ISSN: 0976-3104

REGULAR ISSUE
Shobhana and Radhakrishnan

ARTICLE OPEN ACCESS

THE IIOAB JOURNAL
ISSN: 0976-3104

ESTIMATION OF SEMANTIC SIMILARITY BETWEEN CONCEPTS AND FUZZY RULES OPTIMIZATION WITH MODIFIED GENETIC ALGORITHM (MGA)

B. Shobhana[1*] and **R. Radhakrishnan**[2]
[1]*Anna University, Chennai, Tamil Nadu, INDIA*
[2]*Sasurie College of Engineering, Vijayamangalam, Tirupur, Tamil Nadu, INDIA*

ABSTRACT

Aims: *Semantic similarity estimation is an important component of analyzing natural language resources like clinical records. In the past, several approaches for assessing word similarity by exploiting different knowledge sources have been proposed. Some of these measures have been adapted to the biomedical field by incorporating domain information extracted from clinical data or from medical ontologies.* ***Materials and methods:*** *In the recent work semantic similarity results from medical ontologies is evaluated based on some rules and assumptions. However these rules generation is completely performed based on the concepts only not by consideration of semantic similarity measures on the concepts. Here a new similarity measure is proposed and it is combining both super concepts of the assessed concepts and their common specificity feature. The similarity measure is performed based on Information Content (IC) and context vector. Then Fuzzy Rule base Modified Genetic Algorithm (DLCS-FRMGA) approach is introduced to rule optimization phase.* ***Results:*** *Using the MGA, the problem of finding an optimal rule base can be reduced to improve their applicability and accuracy. Thus the deterioration problem never happens since the best solution from the current generation will be superior to or at least the same with the past.* ***Conclusion:*** *Using MeSH and SNOMED CT as the input ontology, the accuracy of DLCS-FRMGA proposed method is evaluated according to a standard benchmark datasets of manually ranked medical terms.*

Received on: 1st - April -2016
Revised on: 7t-h - May-2016
Accepted on: 22nd – May-2016
Published on: 5th– August-2016

KEY WORDS

Semantic similarity, Super concepts, Evaluated concepts, Least Common Subsumer (LCS), Information Content (IC), Modified Genetic Algorithm (MGA), Fuzzy Rulebase Modified Genetic Algorithm (DLCS-FRMGA)

***Corresponding author: Email:** sobhanab.scholar@gmail.com; **Tel:** +91-9443901212

INTRODUCTION

An information retrieval process begins when a user enters a query into the system. User queries are matched against the database information. However, as opposed to classical SQL queries of a database, in information retrieval the results returned may or may not match the query, so results are typically ranked. This ranking of results is a key difference of information retrieval searching compared to database searching [1]. Depending on the application the data objects may be, for example, text documents, images [2], audio [3], mind maps [4] or videos. Often the documents themselves are not kept or stored directly in the IR system, but are instead represented in the system by document surrogates or metadata [5].Semantic search seeks to improve search accuracy by understanding the searcher's intent and the contextual meaning of terms as they appear in the searchable data space, whether on the Web or within a closed system, to generate more relevant results [6].

Ontology is “a formal explicit specification of a shared conceptualization”. Ontology provides a common understanding of a term and also its relationship with other terms. Thus a hierarchy can be formed with the related terms. Ontology compartmentalizes the variables needed for some set of computations and establishes the relationships between them [7] [8]. The fields of artificial intelligence, the Semantic Web, systems engineering, software engineering, biomedical informatics, library science, enterprise bookmarking, and information architecture all create ontologies to limit complexity and to organize information.

Semantic Information Retrieval has become the core part of any search engine. Many papers deal with SWS that uses the OWL language for constructing ontology. DySE System (Dynamic Semantic Engine) [9] implements a context-driven approach in which the keywords are processed in the context of the information in which they are retrieved, in order to solve semantic ambiguity and to give a more accurate retrieval based on user interests. Ontology Construction in Education Domain [10] deals with the construction of Ontology for specific University constructing instances specifically. Here the usage of Protégé tool for constructing the ontology is illustrated. Query sentences as semantic networks [11] paper describes procedure for representing the queries in natural language as

THE IIOAB JOURNAL
ISSN: 0976-3104

semantic networks. Here a syntactic analysis of the query is done by parsing the query using Stanford parser to tag each and every word with their corresponding parts of speech.
Semantic Information Retrieval System [12] is mainly concerned with retrieving information from a sports ontology using the SPARQL query language. Here specific information is retrieved from the ontology. The sports related information is queried from the ontology and it is done using SPARQL language. A Relation-Based Page Rank Algorithm for Semantic Web Search Engines [13] proves that relations among concepts embedded into semantic annotations can be effectively exploited to define a ranking strategy for Semantic Web search engines. This sort of ranking behaves at an inner level and can be used in conjunction with other established ranking strategies to further improve the accuracy of query results. The overview of the existing systems gives multitude approaches for semantic information extraction. Though these above systems perform a semantic analysis, it has been implemented in a more generic way. Hence in order to further enrich this process to retrieve more promising results a system has been proposed for queries relating to university domain (SIEU).

