Entity Resolution for Symptom Vs Disease for Top-K Treatments

K.A.Vidhya
Dept. of Computer Science,
Anna University, Chennai, India
avidhya06@gmail.com

R.Soorya
Dept. of Computer Science,
Anna University, Chennai, India
Soorya9795@gmail.com

Saranavan.N
Dept. of Computer Science,
Anna University
Chennai, India
saravanan211094@gmail.com

T.V.Geetha
Dept. of Computer Science,
Anna University, Chennai, India
tv_g@hotmail.com

Singaravelan.M
Dept. of Computer Science,
Anna University, Chennai, India
velanceg150795@gmail.com

Abstract—**The sufficient information on the web of data calls for an efficient entity resolution techniques in biomedical records, symptom vs. disease where a particular symptom, subjected to ambiguity. There may be several terms that refer to the same symptom. Thus, Entity Resolution becomes an essential task to identify a particular disease, for a given symptom. This work aims at suggesting the best alternate treatments to the health care professionals based on the patient's disease. A hybrid recommender system that recommends alternate treatments to the healthcare professionals based on their patient's disease, symptoms, age, and gender is designed and developed. Nowadays, there are new on-going treatments which are much successful know from the clinical trials for a particular disease. The content-based filtering, find the different treatments that are available for user's disease based on their outcome obtained by sentiment analysis. The collaborative filtering uses the similarity measure to find the similarity between the user and the patients by considering their age, gender, location, symptoms, and diseases. The treatments obtained from both these modules are then ranked by assigning a score based on their effectiveness and side effects. Finally, the top-k treatments for the disease are recommended to the health care professionals.**

Keywords—Entity Resolution, Linked Open Data, Clinical Trails, Naïve Bayes

I.INTRODUCTION

Nowadays Data Mining plays a major role in the healthcare industry to enable health systems to use data and analytics to identify the ineffectiveness and other effective methods to improvise care and also by decreasing the expenditure. Many Experts in the health care industry believe that the opportunities to improve care and reduce costs concurrently constitutes to thirty percent of overall health care spending. The main issue is that the industry falls behind in implementing the data mining and analytic techniques for such abundant processing such abundant information that is on the web. Recommender systems apply the techniques from areas such as Human-Computer Interaction and Information Retrieval where the sequence of Data mining the processes such as Data pre-processing, Data Analysis and Result Interpretation are carried out in sequence. Mining the important information from the health informatics data that is available on the web and doing an appropriate predictive analysis is an evergreen research as treatment keeps changing based on the new drug that has been coming in the market. To make aware of the current trends in the treatment and the side effects of the drug based on user specific, location specific and symptom specific. The huge volume of information created by the health care system is to be put into a course of action and analyzed and the usual procedures are required to be processed to arrive apt information for the health care professionals to make an overall decision.

This system helps to recommend the best alternative treatment for the clients based on their disease, symptoms, age, and gender. Healthcare professionals who do not have sufficient experience can use this system as a support. Also, this system can be helpful for the professionals to know about many alternative and treatments available. It also provides the health care professionals with new treatments. The recommendation is made based on both content-based filtering and collaborative filtering,[9] [16] [13] [14] in content-based filtering. The data is mined from linked open data which is the large set of RDF triples that helps to find the different treatments that are available for particular disease along with their outcomes. Linked Open Data (LOD), LinkedCT [17], which has clinical trial information is queried using the SPARQL querying language, for the primary outcome status and secondary outcome and the impact of the treatment for a particular disease. For a given set of symptoms, there is a need for Entity Resolution (ER) which is aka object resolution or duplicate detection which is being applied for symptom and disease names to identify the exact disease for the given symptom, helps to group the records that refer to the same entity. The collaborative filtering uses the similarity measure to obtain the treatment statistics for the diseases such as some users benefited from the treatment, side effects of the treatment, and effectiveness of the treatment. Then the treatment obtained from both these modules is ranked by considering some specific features. Then the top-k alternate treatments for the disease are recommended.

