International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 04, April 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

RELATIONAL SIMILARITY SEARCHING APPROACH ON CHEMINFORMATIC DATASET

Anjali Raut[1], Savita Pansare[2], Renuka Hazari[3]

Computer, Genba Sopanrao College of Engineering Pune, (India)

ABSTRACT

Measuring the similarity between words, sentences, paragraphs and documents is an important component in various tasks such as information retrieval, document clustering, word-sense disambiguation, automatic essay scoring, short answer grading, machine translation and text summarization. This survey discusses the existing works on text similarity through partitioning them into three approaches; String-based, Corpus-based and Knowledge based similarities. Furthermore, samples of combination between these similarities are presented.

Text similarity measures play an increasingly important role in text related research and applications in tasks such as information retrieval, text classification, document clustering, topic detection, topic tracking, questions generation, question answering, essay scoring, short answer scoring, machine translation, text summarization and others. Finding similarity between words is a fundamental part of text similarity which is then used as a primary stage for sentence, paragraph and document similarities. Words can be similar in two ways lexically and semantically. Words are similar lexically if they have a similar character sequence. Words are similar semantically if they have the same thing, are opposite of each other, used in the same way, used in the same context and one is a type of another. Lexical similarity is introduced in this survey though different String-Based algorithms, Semantic similarity are introduced through Corpus-Based and Knowledge-Based algorithms. String-Based measures operate on string sequences and character composition. A string metric is a metric that measures similarity or dissimilarity (distance) between two text strings for approximate string matching or comparison. Corpus-Based similarity is a semantic similarity measure that determines the similarity between words according to information gained from large corpora. Knowledge-Based similarity is a semantic similarity measure that determines the degree of similarity between words using information derived from semantic networks. The most popular for each type will be presented briefly. This paper we presents We introduce 12 relational agreement (RA) coefficients for seven metric scales, which are integrated within a group fusion-based similarity searching algorithm. These similarity measures are compared to a reference panel of 21 proximity quantifiers over 3-5 benchmark data sets (UCI), by using informative descriptors, a feature selection stage, a suitable performance metric, and powerful comparison tests.

I. INTRODUCTION

The definition of similarity with respect to molecules is more stringent than that in other fields. Basically it consists of mapping “chemical space” (a representation of a molecule in structural or some property space) to one-dimensional space with entities of real numbers. Ideally similarity measures for molecules behave proportionally to all physical and biological properties of molecules in this representation. In other words, it

ISSN (O) 2454-9665
ISSN (P) 2454-0663

groups together all molecules with very similar physical and biological properties in a confined area of chemical property space. In practice, we are far away from reaching this goal. As we will see in the following paragraphs, molecular representations have to this day only been applied to specific problems of molecular similarity.

Similarity searches complement earlier substructure searches [1] which only consider presence or absence of specific features but did not evaluate global properties and overall shape. Compared to substructure searches, similarity searches are both more general and more comprehensive. They are more general by employing abstract representations of molecules or molecular properties and by being capable of using fuzzy matching techniques. Furthermore they are more comprehensive as they (usually) comprise features derived from the whole molecule under consideration. Molecular similarity calculations are done in three steps: representation of the molecules in descriptor space, feature selection, and comparison. The literature review in the following paragraphs will focus mainly on representation and comparison of molecules.

The group of spectra-derived descriptors uses a "natural" way to derive a one dimensional representation of a molecule. X-ray and electron diffraction as well as infrared spectra have been used in this sense. The resulting spectra have to be converted into descriptor space, e.g. by calculating its zero crossings. The earliest work in this area was done by Soltzberg [2], who used molecular transforms to calculate the diffraction pattern from an X-ray derived three dimensional structure.

A different approach [3] defines fuzzy peak areas to derive molecular features from an infrared spectrum, followed by principal component analysis. Although spectra are a "natural" way to convert a molecule into a one-dimensional representation, small changes often introduce major changes in the spectrum and the representation in descriptor space. These changes often make it difficult to use this approach as a similarity index.

Jain's Compass method [4] is able to take several molecules and several conformations into account, but it needs a user-defined interacting pharmacophore guess. This approach has also been used for selecting library subsets in its extension called Ice pick [Mount 1999], where several conformations of the molecules to be compared are calculated and the three dimensional

structures are docked into each other.

Jain [5] introduced the concept of "morphological similarity" which is defined as a Gaussian function of the differences in molecular surface distances of two molecules at weighted observation points on a uniform grid; compared to field-based methods, this method has the advantage that no alignment is necessary.

Similarity searching is based on the "Similar Property Principle" [6] that states that structurally similar molecules - structures with a "similar" spatial arrangement of "similar" functional groups - tend to have similar properties, physical as well as biological ones. All current drug design efforts are based on this paradigm.

According to the 2003 report by the Tufts Center for Drug Development [7], costs of a single new chemical compound until the point of submission to approval has risen to US$ 802 million. This is due to high failure rates in later stages of drug development. Probably the "easiest cherries have already been picked" – drugs for easily tractable targets have already been found. Furthermore it is well known that in vitro and in vivo screenings are very expensive, compared to so-called in silico approaches.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

A novel method for classifying similarity of molecules is performed by using hash keys of the molecular surface, compared to a panel of reference compounds [8]. Applied to several data sets, the description is found to capture enough information for the prediction of ADME properties

and target binding. Hash codes have already been applied in chemistry before [8], but only for structure storage and not for structure-activity relationships.