METHODS

In this work, firstly, we review and investigate different measures for semantic similarity computation. Then, propose a new measure considering the multiple inheritances in ontologies and the common specificity feature of the evaluated concepts in order to obtain a more accurate similarity between concepts. For measured semantic similarity score values, rules are formed based on fuzzy function. Created fuzzy rules then optimized by *Modified Genetic Algorithm (MGA).* Finally, evaluate the proposed DLCS-FRMGA approach using two datasets of biomedical term pairs scored for similarity by human experts and exploiting SNOMED CT as the input ontology. Compare the correlation obtained by our measure with human scores against other measures. The experimental evaluations confirm the efficiency of the proposed measure.

Similarity Measurement

Here concentrated on the Path [14], *Leacock &Chodorow (LCH)* [15], and Wu & Palmer [16] path finding measures that are dependent on the shortest path separating concepts. Consider $p = \text{path}(C_1, C_2)$, the quantity of nodes in the shortest path splitting two concepts, C_1 and C_2. The shortest path among two concepts navigates their Least Common Subsumer $(\text{lcs}(C_1, C_2))$, i.e. their nearby common parent. The depth $(\text{depth}(c))$ of a concept is described as the amount of nodes in the pathway to the root of the taxonomy; and d indicates the maximum depth of taxonomy. Path describes the similarity among two concepts basically as the converse of the extent of the path separating them.

One major complication with this description is that the similarity of a concept with itself is below 1 $(\text{if } C_1 = C_2, \text{then } \text{path}(C_1, \text{lcs}(C_1, C_2)) + \text{path}(c2, \text{lcs}(C_1, C_2)) = 2)$. As an alternative, here adopted the definition of Wu & Palmer employed in the Natural Language Toolkit :
One major complication with this description is that the similarity of a concept with itself is below 1 $(\text{if } C_1 = C_2, \text{then } \text{path}(C_1, \text{lcs}(C_1, C_2)) + \text{path}(c2, \text{lcs}(C_1, C_2)) = 2)$. As an alternative, here adopted the definition of Wu & Palmer employed in the Natural Language Toolkit :

$$\text{sim}_{wp}(C_1, C_2) = \frac{2 \times \text{depth}(\text{lcs}(C_1, C_2))}{p - 1 + 2 \times \text{depth}(\text{lcs}(C_1, C_2))} \quad (1)$$

Based on this, $\text{if } C_1 = C_2, \text{then } p - 1 = 0$, and the similarity measure evaluates to 1.

Measures based on Information Content

Resnik [17] formulated to balance the taxonomical arrangement of ontology with the information distribution of concepts assessed in input corpora. Here also utilized the notion of IC, by means of associating appearance probabilities to every concept in the taxonomy, calculated from their occurrences in a particular corpus. IC of a term a is calculated in accordance with the negative log of its probability of occurrence, $p(a)$ (5). In this way, uncommon words are considered more useful than common ones.

$$IC(a) = -\log P(a) \quad (2)$$

This procedure is typically done manually in order to guarantee the appropriateness of the tagging, applicability and hampering the scalability of this scheme with huge corpora. Furthermore, when either the taxonomy or the corpus transformations, re-computations are required to be recursively implemented for the affected concepts. As a result, it is essential to carry out a manual and time consuming investigation of corpora and resultant probabilities would depend on the size and temperament of input corpora. By taking the drawbacks of IC-based schemes because of their dependency on corpora, a few authors attempted to intrinsically derive IC values from ontology. These works depends on the supposition that the taxonomic arrangement of ontologies like WordNet is put in order in a significant manner, in accordance with the rule of cognitive saliency [18]. This confirms that humans specialized concepts when they require distinguishing them from previously existing ones. As a result, concepts with several hyponyms (i.e., specializations) are extremely common and offer fewer details than the concepts in the place of leaves of the hierarchy. Seco et al., [19] and Pirró and Seco [20] based IC computations on the amount of hyponyms. In view of the fact that, $\text{hypo}(a)$ the amount of hyponyms of the concept a and max_nodes the amount of hyponyms of the root node, they calculate IC of a concept in the following manner (10):

THE IIOAB JOURNAL
ISSN: 0976-3104

$$IC_{seco}(a) = 1 - \frac{\log(hupo(a) + 1)}{\log(max\_nodes)} \quad (3)$$

The denominator guarantees that IC values are normalized in the limit [0...1]. This scheme only takes hyponyms of a particular concept in the taxonomy; as a result, concepts with the similar number of hyponyms however different degrees of generality come out to be equally comparable. With the aim of dealing with this complication effectively, and in the similar manner as for edge-counting measures, Zhou et al., [21] formulated to balance hyponym-based IC computation with the associated deepness of each concept in the taxonomy. The IC of a concept is found as given below:

$$IC_{zhou}(a) = k\left(1 - \frac{\log(hupo(a) + 1)}{\log(max\_nodes)}\right) + (1 - k)\left(\frac{\log(depth(a))}{\log(max\_depth)}\right) \quad (4)$$

Besides $hypo$ and max_nodes, which has the similar meaning as eq. 20, $depth(a)$ corresponds to the deepness of the concept a in the taxonomy and max_depth indicates the maximum deepness of the taxonomy. The factor k fine-tunes the weight of the two features engaged in the IC assessment. Here used $k = 0.5$.

Context Vector Measures of Semantic Relatedness

Furthermore, the WordNet glosses can be considered as a corpus of contexts comprising of around 1.4 million words. Subsequently, the gloss vector measure got the maximum correlation concerning human judgment by means of different benchmarks.