II.RELATED WORK

Freitas et al. [1], in their study, focused on querying linked data graphs with the use of semantic relativity, thereby linked data fetch intrinsic challenges while the users and the applications get through the existing available data. The Data that are scattered over the web are heterogeneous, complex and even distributed data sets. The users who were working with the linked data over the web could be able to search in the simplest manner. Linked Data query mechanism must be able to abstract users from the representation of data. They also concentrate on by checking the natural language query mechanism for an independent vocabulary term on Linked data. For which they have approached with the combination of the entity query based search, spreading activation. This has been arrived by taking into consideration of the existing study works that are based on the similarity measures or the term expression on WordNet. The query methods or mechanism in the present study got the input from the natural language query and the output retrieved is a set of triple paths that are corresponding to the answers mingled into a connected graph. After this process is getting over, the key entities and the pivots are identified and determined and the user used work of the natural language query is analyzed for query parsing component. The output conceived out of this component is known as Partial Order Dependency Structure and the algorithm process made out of the query continues with a spreading activation search.

MacKellar et al. [2] in their work identified patient-related clinical attempts and search through the semantic amalgamation of Linked Open Data. Patients facing a serious disease, want to be able to search for relevant treatment [6] for new or more effective alternative treatments. The NIH (National Institutes of Health) [17] makes all of its trials available on a website, in fact, for this purpose. In NIH the facility for traversing a particular history of the trail is much difficult as the patient have to swift through lengthy text reports for relevant information [22]. The author has focused on developing patient-focused clinical trial [10][12][15] to build a proper clinical trial facility. An integration of various linked open data has been aimed to improve the semantic understandability using RDF(Resource Description Framework) triples. LOD sources such as the LinkedCt [11] clinical trials dataset, the SIDER [18] drugs and side effects dataset, and the Unified Medical Language System (UMLS) medical ontology are leveraged for consistent semantics across concepts used in different data sources. The system architecture includes a knowledge representation module where RDF triples are generated by extracting data from different sources and linking them with semantic information.The author facilitates the work in such a way that provides patients with powerful semantic search capabilities that use query processor and semantic reasoning components; and semantic-link browsing where the navigation from one concept to another can help users with visual search and exploration of clinical trials and related information[11].

Use cases to illustrate the system functions such as query processing steps, semantic-search, and semantic-link browsing is presented. Madhavi et al. [3] proposed a Recommender System on Medical Recognition and Treatment. Off Late, the intrusion of the computer systems and the decision support systems being developed has been an important issue in medical science. The author has developed a medical recommender system for disease recognition and treatment. So, a recommendation concept and its definition are dealt here. This system can learn the information cashed from patients. Collaborative filtering is a method that is used to filter.

The last feature of this system, its ability to understand the required period of treatment. Based on this matter, the system should anticipate how long is needed to finish the treatment. One part contains important characteristics, and another one is digit information scope. Each of them has its effects and functionalities. Similarity sim(a, b) of patients a and b, given from the rating matrix R, is defined in the formula. In fact, ignores many noises of the system caused by variety in patient ratings which makes it possible for prediction. Q.Wang et al. designed an ER framework using a locality sensitive hashing (LSH) techniques, that effectively merges the semantic features and textual features into ER blocking the process. To understand how the similarity metrics is possible to affect the efficiency of the ER blocking, the robustness of the similarity metrics and their functions with respect to LSH [4][5][7][8] families is studied.

III.PROPOSED WORK

For recommending treatments using both content-based and collaborative method requirement for recommender system consists of three phases. The first phase covers the data collection where the dataset is collected from Linked Open Data (LOD), and the sentiment analysis is done based on the primary outcome and the secondary outcome with the short text as positive and negative. The patient’s data is crawled from websites like PatientsLikeMe [19] from which the similar patients are clustered with an effective score and the side effects.

The ER is applied on the set of symptoms obtained from the patient symptoms where ER is done by novel modified blocking technique where each entity is placed in the set of similar blocks. In this technique, the semantically similar symptoms are grouped together into the same block. Using the wordnet similarity measure taking into account the path length and the word distance measure, the semantic similarity is found between the blocks. With the blocking method, the symptom Vs disease are indexed based on the measure obtained above. An ER based Hybrid Recommendation System is proposed using all the information mined from the web of health care data, and here the system obtains the treatments from both content-based and collaborative filtering. The treatments are then ranked based on their effectiveness and side effects. Finally, top-k treatments are recommended to the users.

