: 40402/C290507, isPreferred: false], ]
], AnnotationBean [
                score = 20
                concept = [localConceptId: 40402/C290507, conceptId:
14138290, localOntologyId: 40402, isTopLevel: 0, fullId:
http://purl.bioontology.org/ontology/NDFRT/C290507, preferredName:
INFORMATION, definitions: [], synonyms: [INFORMATION], semanticTypes:
[[id: 17572510, semanticType: T999, description: NCBO BioPortal
concept]]]
                context = [MGREP(true), from = 48, to = 58, [name:
INFORMATION, localConceptId: 40402/C290507, isPreferred: true], ]
]]
```

Step 9: We parse the relevant information from the file obtained in Step 8 and Update the Publications with the annotations:

```
<!--
http://www.semanticweb.org/ontologies/2011/8/MedicalPapers.owl#paper11
-->

    <medicalpaper rdf:about="#paper11">
        <rdf:type rdf:resource="&owl;Thing"/>
```

<abstr>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. A major morbidity in asthmatic patients occurs with exacerbations and in patients with severe disease. Novel approaches to treatment for these conditions will be an important advance to reduce the morbidity associated with asthma.</abstr>

<annotation>Asthma</annotation><annotation>disease</annotation><annotation>diagnosis</annotation><annotation>Bronchial hyperresponsiveness</annotation><annotation>Asthmatic</annotation><annotation>Approaches</annotation><annotation>guideline</annotation><annotation>use</annotation><annotation>treatment</annotation><annotation>treatment</annotation><annotation>management</annotation><annotation>Patients</annotation><annotation>INFORMATION</annotation><annotation>severe</annotation><annotation>morbidity</annotation><annotation>new</annotation><annotation>CARE</annotation><annotation>Obstruction</annotation><annotation>associated</annotation><annotation>associated with </annotation><annotation>Reversibility</annotation><annotation>Clinical </annotation><annotation>Individual</annotation><annotation>Clinical</annotation><annotation>to reduce </annotation><annotation>Bronchial</annotation><annotation>Guideline</annotation><annotation>Antibiotics</annotation><annotation>Obstruction</annotation><annotation>Individual</annotation>
<url>
http://www.sciencedirect.com/science/article/pii/S0091674911013145</url>
<title>Asthma diagnosis and treatment: Filling in the information gaps</title>
<publishing_date>2011</publishing_date>
<author>Busse WW.</author>
</medicalpaper>

Step 10: Once both the Ontologies are populated; we can begin the matchmaking and ranking algorithm

Step 11: We can now run the Matchmaking and Ranking algorithms; here are the results obtained:

Here are the results related to : Robin Woods
Patient Record Number:1235
Disease:Asthma

Rank is:7
Link is:
http://www.sciencedirect.com/science/article/pii/S0954611111002526

Rank is:7
Link is:
http://www.sciencedirect.com/science/article/pii/S1081120611004261

Rank is:8
Link is:
http://www.sciencedirect.com/science/article/pii/S1081120611004273

Rank is:7
Link is:
http://www.sciencedirect.com/science/article/pii/S1081120611003929

Rank is:8
Link is:
http://www.sciencedirect.com/science/article/pii/S0091674911013145

Rank is:7
Link is: http://www.ncbi.nlm.nih.gov/pubmed/21720220

Rank is:7
Link is: http://chestjournal.chestpubs.org/content/139/2/311.long

Rank is:7
Link is: http://www.ncbi.nlm.nih.gov/pubmed/21658314

Rank is:10
Link is: http://www.ncbi.nlm.nih.gov/pubmed/19505390

Rank is:6
Link is: http://www.ncbi.nlm.nih.gov/pubmed/17509852

Rank is:6
Link is: http://www.ncbi.nlm.nih.gov/pubmed/21359665

Rank is:6
Link is: http://www.ncbi.nlm.nih.gov/pubmed/21573267

Rank is:9
Link is: http://www.ncbi.nlm.nih.gov/pubmed/12207199

CHAPTER 9

PRELIMINARY EVALUATION

In order to evaluate the functionality of our system, we did an evaluation of our results vs. the results of PubMed. PubMed provides a user interface to search for publications related to the terms entered. We use the same interface to enter the disease name, symptoms or medications and retrieve results. On the other hand, we use our system and find related papers to a particular record (patient) who is suffering from the same disease, symptoms and takes the same medications. This allowed us to do a comparison on both the results obtained and conclude the results. We used our test scenario number 2 that was explained in the above section for the evaluation purposes.