Gloss vectors for all concepts in WordNet can be computed in this manner. The relatedness of two concepts is then determined as the cosine of the normalized gloss vectors corresponding to the two concepts:

$$related_{vector}(c_1, c_2) = \cos(angle(\overrightarrow{v_1}, \overrightarrow{v_2})) \quad (5)$$

Where c_1 and c_2 are the two given concepts, $\overrightarrow{v_1}$ and $\overrightarrow{v_2}$ are the gloss vectors corresponding to the concepts and $angle$ returns the angle between vectors. Using vector products, the above relatedness formula can be rewritten as:

$$related_{vector}(c_1, c_2) = \frac{\overrightarrow{v_1}.\overrightarrow{v_2}}{|v_1||v_2|} \quad (6)$$

The measure of semantic relatedness based on WordNet and MeSH glosses, which is enhanced with information from a large corpus of text.

Proposed Similarity Measurements

Here, the most common similarity among two concepts (or two text nodes) C_1 and C_2 indicates a weighted sum of the similarities of the two features among them, i.e.:

$$Sim(C_1, C_2) = w_1 * (Sim\,C(c_1, c_2) + w2 * SimP(c_1, c_2)) \quad (7)$$

Data Content Similarity (SimC)

It is the Cosine similarity among the term frequency vectors of C_1 and C_2:

$$simC(c_1, c_2) = \frac{V_{c_1} \cdot V_{c_2}}{|V_{c_1}| * |V_{c_2}|} \quad (8)$$

Where V_d indicates the frequency vector of the terms within the concept C, $||V_d||$ indicates the length of V_d, and the numerator is the inner product of two vectors.

Presentation Style Similarity (SimP)

It indicates the average of the style Feature Scores (FS) over the entire six presentation style Features (F) among C_1 and C_2:

$$simP(c_1, c_2) = \sum_{i=1}^{6} FS_i / 6 \quad (9)$$

Where FS_i is the score of the ith style feature and it is defined by $FS_i = 1\, F_i^{c_1} = F_i^{c_2}$ and $FS_i = 0$ otherwise, and is the ith style feature of data unit d.

Common Specificity Feature

Here, a new modified measure for the purpose of semantic similarity by means of combining both the super concepts of the assessed concepts and common specificity feature which can confine further semantic indication. This measure can accomplish better performance than other measures, since it is completely based on structure and continue their simplicity. Consider c_i stand for ith concept of ontology. Subsequently, $N(c_i)$ is defined as the collection of the entire super concepts of c_i including c_i itself. As a result, the number of non-common super concepts of concepts can be determined as follows:

$$Noncomsub(c_1, c_2) = |N(c_1) \cup N(c_2)| - |N(c_1) \cap N(c_2)| \quad (10)$$

THE IIOAB JOURNAL
ISSN: 0976-3104

At this point, the NonComSub value can be a sign of the path length of the two concepts. LCS node of two concepts C_1 and C_2 determines the common specificity of C_1 and C_2 in the cluster. So the specificity of two concepts is computed through the process of finding the deepness of their LCS node and subsequently scaling this deepness as follows:

$$\mathrm{ComSpec}(c_1,c_2) = D - \mathrm{depth}(\mathrm{LCS}(c_1,c_2)) \quad (11)$$

Where D indicates the deepness of the ontology and the ComSpec feature decides the common specificity of two assessed concepts. The lesser the ComSpec value of two concepts, the additional details they share, and as a result the more comparable they are. As a result, the semantic distance among concepts c_1 and c_2 is given as follows:

$$\mathrm{SemD}(c1,c2) = \log(\mathrm{NonComSub} \times \mathrm{ComSpec} + 1) \quad (12)$$

It is to be mentioned that whichever concept can be evaluated with itself.

Footrule Similarity

In this proposed similarity is based on the footrule similarity. The expansion of the Spearman footrule distance by means of including the result of a similarity function described on concepts in U. This measure is described in concepts of a generic similarity function. In actual fact, the option of a specific similarity measure for domain is of huge significance, as well as the progress of similarity measures is an active area in the data mining and pattern recognition communities. At this point, take the similarity measure as specified, and presume that the similarity scores returned are significant inside the particular domain. The footrule similarity distance among two (maybe partial) ranked lists σ and r, provided a similarity function $s(.,.)$, is given as:

$$F_{sim}(\sigma,r,s(c_1,c_2)) = \sum_{C_1 \in \sigma_{\lambda_1 r}} \sum_{C_2 \in r_{\lambda_1 \sigma}} \mathrm{sim}(C_1,C_2)|\sigma_{\lambda_1 r}(C_1) - r_{\lambda_1 \sigma}(C_2)| \quad (13)$$

Specifically, the footrule similarity distance is computed on similarity projections of σ and r. The difference in ranks for items in these resultant lists is weighted through the power of the similarity.
This measure can be applied to the entire concepts. Subsequently, the most common similarity between these semantic similarity and footrule similarity is a weighted sum of the similarities:

$$\mathrm{Sim}(C_1,C_2) = w_1 * (\mathrm{SemD}(c1,c2) + w2 * F_{sim}(\sigma,r,s(c_1,c_2))) \quad (14)$$