Content Mining and Polarity Finding Using LOD and NaiveNet

LinkedCT is a linked data version of clinicaltrials.org containing data about the clinical trials. Clinical trials are the experiments that are carried out on the humans to test the new treatments for a disease. The LinkedCT consists of the RDF data which is in the N-Triples Format. The RDF data consists of a subject, predicate, and object and are linked with the other LOD's such as diseasome [20] and sider. The general description of LinkedCT RDF triples, which has a set of attributes for the clinical trials conducted and links many data sources. The feature vector is computed for retrieving the top-k treatment recommendations are the positive outcome, negative outcome, side effect and clinical trail. These concepts are interlinked with the other linked open data like Sider and Diseasome.

This is a clinical trial, NCT01786512, which is aimed at COSMIC-HF which is usually a Chronic Oral Study of Myosin Activation to Increase Contractility in Heart Failure. The first resource shown is a treatment, Omecamtiv mecarbil, this treatment is used in the intervention protocol of the clinical trial for COSMIC-HF - Chronic Oral Study of Myosin Activation to Increase Contractility in Heart Failure, which is represented by the link. Thus LOD is in the form of a graph needs to be converted into a suitable format for further processing. So, this module extracts the required fields from the N-triples file by using the querying language SPARQL. The dataset contains Primary outcome and Secondary outcome for each trial and the classification is done using the outcome as positive or negative.

```
String query = " PREFIX rdftrial:
<http://data.linkedct.org/resource/trial/> \n" +
  " PREFIX rdflabel: <http://www.w3.org/2000/01/rdf-
schema#> \n" +
 " PREFIX rdfgen:
<http://data.linkedct.org/vocab/resource/> \n" +
 "SELECT ?id ?title ?condition ?phase ?prim_measure
?prim_desc ?sec_measure ?sec_desc ?interv_type
?interv_name ?interv_desc  \n" +
 "WHERE
 { \n" +
   "?trial rdfgen:trialid ?id . \n" +
   "?trial rdfgen:brief_title ?title . \n" +
   "?trial rdfgen:trial_condition ?condition . \n"
   "?trial rdfgen:trial_primary_outcome ?primary_link . \n"
+
   "?trial rdfgen:trial_secondary_outcome ?secondary_link
.  \n" +
   "?trial rdfgen:trial_intervention ?intervention_link . \n"
+
   "?primary_link rdfgen:outcome_measure
?prim_measure . \n" +
   "?secondary_link rdfgen:outcome_measure
?sec_measure . \n" +
   "?primary_link rdfgen:outcome_description ?prim_desc
. \n" +
   "?secondary_link rdfgen:outcome_description
?sec_desc . \n" +
   "?intervention_link rdfgen:intervention_description
?interv_desc . \n" +
   "?intervention_link
rdfgen:intervention_intervention_name   ?interv_name .
\n" +
    "?intervention_link
rdfgen:intervention_intervention_type ?interv_type . \n" +
              "?trial rdfgen:phase ?phase . \n" +    "}";
```

Based on the short review comments, the effectiveness of the treatment for the clinical trial has to be determined. The polarity finding has been carried out using the NaiveBayes [21] conditional probability method and we have constructed feature vector called Naivenet for finding the polarity of the short text in the LOD. Based on the performance of the Naivenet, the system decides whether it's positive review or a negative review about the clinical trials. For doing sentiment analysis the machine learning approach has been used to analyze the statement which falls in the classification method to train the model using training dataset. Then, the features from the outcome are extracted and given to the classifier model (Naive Bayes) to determine the type of the outcome. Thus, based on the user's disease the treatments that have positive outcomes are taken and given to ranking module along with the results from collaborative filtering module. After querying from the web of data, the words in the outcome description are mined for analyzing it as positive and negative words. Then the sentiment analysis is done for the words with Naïve Bayes algorithm which is the supervised machine learning algorithm, the positive word list and the negative word list which are integrated into the system beforehand. The classification [23] is done by the conditional probability which is calculated by:

$$p\ \ c_j\ \ d\ \ = \frac{p\ \ d\ \ c_j\ \ p(c_j)}{p(d)} \qquad (1)$$

The naïve independence is used as the criteria for training the system and deciding whether this belongs to the positive description or the negative description based on the probability of the word is decided. The higher the probability, the higher the word belongs to any of the group. To improvise the results, the co-occurrence word is also considered as the word "not-hygienic" gives a negative meaning, but when considered as hygienic it takes positive meaning.

