Semantic Discovery and Retrieval of Relevant Medical Knowledge
A. Asmita Rahman ¹, B. Budak Arpinar ¹, and C. Singaram Sundar ¹
¹Computer Science Department, University of Georgia, Athens, Georgia, USA 30602

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 an ordinary person to keep up with all the new discoveries and updates in this domain. This paper describes an approach to semantically discover and retrieve relevant medical data/information for respective health records (people). This system comprises of sample Electronic Health Records (EHRs) and Health Publications from PubMed. Our system implements a semantic matchmaking algorithm to find the relevant publications in PubMed for any particular health record (profile) using BioPortal Ontologies and UMLS. It then 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 the relevant information better than syntactic searches.

Keywords: Semantic Matchmaking, Matchmaking algorithm, Knowledge Discovery, Electronic Health Records, Health Publications, Ontology.

1 Introduction

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 are being introduced, many old medications are being replaced, several new discoveries are 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 Motivating Scenario

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 previous histories of heart attacks. In March 2009, a study appeared in the Journal of American Medical Association, which indicated that a combination of drugs Clopidogrel (Plavix is the brand name of Clopidogrel) and proton pump inhibitors (PPI) Prilosec 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 [20].

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

2 Related Work

Related works include research done in the field of health and science for clinical trials. It also includes researches that use semantic graphs and relationships for retrieval of information.

TrialX: It is a system, a third party tool that is built for recruiting related health 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.

Semantically Connected Named Entities and Relationships (SCOONER): Domain specific searches comprises of knowledge about the domain which serves as the basis of the search. However, there are three major concerns about such available knowledge: (1) exists only for few well known broad domains; (2) is of a basic nature: either purely hierarchical or involves only few relationship types; and (3) is not always kept up-to-date and is missing insights from recently published results. Kno.e.sis at WSU developed a framework addressing the above concerns. Their implementation provides an up-to-date knowledge based search system called SCOONER. The knowledge is extracted from recent bioscience abstracts. It uses a populated ontology (also called knowledge base) for semantic metadata extraction.

Retrieval of Similar Electronic Health Records Using UMLS Concept Graphs [17]. 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.

3 Our Approach

This system consists of the following two major parts: semantic Matchmaking and Semantic Ranking. This research paper focuses on the semantic Matchmaking. The matchmaking performs all the core operations of finding the relevant publications for any particular health record. Once the results are found, the Semantic Ranking provides us a way of calculating the relevance of the publications to a particular record.

The matchmaking and the ranking process are performed semantically where the system uses ontology mappings, synonyms calculation and hierarchy verification for calculating relevant results. Here is a diagram showing the overview of the functionality of our system:

![Figure 2: Overview of the System.](image)

Our health record consists of the following personal information: (1) Name, (2) Address, (3) ID, (4) Age, (5) KnownDisease, (6) Medications (7) Gender, (8) Symptoms, (9) PrimaryPhysician, (10) PhysicianId, (11) PrimaryPharmacy and (12) PrimaryPharmacyId.

A sample 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. The 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 ontology for semantic matchmaking. 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. We need annotations for populating both the ontologies. The publications from the PubMed including their title and abstract would be annotated. Also, the same process would be followed to annotate the information in the health records. Once both the annotations are received, both the ontologies are updated with the respective annotations and matchmaking procedure advances to the next step.

4 Building Blocks

In order to understand the matchmaking process completely, we must examine the following building blocks of the system. These components played a key role in the implementation of the system; (a) Electronic Health Records (b) PubMed (c) UMLS (d) NCBO BioPortal.
4.1 Electronic 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, sample health records were created for testing purposes based on Google Health’s format (Google Health is discontinued now). Here is a sample of the Google Health Record [9]:

\[\text{Figure 3: Google Health Samples [9]}\]

Sample health records (75) in XML format, similar to Google Health records for testing purposes were generated. These sample health records would enable the application to work properly even when fed with real health records. Here are a few samples of the health records that were generated:

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

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 of 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. 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) was populated. This allows the system to do accurate 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]: (a) Metathesaurus, (b) Semantic Network and (c) SPECIALIST Lexicon and Lexical Tools.