Fuzzy Rules for Semantic Similarity Score Evaluation

Let us introduce some precise definitions of what is meant by the rule base solution representation. First of all here given L linguistic variables that is L=3 ontologies, MeSH , SNOMED and MeSH and SNOMED $\{A^1, ..., A^L\}$ and their semantic score value for concept 1 to concept n. Each linguistic variable A_i has M_i linguistic descriptions $\{A_1^i, \ldots, A_{M_i}^i\}$ that arerepresented by triangular membership functions $\mu_j^i, j = 1, \ldots, M_i$. A fuzzy rulehas the form [22]: $if A^{i_1}$ is $A_{j_1}^{i_1}$ and A^{i_2} is $A_{j_2}^{i_2}$ and A^{i_k} is $A_{j_2}^{i_k}$ then concept C
Where $i_1, \ldots i_k \in \{1, \ldots L\}, j_k \in \{1, \ldots M_{i_k}\}$and $o \in [0,1]$. A rule base is a set of several rules. Let us assume that we are given a rulebase consisting of n rules: $if A^{i_1^1}$ is $A_{j_1}^{i_1^1}$ and $A^{i_2^1}$ is $A_{j_2}^{i_2^1}$ and $A^{i_k^1}$ is $A_{j_2}^{i_k^1}$ then concept c_1
$if A^{i_1^2}$ is $A_{j_1}^{i_1^2}$ and $A^{i_2^2}$ is $A_{j_2}^{i_2^2}$ and $A^{i_k^2}$ is $A_{j_2}^{i_k^2}$ then concept c_2
$if A^{i_1^m}$ is $A_{j_1}^{i_1^m}$ and $A^{i_2^m}$ is $A_{j_2}^{i_2^m}$ and $A^{i_k^m}$ is $A_{j_2}^{i_k^m}$ then concept c_n
where $i_1^m \in \{1, \ldots, L\}$and $j_1^m \in \{1, \ldots, M_{i_1^m}\}$. Given a semantic vector from the multiple ontologies $sev \in \mathbb{R}^L$of observed semantic values, whose components are values for the linguistic variables A^1 , ..., A^L, can evaluate the rule base as follows [23]: the function ρ describes the way the rule base interprets semantic values observations $sev$to produce a single output value that is selected semantic vector belongs to concept one or not . However the above mentioned rules only satisfied to semantic measurement of the single attribute only so this can be changed as follows,

$if A^{i_1^1}$ is $A_{j_1}^{i_1^1}$ and $B^{i_2^1}$ is $A_{j_2}^{i_2^1}$ and $B^{i_k^1}$ is $A_{j_2}^{i_k^1}$ then concept $c_1 \& c_2$

$if A^{i_1^2}$ is $A_{j_1}^{i_1^2}$ and $B^{i_2^2}$ is $B_{j_2}^{i_2^2}$ and $B^{i_k^2}$ is $B_{j_2}^{i_k^2}$ then concept $c_2 \& c_2 or c_2 \& c_1$

$if A^{i_1^m}$ is $A_{j_1}^{i_1^m}$ and $B^{i_2^m}$ is $B_{j_2}^{i_2^m}$ and $A^{i_k^m}$ is $A_{j_2}^{i_k^m}$ then concept $c_n \& c_{n-1}$ or $c_n * c_{n-1}$

This value has an application specific meaning and can be taken to be a real number. More precisely, $\rho: \mathbb{R}^L \to \mathbb{R}$ is defined as follows:

$$sev = \begin{pmatrix} sev^1 \\ sev^2 \\ \vdots \\ sev^L \end{pmatrix} \to \frac{\sum_{m=1}^{n} we(c^m) \prod_{j=1}^{k_m} \mu_{j_1^m}^{i_1^m}(x^{i_1^m})}{\sum_{m=1}^{n} c^m} \quad (15)$$

Evaluation Function

Consider an evaluation function (to minimize) that measures the minimum difference values between ontologies semantic values and as well as their concepts when training a rule base to fit a given biomedical dataset samples. This observed semantic values consists of a set $\{sev_i, y_i\}_{i=1,..N}$, where each $sev_i = \begin{pmatrix} sev_i^1 \\ sev_i^2 \\ \vdots \\ sev_i^L \end{pmatrix}$

THE IIOAB JOURNAL
ISSN: 0976-3104

is a vector that has as many components as there are linguistic variables, i.e. $sev_i \in \mathbb{R}^L \forall i = 1,\dots,N$, and each y_i is a real number, i.e. $y_i \in \mathbb{R} \forall i = 1,\dots,N$. Then the evaluation function has the form.

$$diff = \sum_{i=1}^{N} (\rho(sev_i) - y_i)^2 = \sum_{i=1}^{N} \left(\frac{\sum_{j=1}^{n} a_{ij}\, we(c^j)}{\sum_{j=1}^{n} we(c^j)} - y_i \right)^2 \quad (16)$$

Where $a_{sm} = \prod_{l=1}^{k_m} \mu_{j_l}^{q_m}\left(x_s^{q_m}\right)$ (17)