Then the top-k treatment is fetched used the other features like patient age, gender, and location. The content-based filtering uses the treatment database to find the suitable treatments based on the user's disease. It uses the naive Bayes algorithm to perform sentiment analysis on the outcome of the treatment considering different measures. Since the dataset is voluminous in size indexing is performed on the disease name and the top treatments for the disease are obtained. The patient database is used to perform the user-user collaborative filtering. We use the Jaccard similarity measure to calculate the similarity between the end user and the other users in the database. The person's age, gender and location are considered while calculating the similarity. Thus, the treatments taken by the top-k similar users are obtained.

Entity Resolution For Symptom Vs. Disease

Entity Resolution can be defined as identifying the two different real-world entities that refer to the same entity. In bio-medical literature, the term entity resolution can be referred as an abbreviation of disease name, two different symptoms referring the same disease. In this work deciphering, the symptoms are much needed as the disease has to be identified for the given symptom and the top-k treatments has to be suggested. The symptoms are subjected to uncertainty where one symptom corresponds to one or more disease, and there may be several terms that refer to the same symptom. Thus, entity resolution becomes essential for deciphering the symptom where this module helps to group the different terms that refer to the same symptom. The entity resolution is performed by using a blocking technique where the different terms that refer to the same symptom are placed in the same block.

The semantic and word order similarity measures are considered in the blocking technique. The semantic vector for each symptom is constructed based on the path length and depth between the words in the WordNet hierarchy. The importance of the word in the sentence is also considered while calculating the similarity. We also consider the word order similarity where the order of the words in the sentences is used to find the similarity. Thus, both these measures are combined to give the final similarity value. If the overall similarity value exceeds the threshold, then the symptoms are placed in the same block. The overall similarity is the combination of both the semantic and word order similarity and is defined by:

$$S\ T_1, T_2 \ = \delta S_s + (1-\delta) S_r \tag{2}$$

$$S\ T_1, T_2 \ = \delta \frac{S_1.S_2}{|\ S_1\ |.|\ S_1\ |} + (1-\delta)\frac{|\ r_1 - r_2\ |}{|\ r_1 + r_2\ |} \tag{3}$$

Delta is the value chosen between 0.5 and 1. s1; s2 are the semantic vectors. r1; r2 are the word order vectors for the input sentences. Ss is the Semantic similarity. The entity resolution phase helps us to identify the terms that refer to the same symptom using Blocking algorithm. There may be cases where the end users symptom may have the same meaning as that of another symptom in the patient's database. Thus, while performing the user-user collaborative filtering these symptoms should not be considered different as they are semantically similar. This leads to the need for entity resolution where the semantically similar symptoms are being placed in the same block.

Sample Symptoms for a particular disease

Skin pain- Chest pain- Joint pain- Pain in lower back-
Back pain-Dry skin- Muscle pain- Pelvic pain -
Pain
Stomach pain- Muscle and joint pain- Rash or skin problems- Abdominal pain
Delusions- Paranoia - Visual hallucinations-
Auditory hallucinations
Deteriorating mental ability (dementia)- Excitability-
Emotional lability
Craving alcohol- Craving nicotine- Guilt related to addiction
Vomiting- Nausea and vomiting

This phase constructs a semantic vector for the given symptoms using wordNet hierarchy and uses cosine similarity measure to find the semantic similarity between them. This measure considers only the similarity between the words in the sentence but not the order of the words in the sentence. This leads to the use of word order similarity. The result from both these similarity measures is combined in the blocking technique. The following diagram shows the set of symptoms grouped together using Blocking technique.

Collaborative Filtering

The patient's data is crawled from *PatientsLikeMe* website which contains the patient's age, location, gender, disease, symptoms and the treatments they had taken along with their side effects and effectiveness. This module performs the user-user collaborative filtering by finding the similarity between the end user and the other users in the dataset by using Jaccard Similarity. The similarity is calculated by considering the factors such as disease, symptoms, age, gender, and location.

$$J\ S_1, S_2 \ = \frac{|s_1 \cap s_2|}{|s_1 \cup s_2|} = \frac{|s_1 \cap s_2|}{s_1 + s_2 - |s_1 \cap s_2|} \tag{4}$$

If s_1 and s_2 are both empty, we define J (s_1, s_2) = 1.$\{0 \le J(s_1, s_2) \le 1\}$

Where S_1 and S_2 are the set of symptoms. Here entity resolution is applied for the symptoms so that different terms referring to the same symptom are considered as the same entity. This helps us to find the similar users more accurately. The top-k similar users are obtained, and their treatments are given to the ranking module.