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 browsing, finding, filtering, searching ontologies. It can also be used for submitting new ontologies and for exploring mapping between ontologies.

BioPortal provides access to one of the largest repositories of biomedical ontologies. We can access these by web browsers or via web services (RESTful services). The BioPortal library consists of the following:

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

4.4.1 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 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. The expanded semantic annotations are obtained by considering the transitive closure, semantic distance and ontology mappings. Here is the workflow of the annotator web service:

![Annotator Workflow](image)

Figure 4: Annotator Workflow [16]

5 Semantic 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 paper 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. Semantic matchmaking is different from any other matchmaking in a way that the results are obtained in light of a shared conceptualization for the knowledge domain at hand, which we call ontology.

In order to obtain relevant results, we must ensure that the semantic annotations are 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.

5.1 Health Record Ontology

This ontology contains all the patients information with all the results obtained after the annotation process as follows:

(a) Name, (b) ID (unique), (c) Age, (d) Gender, (e) Known disease, (f) Medications, (g) Symptoms, (h) Annotations results for known disease (with synonyms), (i) Annotations results for medications (with synonyms), (j) Annotations results for symptoms (including synonyms)

5.2 Paper Publication Ontology

This ontology contains all the paper publications information. About 150 publication 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:

(a) Title, (b) Abstract, (c) Publication date, (d) Authors names, (e) Annotations for title, (f) Annotations for abstract

6 Matchmaking Algorithm

As seen in the Figure 5, 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:

![Matchmaking Workflow](image)

Figure 5: Matchmaking Workflow

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

For the Heath Records: (a) Disease name, (b) Annotations and synonyms of the disease names, (c) Medications, (d) Annotations and synonyms of the medication names, (e) Symptoms and (f) Annotations and synonyms of the medication names.
Our system now performs the matchmaking and provides the results accordingly. In this process, the system not only performs the keyword matching, but also takes into consideration the semantic hierarchy, transitive closure, ontology mappings, semantic distance, 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.

7 Testing the Matchmaking Algorithm

Let us consider motivating scenario mentioned in Motivation section. 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 effects of both drugs when taken together.

Results: Here are the results related to: Mathew burton
Patient Record Number: 1284
Disease: Heart Attack
Rank is: 8
Link is: http://www.ncbi.nlm.nih.gov/pubmed/22053225
Rank is: 7
Link is: http://www.ncbi.nlm.nih.gov/pubmed/22053219
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

Here is a Snapshot of the User Interface results:

Figure 6: Snapshot of results of test case

8 Comparison with Syntactic Matchmaking

8.1 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.

8.2 Discovery of Medication Side Effects

Our system enables a user to not only get the related publications based on the disease that they are suffering from, but also enables them to discover any side effects of the medications and drugs they are taking. The results are not just limited to the disease’s name because of using UMLS and 173 ontologies during the matchmaking process. 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 user, 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

In our system, the side effects of drugs are discovered whether they appear directly or indirectly as it checks the annotations, synonyms etc. This is something that cannot be achieved by syntactic matchmaking.

8.3 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, which may have some synonyms of the current...
medications or symptoms. This search goes beyond the keyword search and retrieves the papers semantically. 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
Title: Vomiting

Rank is: 6
Title: The patient with haematemesis and melaena

Rank is: 6
Title: Gastric duplication cysts as a rare cause of haematemesis

### 8.4 Knowledge Discovery without Specific Input

Our system allows a user to discover the papers related to them without having particular information about the disease that they might be suffering from. Since the search can be done with any one of the parameters (medications, disease, symptoms etc.), the complete information is not mandatory. A person might search based on his 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.

### 9 Preliminary Evaluation

In order to evaluate the functionality of our system, we compared our results with 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:**

Here is snapshot of the results obtained from our system:

**Name:** Mathew Burton
**Patient Record Number:** 1284
**Disease:** Heart Attack

We can see that our system, gave the results of papers discussing the combined effects of both the drugs Prilosec and Plavix 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 or keyword, but it also takes the semantic relationships between the provided information into consideration. Thus in the above example, we did not get results related only to Heart attack, but also results related to symptoms, medications, side effects of medication, combined effect of two medications etc.

10 Conclusions and Future Work

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. 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 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.

11 References