The major aim of this work is to optimize the rules base in such a way that the evaluation function $diff$ becomes minimal. This involves two separate problems. Firstly, the form of the membership functions μ_j^i may be varied to obtain a better result. Secondly, the rule base may be varied by choosing different rules or by varying the weights $we(c_i)$. In this paper we consider both problems as important however for rule optimization varying the weights $we(c_i)$ is determined via the use of the Modified Genetic Algorithm (MGA). In the MGA the difference between the two ontologies semantic meaning is less means then weight becomes very high or it becomes less, if the difference is less than those concepts are similar and those rules are considered as major important rule in fuzzy theory. Concentrate on the second problem, taking the form of the membership functions to be fixed. For example, we can standardize the number of membership functions for each linguistic variable A^i to be $M_i = 2n_i - 1$ and define

$$\mu_j^i = \begin{cases} 0 : sev \le \dfrac{j-1}{2n_i} \\ 2n_i sev + 1 - j : sev \in \left[\dfrac{j-1}{2n_i}, \dfrac{j}{2n_i}\right] \\ -2n_i sev + 1 - j : sev \in \left[\dfrac{j}{2n_i}, \dfrac{j+1}{2n_i}\right] \\ 2n_i sev + 1 - j : sev \le \dfrac{j-1}{2n_i} \end{cases} \quad (18)$$

for $j = 1,..2n_i - 1 = M_i$

Search Space

The search space is the set of all potential rule base solutions. Let us first of all compute the maximum number of rules n_{max} then each rule can be written in the form

If $A^1 is A_{j1}^1 and\ A^2 is A_{j2}^1$…. And $A^L is A_{jL}^L$ then concept

where in this case $j_i \in \{0, 1, \dots, M_i\}$ and $j_i = 0$ implies that the linguistic variable A_i does not appear in the rule. Then we have

$$n_{max} = (M_1 + 1) \times (M_2 + 1) \dots \times (M_L + 1) - 1 \quad (23)$$

Note that we have subtracted 1 to exclude the empty rule. If include the possible choices of weights $we(c_i)$ with discretization $we(c_i) \in \{0, \frac{1}{d}, \dots, 1\}$, then we have a system of $(d+1)^{n_{max}}$.

Modified Genetic Algorithm (MGA)

MGA possesses a structure similar to GA. However, the MGA [24] has been distinguished from the GA in that the reproduction of the optimized rules is processed after the completion of both the crossover and mutation. Thus the deterioration problem never happens since the best solution from the current generation will be superior to or at least the same with the past. However, in the recent work, rules typically consist of diverse sets of classes and properties. In addition, this rule base analysis does not execute the rules. So the measurement of semantic similarity between concepts requires much time to complete the task, to solve this problem grouping is introduced in the next approach.

In the beginning the MGA creates an initial population based on the semantic score values. In the next step the algorithm evaluate the objective values (difference value of the semantic score between two ontologies in the multiple ontology stage) of the current population. After that weights are reproduced. During the reproduction, crossover first occurs. New Weights from rules combine to form a whole new rule. The newly created weight values of the concepts then mutates. Mutation means that the elements of chromosome are a bit changed. These changes are mainly caused by errors in copying weights from fuzzy rule base. Then MGA ranked weights represented by their associated cost, to be "minimized", and returns the corresponding individual fitness. Next the most fitted weights from fuzzy rules are selected. Here the objective values of the fuzzy rules in the offspring are evaluated and re-insertion of fuzzy rules in rule phase replacing parents is done. The MGA is terminated when some criteria are satisfied, e.g. a certain number of generations, a mean deviation in the population, or when a particular point in the search space is encountered.

THE IIOAB JOURNAL
ISSN: 0976-3104

RESUTLS

There are no standard human rating datasets for semantic similarity in biomedical domain. Carried out an experiment similar to the one proposed. Two ontologies are used as background knowledge: WordNet and MeSH. WordNet is a lexical database that describes and structures more than 100,000 general concepts, which are semantically structured in an ontological way. The *Medical Subject Headings (MeSH)* contains a hierarchy of medical and biological terms defined by the U.S National Library of Medicine. Use the benchmarks of Hliaoutakis et al.[25] and Pedersen et al.[26] to evaluate our methods. For the first one, we have taken the ratings given by the 3 physicians, 9 medical coders and the average of both. In this manner, a suitable LCS between the two ontologies must be discovered to enable the similarity assessment. The performance of proposed Fuzzy Rulebase Modified Genetic Algorithm (DLCS-FRMGA) based similarity measure evaluation is compared to existing higher similarity method of Al Fengq in Yang et al and Rule Based Semantic Score Evaluation (RSSE).Precision measure is calculated based on the formula

$$\text{Precision} = \frac{T_p}{(T_p + F_p)} \quad (24)$$

Recall is calculated based on the formula

$$\text{Recall} = \frac{T_p}{(T_p + F_n)} \quad (25)$$

Accuracy is calculated based on the formula

$$\text{Accuracy} = \frac{T_p}{(T_p + F_p + F_n)} \quad (26)$$

Where T_p – True Positive (Correct result), T_N – True Negative (Correct absence of result), F_p– False Positive (Unexpected Result) , F_n– False Negative (Missing result).
F-Measure is calculated based on the formula

$$F=2.\frac{\text{precision}.\text{recall}}{\text{prrecision}+\text{recall}} \quad (27)$$

The simulation results for the evaluation of the proposed approach against various performance measures like Precision, Recall, Accuracy and F-Measure.

Accuracy: The proposed similarity measure evaluation using Wordnet Fuzzy Rulebase Modified Genetic Algorithm (DLCS-FRMGA) produced better accuracy rate shown in **[Figure -1]** in which When the number of concepts increases the accuracy of the result is increases.