Algorithm for Entity Resolution

```
For each tuple $T_i$=1 to n do
    Get the user info $U_i$= {$u_1$,$u_2$,$u_3$ ,$u_4$,$u_5$…..$u_n$}
    Get the set of symptoms {$s_1$,$s_2$,$s_3$,$s_4$,$s_5$,$s_6$,$s_7$….$s_n$}
For each Symptom {$S_1$….$S_n$} do
Assign Set of blocks{$B_{1,}B_{2.....}B_n$}for similar entities
Calculate Semantic Similarity using
        $S\ T_1, T_2 \ = \delta S_s + (1-\delta) S_r$
        $S\ T_1, T_2 \ = \delta \frac{S_1.S_2}{|\ S_1\ |.|\ S_1\ |} + (1-\delta)\frac{|\ r_1 - r_2\ |}{|\ r_1 + r_2\ |}$
```

```
Calculate Word Distance using WordNet
 sim = semantic similarity + word distance
End For
For each Patient {$P_1$…$P_j$} do
    <Set> $P_i$ = {$P_{age}$,$P_{gender}$ , $P_{location}$}
     similarity = JACCARDSIMILARITY(Patient; User)
end for
     SORT(similarity; reverse = True)
     SP <Set> Top ←KPatients
     SP ={$T_1$,$T_2$,$T_3$….$T_n$}
for each patient in similarpatients do
      treatments = {<Set>$P_i$.<set>$T_i$}
      APPEND(similarpatienttreatements; <Set> $T_i$)
end for
return similar patient treatments
```

Ranking

Ranking helps to recommend the most suitable treatment to the user based on the health records as well as the statistics of the particular treatment. Here, the treatments obtained from both the content based and collaborative module are ranked by considering their effectiveness and the side effects. A score is assigned to each treatment based on these two measures. This helps us to find the best treatments for the disease. Finally, these treatments are recommended to the user and are considered to be compatible with the user's health condition.

Algorithm for Ranking

```
AT<set> ← Treatments from ContentBasedFiltering
SPT <set> ←Treatments from CollaborativeFiltering
      score = 0
      for each treatment in {$T_1$,$T_2$,$T_3$….$T_n$} do
                 currEffectiveness=treatment.effectiveness
     APPEND(effectiveness;currEffectiveness)
currSideEffects = treatment.sideEffects
APPEND(sideEffects;currSideEffects)
SORT(effectiveness; reverse = True)
SORT(sideEffects; reverse = True)
score+ = SCORE(effectiveness)
score+ = SCORE(sideEffects)
end for
SORT(treatmentList) //
recommendedTreatments ← Top-k Treatements based on score
            return recommendedTreatements
```

IV.EVALUATION AND RESULTS

The system implemented in 16GB ram with intel i5 processor loading the linked open data of 8Gb data is much harder as the no of triples is colossal to be mined, using the option of TTB Loader. The data analysis was done on the both content-based, and collaborative filtering method where the top-k treatments need to be fetched. The Jaccard similarity measure between two sets is the result of a division between the number of features that are common to both divided by the total number of properties considered. If A ∧ B are both empty, we define J(A, B) = 1 where A and B are the set of symptoms. The treatments obtained from the content based, and the collaborative module cannot be recommended to the users without analyzing their statistics. Thus, we use the ranking module to provide the ranking score to the treatments based on their effectiveness and side effect information obtained from PatientsLikeMe website. Finally, the treatments with the top-k scores are considered to be the best for the end users disease and are recommended.

Precision and Recall

In the top-k treatment recommendation, the number of treatments retrieved has to check with standard treatments present in UMLS link. Pattern recognition and information retrieval are calculated with the basic metrics, precision, and recall in the case of binary classification. The precision that is also known as positive predictive value, is the part of the document that is relevant to the required user's information need or query among the retrieved document, whereas, recall which is also identified as sensitivity is the part of the documents that are relevant to the query raised.