Here is snapshot of the results obtained from PubMed:

User Profile:

Name: Mathew Burton

Known Disease: Heart Attack

Symptoms: Arm pain, Acidity

Medications: Prilosec, Plavix, Alprenolol

Query 1:

PubMed Input: Heart Attack, Arm pain, Acidity, Prilosec, Plavix, Alprenolol

PubMed Output: No items found.

Query 2: Prilosec, Plavix, Alprenolol

PubMed Output: No items found.

Query 3: Heart Attack, Arm pain, Acidity

PubMed Output: No items found.

Query 4: Heart Attack

PubMed Output:

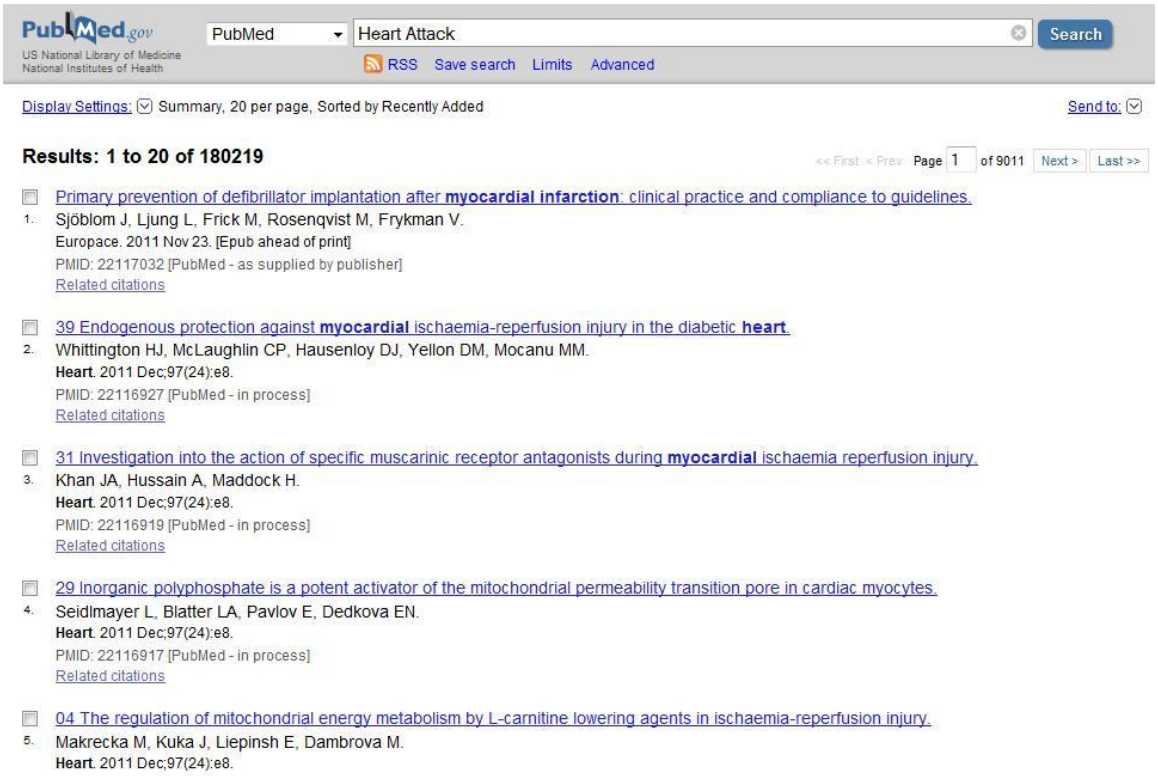


Figure 18: PubMed Results

As seen in the above test queries, PubMed only gives results when one term is entered at a time. When we tried entering all the keywords in a given profile, no results were obtained. However, when we entered one term “Heart attack”, we received several results. In addition, the results are based on syntactic matches on the term “heart attack”, thus the additional relevant information is not obtained, which includes information about medications, side effects, combined effect of drugs etc.