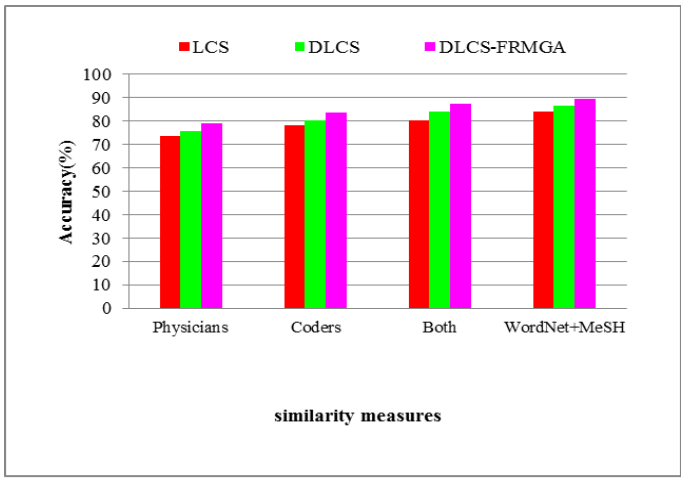

Fig: 1. Accuracy comparison of similarity measures

THE IIOAB JOURNAL
ISSN: 0976-3104

Figure 1 shows the accuracy comparison results where proposed DLCS-FRMGA produces 89.28% accuracy value to WordNet+MeSH which is better When compared to existing semantic similarity measurements with increased percentage of 2-6% for accuracy parameter.

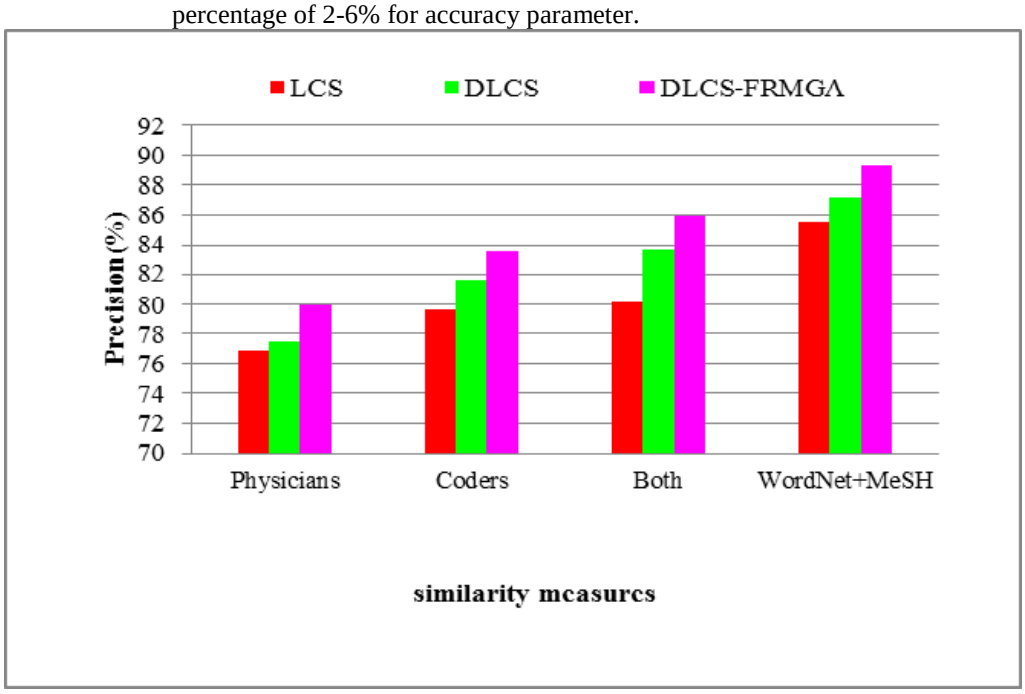

Fig: 2. precision comparison of similarity measures

[Figure- 2] shows the precision comparison results where proposed DLCS-FRMGA produces 89.28% precision value to WordNet+MeSH which performs better with increased percentage of 2-4%.

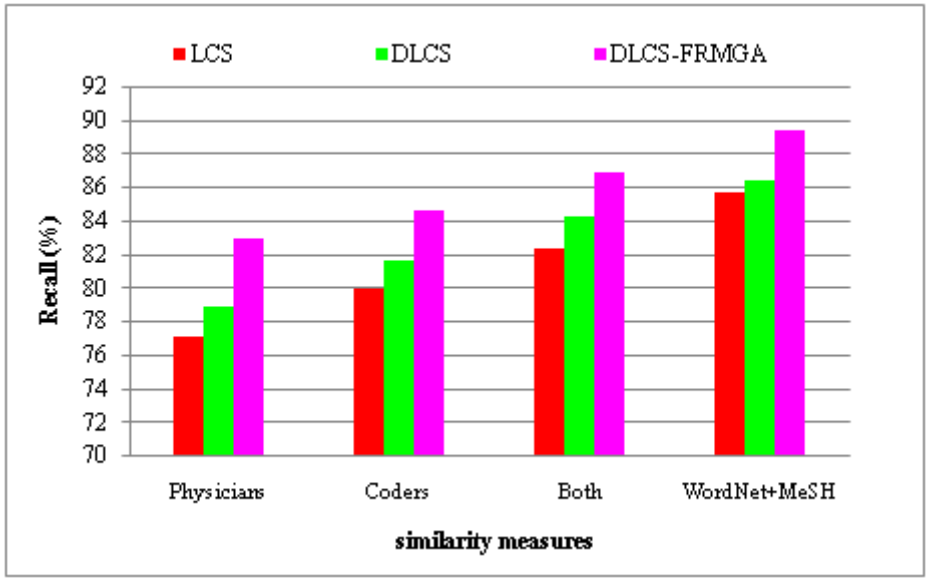

Fig: 3. Recall comparison of similarity measures

[Figure -3] shows the recall comparison results where the proposed DLCS-FRMGA produces 89.36% recall value which performs better with increased percentage of 2-4% for recall parameter.

