ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 4, April 2018

A Survey on Classification and Analysis of Chemical Similarity Search in Large Database

Pradip A. Sarkate[1], Prof. A.V.Deorankar[2]

P.G. Student, Department of Computer Science & Engineering, Govt. College of Engineering, Amravati, India[1]

Associate Professor, Department of Computer Science & Engineering, Govt. College of Engineering, Amravati, India[2]

ABSTRACT:In a data mining similarity searching of a medical field is widespread problem of modern database application in chemical genomics, drug design and chemical research database. Search problem in chemical database are rarely based on exact matches rather than specific notion similarity. In data mining various graph kernel function designed to capture the intrinsic similarity of graphs. The graph kernel function used to similarity search for efficient and accurate similarity search in chemical database. Measure similarity of a graph represented a chemical structure utilize a new kernel function is G hash method. The g-hash method efficient and fast search method achieves k nearest neighbour classification.

KEYWORDS: chemical classification, G-hash function, graph kernel, k-NN search

I. Introduction

In biological similarity search in large chemical drug or medicine database are two techniques for search method 2D structure and 3D structure of biomolecules. 3D structure based approach is complement the long established facilities for 2D substructure searching. In 2D based structure similarity measurement may divided into two categories. Fragment based measure which used to searching chemical structure databases are extremely popular for prediction purposes. It identifies small chemical fragment of molecular structure used to similarity searching system. The fragment based is widely used to similarity measurement and adopted as a default choice in chemical database. Graph based method is utilize a graph model of chemical structure and utilize a different graph similarity measurement in a large chemical structure. It also used to graph edit distance and graph alignment structure algorithm which do not break the chemical structure into fragment and has started to gain popularity. 3D structure based approaches three dimensional shapes used molecular descriptors of chemical structure. Define query specification and docking search used to interaction of two chemical structures when the target is known.

The graph kernel function is used to G-hash method achieves state of the art performance for similarity search in chemical database. The G-hash method is used to bridge gap between graph kernel function and accurate similar search in large chemical database. In this method chemical structure represent two dimensional connecting graphs where node represent atoms and chemical bounds are edges between atoms. We extracting local feature of each node and their neighbour node in a graphs. We used to hash table that defined a graph kernel function to capture the important similarity of graph and fast similar query processing. The G-hash method used to retrieve k nearest neighbours node of graph.

II. Related work

Our work mainly involves two aspects: Substructure search is defining a graph component search which decomposing smaller pieces and chemical similarity search is searching chemical structure of graph, node and atoms, edges and chemical bounds.

DOI: 10.15680/IJIRCCE.2018.0604006

A. Substructure search:

In 2D structure graph which are chemical structures decomposing into smaller set pieces. Each piece as descriptor and building a descriptor based index structure of sub component graph query. The daylight fingerprints algorithms have all path up to fixed length are retrieved as descriptors [3]. Molecules are representing as bit string indexing by descriptors.
GDIndex [10] also used to include a hash table of sub graph for fast searching method. The main drawback of substructure component search is that no quantitative similarity measurement provided difficult to rank the search result.

B. Chemical similarity Search:

Elucidate the subcomponent strategy of a similar search in huge chemical database are non-trivial. Daylight fingerprint treats a chemical compound as a bit string use various similar metric for bit string [9].Fragment based method conventional similarity searching is appropriate identify complete structure [11]. Maximal common substructure utilize the similarity of target structure of chemical atoms properties to improve the similarity search algorithm, this method shows better accuracy prediction technique.
Graph alignment and graph edit distance used to measure of similarity between two graphs [11]. The graph are represent a chemical compounds are organic molecules that are commonly model by graph. In chemical structure the graph denoted nodes and edges are defines nodes are atoms and edges are chemical bonds in chemical structure.

III. PROPOSED SYSTEM

In this section, we proposed the graph kernel for similarity measurement of chemical database system. We used g-hash method defined similarity of chemical structures. The G-hash method achieves state of the performance for k nearest neighbour classification.

G hash function:

The G hash function introduces feature extraction process, and kernel function for similarity measurement.
In node feature extraction we will extract the features associated with a node and other local feature extraction are extract features in local region centred specific node.
Kernel function is used to better similarity measurement, the G hash method defined the similarity based on wavelet analysis and used a hash table as index structure to speed up the graph similarity. We used to following chemical structure to compute the similarity of chemical structures in structured matching kernel. Wavelet analysis method shows good definition of similarity between graphs through graph kernel function

Example of graph

K-NN query processing:

InkNNquery processing we calculated the distance graph query and all graph in databases. kNNquery processing based on tree structure kNNsearch performed by hashing function as per hash table and then kNN rank discover rank object by their distance to query point.

IV. CONCLUSION

Similarity of query processing of huge chemical database is critical issue, since we need to balance running time efficiency and better accuracy result.Similarity search in chemical database, we used graph kernel function to measure similarity of graph represented chemical. We utilize the G hash method of graph kernel function efficient and fast search. It will improve the performance of k nearest neighbour query processing algorithm. The key features for G hash method arekNN query time is scalable to large database and better accuracy in classification.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 4, April 2018

REFERENCES

1. T. Girke, L. Cheng, and N. Raikhel, "Chemmine. a compound mining database for chemical genomics," Plant Physiology, vol. 138, pp. 573–577,2005.

2. P. Willett, J. M. Barnard, and G. M. Downs, "Chemical similarity searching,'' *J. Chem. Inf.Comput. Sci.*, vol. 38, no. 6, pp. 983_996, 1998.

3. "Daylight fingerprints," 2008, software available at http://www.daylight.com.

4. Y. Cao, T. Jiang, and T. Girke, "A maximum common substructure based algorithm for searching and predicting drug-like compounds," Bioinformatics, vol. 24(13), pp. i366-74, 2008

5. X. H. Wang, A. Smalter, J. Huan, and G. H. Lushington, "G-hash: towards fast kernel-based similarity search in large graph databases." In Proc. 12th Int. Conf. EDBT., 2009, pp. 472-480.

6. Smalter, J. Huan, and G. Lushington, "Graph wavelet alignment kernel for drug virtual screening," in Proceedings of the 7th Annual International Conference on Computational System Bioinformatics, 2008.

7. D. Shasha, J. T. L. Wang, and R. Giugno, "Algorithmics and applications of tree and graph searching," in Proceeding of the ACM Symposium on Principles of Database Systems (PODS), 2002.

8. X. Yan, P. S. Yu, and J. Han, "Graph indexing based on discriminative frequent Structure analysis," in ACM Transactions on Database Systems (TODS), 2005.

9. R. Jorissen and M. Gilson, "Virtual screening of molecular databases using a support vector machine," J.Chem. Inf. Model., vol 45(3), pp.549-561, 2005.

10. D. Williams, J. Huan, and W. Wang, "Graph database indexing using structured graph decomposition," in Proceedings of the 23rd IEEE International Coference on Data Engineering (ICDE), 2007

11. J. Vert, "The optimal assignment kernel is not positive definite," French center for Computational Biology, Tech. Rep. HAL-002182778, 2008.

DOI: 10.15680/IJIRCCE.2018.0604006