Here is snapshot of the results obtained from our system:

User Profile:

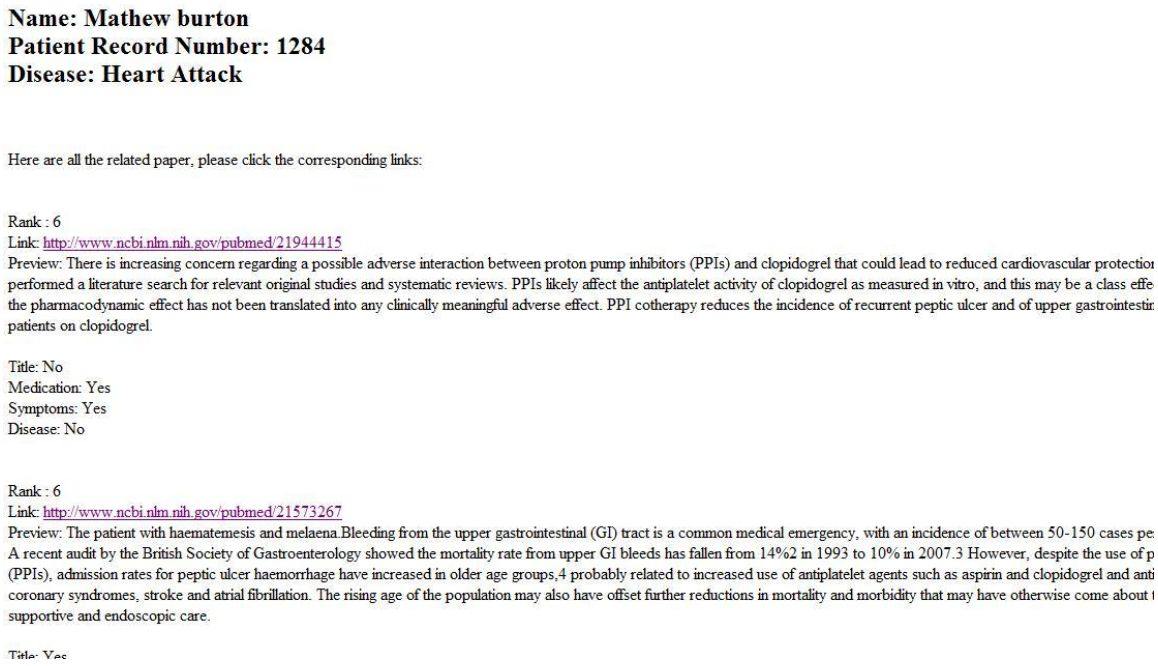
Name: Mathew Burton

Known Disease: Heart Attack

Symptoms: Arm pain, Acidity

Medications: Prilosec, Plavix, Alprenolol

Results:



Name: Mathew burton
Patient Record Number: 1284
Disease: Heart Attack

Here are all the related paper, please click the corresponding links:

Rank : 6
Link: http://www.ncbi.nlm.nih.gov/pubmed/21944415
Preview: There is increasing concern regarding a possible adverse interaction between proton pump inhibitors (PPIs) and clopidogrel that could lead to reduced cardiovascular protectio
performed a literature search for relevant original studies and systematic reviews. PPIs likely affect the antiplatelet activity of clopidogrel as measured in vitro, and this may be a class effe
the pharmacodynamic effect has not been translated into any clinically meaningful adverse effect. PPI cotherapy reduces the incidence of recurrent peptic ulcer and of upper gastrointestir
patients on clopidogrel.

Title: No
Medication: Yes
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 pe
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 p
(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 anti
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
supportive and endoscopic care.

Title: Yes

Figure 19: Results Snapshot

We can see that our system, gave the results of papers discussing the combined effects of both the drugs Prilosec and Plivax together, while there was no implicit information given. Our system was able to discover the semantic relationship between the two drugs and thus showed the related papers in the result which were not found in the PubMed results.

From the above example it is evident that our system performs better than the searches done at PubMed. Our system not only allows us to search based on the profile

and not one keyword, but it also takes the semantic relationship between the provided information into consideration. Thus in the above example, we did not only get results related to Heart attack, but also results related to symptoms, medications, side effects of medication, combined effect of two medications etc.

CHAPTER 10

CONCLUSION AND FUTURE WORKS

The amount of knowledge in the medical domain is growing exponentially. With this growth, it is becoming a very hard for physicians or the patients to keep track of all the new discoveries. Our system addresses this issue and makes this knowledge discovery easier.

Our system performs semantic matchmaking for knowledge discovery and then semantic ranking to rank the results for a particular patient. This can be used by physicians or by patients to discover resources related to their Personal Health Record. Since the system performs semantic matchmaking, the results are more precise and accurate. As seen in the above two motivating examples; our system enables the user to discover papers/knowledge that would not have been possible to discover via syntactic matchmaking.

Future works on this system might include taking geographic location, age and gender into consideration when ranking the results for any particular patient. Geographic location may affect the results as some diseases are more common in some countries than other. In addition, age and gender may also affect the results as some diseases and publications are for a particular age group or gender.

Also, an extended evaluation in form of usability studies can be done with the help of doctors and physicians to identify the accuracy of the results.

REFERENCES

1. Tommaso Di Noia, Eugenio Di Sciascio, Francesco M. Donini. "A Non-Monotonic Approach to Semantic Matchmaking and Request Refinement in E-Marketplaces" International Journal of Electronic Commerce, 2008.