THE IIOAB JOURNAL
ISSN: 0976-3104

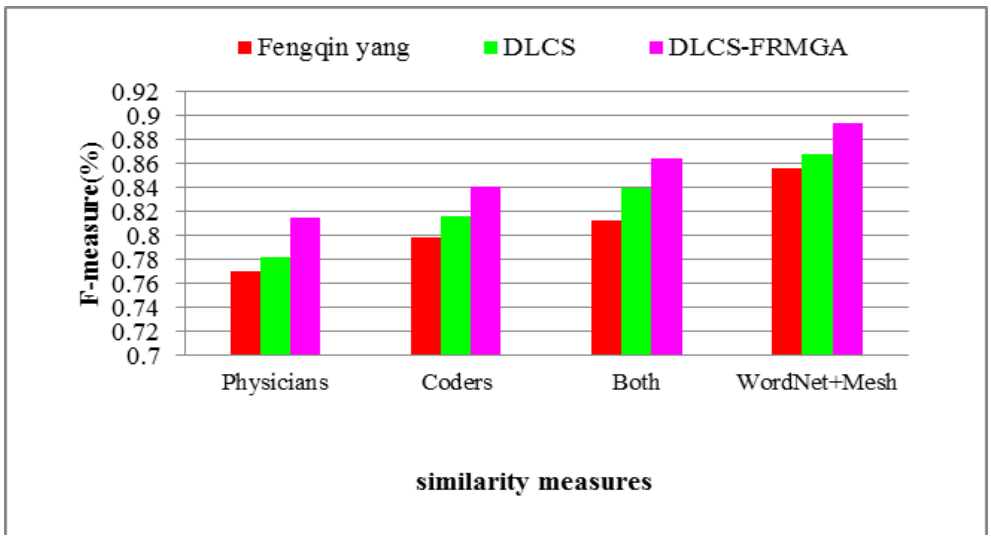

Fig: 4. F-measure comparison of similarity measures

The proposed similarity measure using Fuzzy Rule base Modified Genetic Algorithm (DLCS-FRMGA) produced high F1-measure shown in **[Figure -4]** which is higher than the existing similarity methods.

CONCLUSION AND FUTURE WORK

In the last few years, the amount of clinical data that is electronically available has increased rapidly. Digitized patient health records and the vast amount of medical and scientific documents in digital libraries have become valuable resources for clinical and translational research. With the intention of sorting them tranquil to understand and progress their applicability and precision, anticipate a Fuzzy Rulebase Modified Genetic Algorithm (DLCS-FRMGA) method deal with in fuzzy theory that permits the procedures deliberate to be consistently redefined. The common specificity feature considers the depth of the Least Common Subsumer (LCS) of two concepts and the depth of the ontology to obtain more semantic evidence. In addition the proposed DLCS-FRMGA method, some other features such as Data content and presentation features also extracted from biomedical text domain. For extracted features then data content similarity and Presentation style similarity is also measured. The weight of similarity between two concepts is calculated by using aggregation criteria. The similarity results of these are evaluated based on some rules and assumptions. For this purpose a DLCS-FRMGA approach is introduced to rule optimization phase. However rules in the rule generation phase is optimized using Modified Genetic Algorithm (MGA). Also anticipate calculating DLCS-FRMGA method in a accessible and effectual way from the taxonomical knowledge demonstrated in biomedical ontologies. As a outcome, new-fangled DLCS-FRMGA approach increases the semantic similarity measures expressed in terms of concept Information Content are presented. These measures are evaluated and compared to related works using a benchmark of medical terms and a standard biomedical ontology. The correlation amongst the outcomes of the estimated events and the human authorities' assessments demonstrates that DLCS-FRMGA method outstrips further most of the a fore mentioned events evading, roughly some of their confines.

CONFLICT OF INTEREST

The author declares no competing interest in relation to the work.

ACKNOWLEDGEMENT

Author is thankful to Dr. Radhakrishnan, Principal, Sasurie College of Engineering, Tamilnadu, India for his technical guidance in bringing out this paper efficiently. The author is also thankful to Dr.Makkal.G.Rajan, Chairman, Sri **Rajiv Gandhi polytechnic college**, Tamilnadu for providing moral support.

FINANCIAL DISCLOSURE

Nil

THE IIOAB JOURNAL
ISSN: 0976-3104

REFERENCES

[1] Jansen BJ, Soo YR, [2010]. The seventeen theoretical constructs of information searching and information retrieval. *Journal of the American Society for Information Science and Technology* 61(8): 1517-1534.

[2] Goodrum, Abby A, [2000]. Image information retrieval: An overview of current research. *Informing Science* 3(2): 63-66.

[3] Foote, Jonathan,[1999]. An overview of audio information retrieval. *Multimedia systems* 7(1): 2-10.