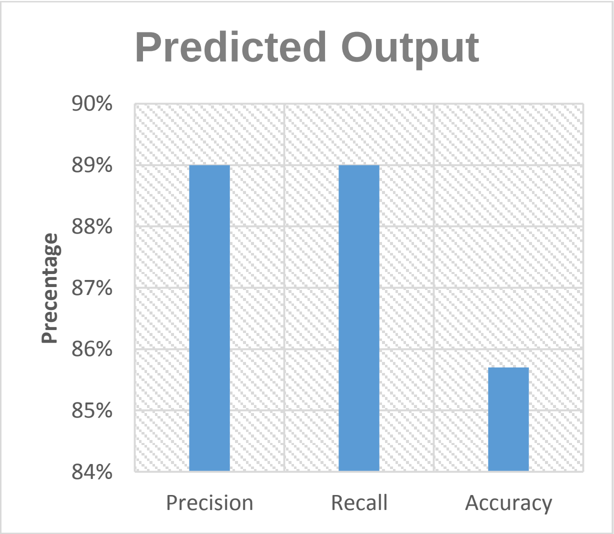

Fig 1: Predicted Output

Hence, we could identify that both recall and precision works based on the mutual understanding of the query raised and the measure of relevance.In this work, we use True Positive, True Negative, False Positive and False Negative values to find the accuracy of the work.

Discounted Cumulative Gain

The recommendation system uses the ranking measure, Discounted Cumulative Gain (DCG). The recommendation system retrieves the top-k treatments which are often used to measure the effectiveness of related treatments. In a term search process, a graded relevance scale of treatment is used to achieve the target result set. DCG identify and calculate the efficacy, or gain, of treatments depending upon its grade or rank placement in the list of result arrived. The value is gathered from the top list of the outcome result to the bottom with the value or gain of each and every result dropped out at lower ranks.

The basic principle of DCG is that the extremely relevant documents appearing lower in a search query result list should be taken for granted in another way as the graded relevance value is reduced logarithmically proportional to the location of the result. The discounted CG accumulated at a particular rank position P is defined as:

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2(i)} \qquad (5)$$

$$nDCG_p = \frac{DCG_p}{IDCG_p} \qquad (6)$$

Note that in a perfect ranking algorithm, the DCG_p will be the same as the $IDCG_p$ producing an $IDCG_p$ of 1.0. All nDCG calculations are then relative values on the interval 0.0 to 1.0 and so are cross-query comparable.

Evaluation Results - Precision and Recall

Here we took 74 sample test cases. Out of 14 test cases, there are 9 positive cases and 5 negative cases.

Precision = TP / (TP+FP) = 8/9 = 0.89 = 89%

Recall = TP / (TP+FN) = 8/9 = 0.89 = 89%

Accuracy = (TP + TN) / (TP + TN + FP + FN) = 12/14 = 0.857=85.7%

DISCOUNTED CUMULATIVE GAIN

Disease: Multiple Sclerosis
Symptoms: nervous, bowel problem, pain, tiredness, sexual dysfunction
Gender: Female
Age:50

Output:

1. Natalizumab - Prescription Drugs
2. Teriflunomide - Prescription Drugs
3. Dimethyl fumarate - Prescription DrugsInterferon beta-1a SubQ injection - Prescription Drugs
4. Fingolimod - Prescription Drugs
5. Atorvastatin Therapy – Treatments

Results

The system produces the result by ranking the treatments. Each treatment is to be judged on a scale of 0-3 with 0 meaning irrelevant, three meaning completely relevant, and 1 and two meaning ‘somewhere in between’. The relevance score for the above test case output is

TABLE1: VALUE OF DISCOUNTED CUMULATIVE GAIN

i	*rel (i)*	*log2 (i)*	*rel(i) / log2 (i)*
1	3	0	N/A
2	2	1	2
3	3	1.585	1.892
4	1	2.0	0.5
5	2	2.322	0.861

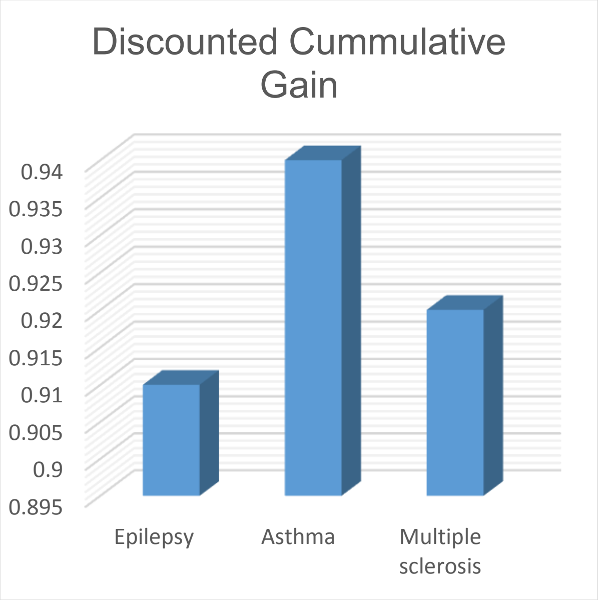

Fig 2: Discounted Cummulative Gain

That is treatment one is completely relevant, two is somewhere in between and so on. The cumulative gain is:
CG = 3+2+3+1+2+3 = 14