2. Jonquet, Clement. "Semantic Annotations of BioMedical Data". http://www.slideshare.net/jonquet/semantic-annotation-of-biomedical-data-7281656. March 2011.

3. Good, Benjamin. "NCBO Annotator versus MetaMap on GO concept detection". http://i9606.blogspot.com/2010/12/ncbo-annotator-versus-metamap-on-go.html. December 02, 2010.

4. Jonquet, Clement. Musen, Mark A and Shah, Nigam. "A System for Ontology-Based Annotation of Biomedical Data". Proceedings of the 5th international workshop on Data Integration in the Life Sciences. "http://bmir.stanford.edu/file_asset/index.php/1316/Article-DILS08_Jonquet_Musen_Shah_published.pdf." 2008.

5. Patricia L. Whetzel, Nigam H. Shah, Natalya F. Noy, Clement Jonquet, Adrien Coulet, Nicholas Griffith, Cherie Youn, Michael Dorf and Mark A. "Ontology Web Services for Semantic Applications". Stanford University, Stanford CA, USA. "http://www2.lirmm.fr/IC//Supports/FMIN113-ProjetTutore/2010_11/ExemplesPosters/CJ2.pdf." 2008.

6. U.S. National Library of Medicine. "MetaMap". "http://www.nlm.nih.gov/research/umls/implementation_resources/metamap.html" . September 2011.

7. Alan R. Aronson, PhD and François M. Lang, MSE. "The Evolution of MetaMap, a Concept Search Program for Biomedical Text". Lister Hill National Center for Biomedical Communications. "http://www.lhncbc.nlm.nih.gov/lhc/docs/published/2009/pub2009041.pdf". 2009.

8. Aronson, Alan R. Lang, François-Michel. "An overview of MetaMap: historical perspective and recent advances" J Am Med Inform Assoc (JAMIA). "http://www.lhncbc.nlm.nih.gov/lhc/docs/published/2010/pub2010033.pdf". 2010

9. Google Health Samples. "http://code.google.com/p/googlehealthsamples/source/browse/trunk/CCR_sa

mples/". Retrieved on September 2011.

10. PubMed.com "http://www.ncbi.nlm.nih.gov/pubmed/" 2011.

11. PubMed Quick Start. U.S. National Library of Medicine National Institutes of Health. "http://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.PubMed_Quick_Start". Retrieved on Aug 2011.

12. PubMed FAQS. U.S. National Library of Medicine National Institutes of Health. "http://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.FAQs". Retrieved on Aug 2011.

13. UMLS. Open Clinical Knowledge Management for Medical Care. "http://www.openclinical.org/medTermUmls.html " Retreived on Aug 2011.

14. UMLS. Wikepedia. "http://en.wikipedia.org/wiki/Unified_Medical_Language_System" Retrieved on Aug 2011.

15. Chintan Patel, Sharib Khan, and Karthik Gomadam. "TrialX: Using semantic technologies to match patients to relevant clinical trials based on their Personal Health Records". In Proceedings of the 8th International Semantic Web Conference, 2009.

16. Google Health Vs Microsoft Health Vault. User Centric, Inc. "http://www.usercentric.com/publications/2009/02/02/google-health-vs-microsoft-healthvault-consumers-compare-online-personal-hea" February 2009.

17. Microsoft Health Vault. Microsoft. http://www.microsoft.com/en-us/healthvault/" retrieved on Sep 2011.

18. Life Record Personal Medical Record System. Life Records, Inc. http://www.myliferecord.com/ retrieved on Sep 2011.

19. User Manual Life Record Personal Medical Record System. Life Records, Inc. http://download.liferecord.com/mlruserguide.pdf retrieved on Sep 2011.

20. Page Rank. Wikepedia http://en.wikipedia.org/wiki/PageRank. retrieved on Sep 2011.

21. About Google Health. Google Health. http://www.google.com/intl/en-US/health/about/ retrieved on Sep 2011.

22. OBA Service Workflow. BioOntology Wiki. http://www.bioontology.org/wiki/index.php/File:OBA_service_workflow.png " retrieved on Sep 2011.

23. Laura Plaza, Alber Diaz. “Retrieval of Similar Electronic Health Records Using UMLS Concept Graphs”. Proceedings of the Natural language processing and information systems, and 15th international conference on Applications of natural language to information systems, 2010.

24. Barbara Hayes, William Aspray. “Fighting Diabetes with Information: Where Social Informatics Meets Health Informatics”. iConference, 2010.

25. Semantic Network, U.S. National Library of Medicine National Institutes of Health. “http://www.ncbi.nlm.nih.gov/books/NBK9679/”. Retrieved on Nov 2011.

26. 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.