[4] Beel J, Gipp B, [2010]. Link analysis in mind maps: a new approach to determining document relatedness. *4th International Conference on Uniquitous Information Management and Communication*, pp.38.

[5] Baeza-Yates, Ricardo, William Bruce Frakes, eds. [1992]. Information Retrieval: Data Structures & Algorithms. *Prentice Hall*.

[6] Latzina, Markus, Anoshirwan Soltani, [2011]. Mixed initiative semantic search. *U.S. Patent* 7,987,176, No. 26.

[7] Gruber, Thomas R, [1993]. A translation approach to portable ontology specifications. *Knowledge acquisition* 5(2): 199-220.

[8] Fensel, Dieter, Ontologies, [2001]. In Ontologies, *Springer Berlin Heidelberg*. pp. 11-18.

[9] Antonio M, Rinaldi,[2009]. An Ontology-Driven Approach for Semantic Information Retrieval on the Web. *ACM Transactions on Internet Technologies* 9(3):10:1-10:24.

[10] Malik SK., Prakash N, Rizvi S.A.M. [2010]. Developing a University Ontology in Education Domain using Protégé for Semantic Web. *International Journal of Engineering Science and Technology* 2(9):4673-4681.

[11] Joel Booth, Barbara Di Eugenio, Isabel F, Cruz, Ouri Wolfson, [2009]. Query Sentences as Semantic (Sub) Networks, *IEEE International Conference on Semantic Computing*, pp.89-92.

[12] Jun Zhai, Kaitao Zhou, [2010]. Semantic Retrieval for Sports Information Based on Ontology and SPARQL, *International Conference of Information Science and Management Engineering (ICME)*. 19(2): 315-323.

[13] Fabrizio Lamberti, Andrea Sanna, Claudio Demartini, [2009]. A Relation-Based Page Rank Algorithm for Semantic Web Search Engines, *IEEE Transactions on Knowledge and Data Engineering* 21(1):123-136.

[14] Pedersen T, Pakhomov SVS, Patwardhan S, Chute CG. [2007]. Measures of semantic similarity and relatedness in the biomedical domain. *J Biomed Inform* 40:288–299.

[15] Leacock C, Chodorow M, [1998]. Combining local context with WordNet similarity for word sense identification, WordNet: A Lexical Reference System and its Application.

[16] Wu Z, Palmer M, [1994]. Verbs semantics and lexical selection, Association for Computational Linguistics, *Proceedings of the 32nd annual meeting on Association for Computational Linguistics, Las Cruces, New Mexico*, pp.133–138.

[17] Resnik P, [1995]. Using Information Content to Evalutate Semantic Similarity in a Taxonomy, In C.S. Mellish (Ed.), 14th International Joint Conference on Artificial Intelligence, *IJCAI Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc.*, 1: 448-453.

[18] Blank A, [2003]. Words and Concepts in Time: Towards Diachronic Cognitive Onomasiology, In R.Eckardt, K. von Heusinger& C. Schwarze (Eds.), Words and Concepts in Time: towards Diachronic Cognitive Onomasiology Berlin, Germany: Mouton de Gruyter, pp.37-66.

[19] Seco N, Veale T, Hayes J, [2004]. An Intrinsic Information Content Metric for Semantic Similarity in WordNet, In R. López de Mántaras& L. Saitta (Eds.), 16th Eureopean Conference on Artificial Intelligence, *ECAI 2004, including Prestigious Applicants of Intelligent Systems, PAIS Valencia, Spain: IOS Press*, pp.1089-1090.

[20] Pirró G, Seco N, [2008]. Design, Implementation and Evaluation of a New Semantic Similarity Metric Combining Features and Intrinsic Information Content, In R. Meersman& Z. Tari (Eds.), *OTM 2008 Confederated International Conferences CoopIS, DOA, GADA, IS, and ODBASE Monterrey, Mexico: Springer Berlin/Heidelberg*, 5332:1271-1288.

[21] Zhou Z, Wang Y, Gu J, [2008]. A New Model of Information Content for Semantic Similarity in WordNet, In S.S. Yau, C. Lee & Y.-C. Chung (Eds.), *Second International Conference on Future Generation Communication and Networking Symposia, FGCNS, Sanya, Hainan Island, China*, pp.85-89.

[22] JohnsonR, MelichM, MichalewiczZ, SchmidtM, [2005]. Coevolutionary optimization of fuzzy logic intelligence for strategic decision support. *IEEE Transactions on Evolutionary Computation* 9(6) :682–694.

[23] Esmin A, Lambert-TorresG,[2007]. Evolutionary computation based fuzzy membership functions optimization. *IEEE International Conference on Systems, Man and Cybernetics ISIC*, pp. 823–828.

[24] Roeva O,[2006]. A Modified Genetic Algorithm for a Parameter Identification of Fermentation Processes, Biotechnology & Biotechnological Equipment, 20, 1: 202-209.

[25] Hliaoutakis A, Varelas G, Voutsakis E, Petrakis E.G.M, Milios E.E, [2006] Information Retrieval by Semantic Similarity, International Journal on Semantic Web and Information Systems, 2: 55-73.

[26] Petrakis E.G.M, Varelas G, Hliaoutakis A, Raftopoulou P, [2006]. X-Similarity:Computing Semantic Similarity between Concepts from Different Ontologies. *Journal of Digital Information Management*, 4:233-237.