The last and most recent group of molecular descriptors are the back-projectable descriptors. Those descriptors can be projected back on the molecules that were used to derive the descriptor in the first place and often hint at points where molecules can be optimized with respect to bioactivity. The first back-projectable descriptor was published by Pastor and co-workers [9] and was called GRIND (Grid INdependent Descriptors). First, a set of simplified molecular interaction fields 13 around the probe molecule is calculated. Commonly, a hydrophobic probe (DRY), an oxygen probe (O) and a nitrogen probe (N1) are used to distinguish between hydrophobic, hydrogen bond donor and hydrogen bond acceptor properties, respectively. In the second step, an alignment-independent descriptor based on autocorrelation is calculated. Another descriptor that falls into this area is the MaP (Mapping Property distributions of molecular surfaces) descriptor [9]. This algorithm consists of three steps. Equally distributed surface points are computed first and then molecular properties are projected onto this surface. After that the

distribution of surface points and properties is encoded into a translational and rotationally invariant molecular descriptor which is based on radial distribution functions. An important feature of back-projectable descriptors is that they are easy to interpret.

Sybyl atom types [10] are employed for the derivation of the environments. These are force-field atom types, which implicitly include molecular properties such as geometry. An individual atom fingerprint is calculated for every atom in the molecule. This calculation is performed using distances from 0 up to n bonds and keeping count of the occurrences of the atom types. A count vector is constructed with the vector elements being counts of atom types at a given distance from the central atom. Every atom is described by exactly one count vector resulting in molecular atom environment fingerprints in which the number of atoms in a given molecule equals the number of count vector entries in the fingerprint.

II. HEADINGS

In the proposed research work to design and implement a system which is work base on feature similarity approach using algorithm on health care datasets. The can also having a ability to provide a high level security using AES encryptions scheme.

System first extract the different features from datasets then measure the similarity between objects using cosine base vector similarity and provide the matching results. We also focus on database security attack like SQL injection as well as prevention approaches.

III. RELATED WORK

Classification is most important technique in data mining. Using the classification system can converted unlabeled data into labeled data. By using the classifier system can label the similar data in one group. Anther

ISSN (O) 2454-9665
ISSN (P) 2454-0663

most concept in cloud computing is encryption technique for security. These both techniques are used in different field.

IV. FIGURES AND TABLES

Proposed system consists of three modules, such as user, dataset provider and third cloud. Here first query is fired by client in encrypted form is provided to cloud. Then cloud provides k closest records to the client in encrypted form. Dataset provider provides data to the cloud.

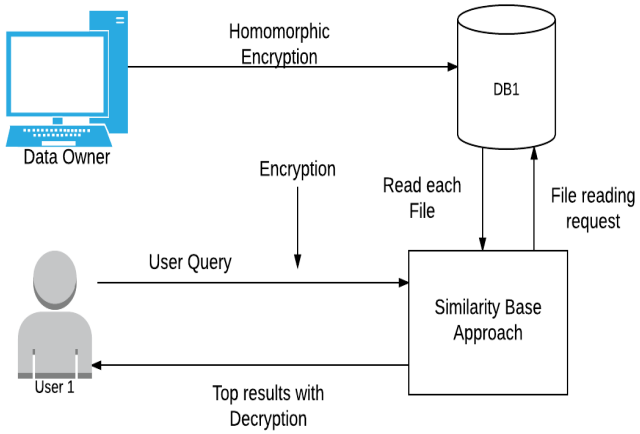

Figure 1: Proposed System Architecture

Let's,

Here, S is the all system module which holds the overall system

CS={C1,C2,C3…Cn}

C1= Authentication and key generation process.

C2= upload file with data encryption and data deduplication checking.

C3= search cipher data

C4 = decryption of data and file download.

C5= analysis graphs.

Doc={D1,D2,D3…….Dn} group of documents

Query ={Q1,Q2,Q3……..Qn} set of queries

DL={Doc1,Doc2…….Docn}

Here R is web base approach which handles the parallel searching, the result of query classified into n number of result pages.

Process state diagram:

International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 04, April 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

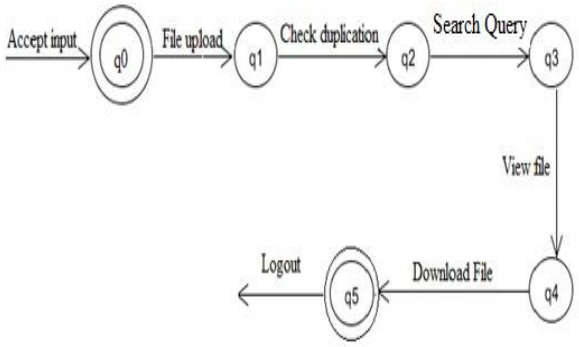

Figure 2: Process state diagram

Where,

qo is process of accepting input.

q1 is process of Uploading a file.

q2 is process of checking duplication.

q3 is process of modifying a file.

q4 is process to view a file.

q5 is process of downloading a file.