DCG = 3 + (2+1.892+0.5+0.861+1.16) = 9.41

3,2,3,1,2,3 nDCG = DCG / IDCG = 9.41 / 10.139 = **0.928**

To normalize DCG values, an ideal ordering for the given treatment is needed. For this example, that ordering would be the monotonically decreasing sort of the relevance scores, which is: 3,3,3,2,2,1

The DCG of the ideal ordering is:

IDCG = 10.139

TABLE2: THE LIST OF TREATMENTS RETRIEVED FOR A GIVEN DISEASE

Disease	Symptoms	Age	Gender	Treatments	Precision	Recall
Multiple sclerosis	nervous, pain	55	Female	Natalizumab, Teriflunomide, Atorvastatin therapy	0.782	0.892
Fibromyalgia	sleeplessness, anxiety	35	Male	The Clinical Effect of Low-Intensity Electromagnetic Field Neurostimulation in Fibromyalgia Syndrome Patients	0.632	0.832
Fibromyalgia	fatigue, menstrual cramps	39	Female	Hydrocodone-Acetaminophen, Milnacipran, Tizanidine	0.784	0.845
Amyotrophic lateral Sclerosis	difficulty in walking, muscle cramps, fatigue	50	Male	Feeding tube, Dextromethorphan with Quinidine, Neuromuscular Ultrasound in ALS	0.872	0.863
HIV	fatigue, sore throat, weight loss	48	Male	Positive Stribild, Efavirenz-Emtricitabine-Tenofovir	0.765	0.856
Asthma	cough, nervous, breathing problem	40	Female	Albuterol, Budesonide	0.876	0.876
Major depressive disorder	anxious mood, fatigue, pain	53	Male	venlafaxine, Fluoxetine	0.932	0.756
Epilepsy	Insomnia, head pain, Anxious	45	Female	zonisamide, phenytoin	0.821	0.734
Diabetes	Tiredness, urinating, Pain	60	Male	Insulin Glargine, Metformin, Glimepiride	0.786	0.695
Thyroid cancer	Pain, Nervous, depressed	44	Female	Clonazepam, Gabapentin, Thyroidectomy	0.654	0.856
Primary Lateral Sclerosis	Yawning, pain, Fatigue	58	Male	Citalopram, Tizanidine, Diazepam	0.653	0.763
Breast cancer	Insomnia,Depressed,Fatigue	65	Female	Tamoxifen,Anastrozole,Hydrocodone-Acetaminophen	.889	.832

V.CONCLUSION AND FUTURE WORK

On the basis of above studies, it is concluded that the system is very versatile and helps in identifying the alternative treatment to the doctors based on their symptoms, diseases, age, and gender. With well-defined data set, the system can handle a large number of diseases, and it can recommend treatments very accurately. The system uses the entity resolution (Modified Blocking technique) to find the symptoms that are similar to the user symptoms. By using the entity resolution, the system can find the most related treatments to the user. The system also performs user-user collaborative filtering to find similarity among the users. Thus, the system finds the most similar user and the treatments that are suitable for the input user. The system also ranks the obtained treatments based on their effectiveness and side effects for user simplicity.

The patient's location and their previous history are not considered in the system. If the location is considered the treatments can be recommended based on that. For example, the Indian patient will get treatments like Ayurveda, Homeotherapy, etc. The US patient will get the treatments like English medicines and the treatments well suited for their environment. By considering the user's history, the system can easily find out the treatments that are already taken and find the possible treatments based on their health condition. By considering the location and the previous history of the patient the system can improve its accuracy and recommend treatments more effectively. Thus, there is a great scope for this work in near future and above said extensions could be made to the system.

ACKNOWLEDGMENTS

I am greatly thankful for UGC in supporting this study, which was funded by DST PURSUE UGC, NEW DELHI, INDIA (9500/PD2/2014-2).

REFERENCES

[1]. Freitas, A., Oliveira, J. G., O'riain, S., Da Silva, J. C., & Curry, E. (2013). Querying linked data graphs using semantic relatedness: An independent vocabulary approach. *Data & Knowledge Engineering*, *88*, 126-141.Chicago

[2]. MacKellar, B., Schweikert, C., & Chun, S. A. (2013, July). Patient-oriented clinical trials search through the semantic integration of Linked Open Data. Cognitive *Informatics & Cognitive Computing (ICCICC), 2013 12th IEEE International Conference on* (pp. 218-225). IEEE.

[3]. Mahdavi, M., Meisamshabanpoor (2012). Implementation of a Recommender System on Medical Recognition and Treatment. *International Journal of e-Education, e-Business, e-Management and e-Learning*, *2*(4), 315.

[4]. Q. Wang, M. Cui and H. Liang, "Semantic-Aware Blocking for Entity Resolution", *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 1, pp. 166-180, 2016.

[5]. Kolb, L., Thor, A., & Rahm, E. (2012). Multi-pass sorted neighborhood blocking with MapReduce. *Computer Science-Research and Development, 27*(1), 45-63.

[6]. C. Bizer, T. Heath, T. Berners-Lee, "Linked Data—The Story So Far," *International Journal on Semantic Web and Information Systems* 5 (3):1–22, 2009.

[7]. Li, L., Li, J., & Gao, H. (2015). Rule-Based Method for Entity Resolution.*Knowledge and Data Engineering, IEEE Transactions on*, *27*(1), 250-263.

[8]. Y. Li, D. McLean, Z. Bandar, J. O'Shea and K. Crockett, "Sentence similarity based on semantic nets and corpus statistics", *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 8, pp. 1138-1150, 2006.

[9]. G. Adomavicius and Y. Kwon, "New Recommendation Techniques for Multicriteria Rating Systems", *IEEE Intell. Syst.*, vol. 22, no. 3, pp. 48-55, 2007.

[10]. D. W. Lonsdale, C. Tustison, C. G. Parker, and D. W. Embley, "Assessing clinical trial eligibility with logic expression queries," *Data & Knowledge Engineering*, vol. 66, no. 1, pp. 3–17, Jul. 2008.

[11]. T. Berners-Lee, Linked Data Design Issues, http://www.w3.org/DesignIssues/LinkedData.html. 2009.

[12]. P. Nadkarni, L. Marenco, R. Chen, E. Skoufos, G. Shepherd, P. Miller, Organization of heterogeneous scientific data using the EAV/CR representation, Journal of the American Medical Informatics Association (1999) 478–493.

[13]. G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems:A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[14]. Z. Luo, R. Miotto, and C. Weng, "A human-computer collaborative approach to identifying common data elements in clinical trial eligibility criteria.," *Journal of Biomedical* Informatics, vol. 46, no. 1, pp. 33–39, Jul. 2012.

[15]. Resource Description Framework (RDF) Model and Syntax Specification, http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/. 2013.

[16]. G. Adomavicius and Y. O. Kwon, "New recommendation techniques for multicriteria rating systems," IEEE Intelligent Systems, vol. 22, no. 3, pp. 48–55, 2007.Z. Luo, S. B. Johnson, A. M. Lai, and C. Weng, "Extracting Temporal Constraints from Clinical Research Eligibility Criteria Using Conditional Random Fields" in Proceedings of the Annual AMIA Symposium, 2011, pp. 843–852.

[17]. https://datahub.io/dataset/linkedct

[18]. http://sideeffects.embl.de/about/

[19]. https://www.patientslikeme.com/

[20]. https://datahub.io/dataset/fu-berlin-diseasome

[21]. K.A.Vidhya, T.V.Geetha and G.Aghila,"Text Document Classification using Rough Set theory and Multi-level Naïve Bayes", International Journal of Applied Engineering Research (IJAER), pp.331-336, Volume 10, Number 75 (2015).

[22]. Vidhya. K. A., and G. Aghila. "Text mining process, techniques and tools: an overview." International Journal of Information Technology and Knowledge Management 2.2 (2010): 613-622.

[23]. Vidhya. K. A., & Aghila, G. (2010, February). Hybrid text mining model for document classification. In Computer and Automation Engineering (ICCAE), 2010. The 2nd International Conference on (Vol. 1, pp. 210-214). IEEE.