V. RESULTS AND DISCUSSION

For the system performance evaluation, calculate the matrices for accuracy. The system is implemented on java 3-tier MVC architecture framework with INTEL 3.0 GHz i5 processor and 8 GB RAM with public cloud Amazon EC2 consol. System also evaluated the computation costs of SkNNm for varying values of k, l and K. Throughout this sub-section, system fix m = 6 and n = 2000. However, system observed that the running time of SkNNm grows almost linearly with n and m.

The below tables 1 shows current system evaluation outcome

Approach	**Data Records**	**Times in Seconds**
Serial input records	**2000**	**35**
	4000	**68**
	6000	**102**
	8000	**132**
	1000	**171**

Table 1: Time Required for query processing when m = 6, k = 5 and K = 512

After the complete implementation of system evaluate with different experiments. For the second experiment system focus on time complexity of cryptography algorithm. The system take use different time for data encryption as well as data decryption purpose. The below figure 3 shows the encryption and decryption time complexity.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

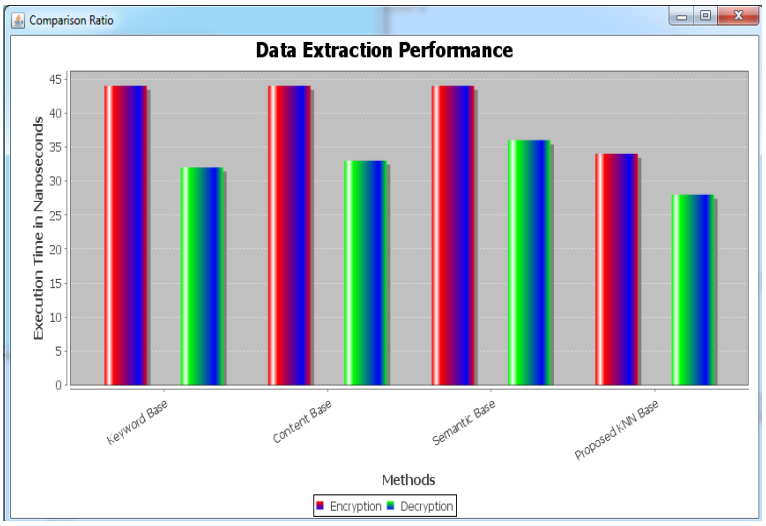

Figure 3: Data encryption and decryption performance with different approaches

VI. CONLUSION

To secure user privacy, numerous privacy-preserving category methods have been suggested over the past several years. The current methods are not appropriate to contracted database surroundings where the information exists in secured form on a third-party server. This paper suggested a novel privacy-preserving VCS classification protocol over secured information in the cloud. Our protocol defends the privacy of the information, user's input query, and conceals the information access patterns. System also analyzed the efficiency of our protocol under various parameter.

REFERENCES

[1] Hagadone, T.R. Molecular Substructure Searching: Efficient retrieval in Two Dimensional Structure Databases. J. Chem. Inf. Comput. Sci. 1992, 32, 515 – 521.

[2] Soltzberg, L.J.; Wilkins, C.L. Molecular Transforms: A Potential Tool . for Structure Activity Studies. J. Am. Chem. Soc. 1977, 99, 439 – 443.

[3] Schoonjans, V.; Questier, F.; Guo, Q.; van der Heyden, Y.; Massart, D.L. Assessing molecular imilarity/diversity of chemical structures by FT-IR spectroscopy. J. Pharm. Biomed. Anal.2001, 24, 613 – 627.

[4] Jain, A.N.; Koile, K.; Chapman, D. Compass: Predicting biological activities from molecular surface properties. Performance comparison on a steroid benchmark.. J. Med. Chem. 1994,37, 2315 – 2327.

[5] Jain, A.N. Morphological similarity: A 3D Molecular Similarity Method: Correlation with Protein-Ligang Interactions. J. Comput.-Aided Mol. Des. 2000, 14, 199 – 213.

[6] Concepts and Applications of Molecular Similarity; Johnson, A.M.; Maggiora, G.M., Ed.; Wiley: New York, 1990.

[7] DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G. The price of innovation: new estimates of drug development costs. J. Health Econ. 2003, 835, 1–35.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

[8] Ghuloum, A.M; Sage, C.R.; Jain, A.J. Molecular hashkeys: A novel method for molecular characterization and its application for predicting important pharmaceutical properties of molecules. J. Med. Chem. 1999, 42, 1739 – 1748.

[9] Stiefl, N.; Baumann K. Mapping Property Distributions of Molecular Surfaces: Algorithm and Evaluation of a Novel 3D Quantitative Structure-Activity Relationship Technique. J. Chem. Inf.Comput Sci. 2003, 46, 1390 – 1407.

[10] Clark, R.D.; Cramer, R.D.; Van Opdenbosch, N. Validation of the General Purpose Tripos 5.2 Force Field. J. Comp. Chem. 1989, 10, 982-1012.