International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307, Volume-1, Issue-3, July 2011

Survey of Finding Frequent Patterns in Graph Mining: Algorithms and Techniques

Vijender Singh, Deepak Garg

***Abstract*— Graphs become increasingly important in modeling complicated structures, such as circuits, images, chemical compounds, protein structures, biological networks, social networks, the web, workflows, and XML documents. Many graph search algorithms have been developed in chemical informatics, computer vision, video indexing and text retrieval with the increasing demand on the analysis of large amounts of structured data; graph mining has become an active and important theme in data mining.**

Index Termes: **Subgraphs, Graph Mining, gSpan**

I. INTRODUCTION

In mathematics, computer science and related subjects an algorithm is an effective method for solving a problem expressed as a finite sequence of instructions. Algorithms are used for calculation data processing and many other fields.

Meaning1. An algorithm operating on data that represent continuous quantities, even enough this data is represented by discrete approximation such algorithm are studied in numerical analysis.

Meaning2. An algorithm in the form of different equations that operates continuous on the data running an analog computer.

A. Apriori-Based Approach

Apriori based frequent substructure mining algorithm share similar characteristics with Apriori-based frequent item set mining algorithms. The search for frequent groups starts with graphs a small "size" and proceeds in a bottom-up manner by generating candidate having an extra vertex, edge or path. The definition of graph size depends on algorithm used.

B. Pattern-Growth Approach

The Apriori-based approach has to use the breadth-first search (BFS) strategy because of its level-wise candidate generation.

II. SURVEY OF TECHNIQUES AND ALGORITHMS

Various algorithms on graph mining were developed by many researchers. Some of them are reviewed in this section. Ullmann [1] in 1976 developed an algorithm for subgraph isomorphism. Subgraph isomorphism determined by means of a brute-force tree search procedure. This algorithm attains efficiency by inferentially eliminating successor's nodes in the tree search. Agarwal and Srikant [2] in 1994 considered the problem of discovering association rules between items in a large database of sales transaction. They presented two new algorithms for solving this problem that are fundamentally different from the known algorithm. Cook and Holder [14] in 1994 discovered a new version of their SUBDUE substructure discovery system is based on minimum description length principle. Holder, Cook and Djoko [3] in 1994 described the SUBDUE system which the minimum description length (MDL) principle is discovered substructures that compress the database and represent structural concepts in the data. In this paper they described the application of SUBDUE and also discussed the minimum description length principle and background knowledge used by SUBDUE can guide substructure discovery in a variety of domain. Blockeel and Raedt [6] in 1998 introduced a first-order framework for top-down induction of logical decision tree. Top-down induction of decision trees is the best known and most successful machine learning technique. It has been used solve numerous practical problems. It employs a divide-and conquers strategy, and in this it differs from its rule-based competitors which are based on covering strategies. Chakrabarti, Dom and Indyk [7] in 1998 developed a new method for automatically classifying hypertext into a given topic hierarchy, using an iterative relaxation algorithm. After bootstrapping off a text-based classifier, they used both local texts in a document as well as the distribution of the estimated classes of other documents in its neighborhood, to refine the class distribution of document being classified. They discussed three area of research: text and hypertext information retrieval, machine learning in context other text or hypertext, and computer vision and pattern recognition.

Inokuchi,Washio and Motoda [9] in 1998 proposed a novel approach name AGM to efficiently mine the association rule among the frequently appearing substructure in a given graph dataset. A graph is represented by adjacency matrices and the frequent patterns appearing in the matrices are mined through the extended algorithm of the basket analysis. Calders and Wisen [10] in 2001 Presented on monotone data mining layer a simple data-mining logic (DML) that can express common data mining tasks like "find Boolean association rules" or "Find inclusion dependencies". Kramer, Raedt, and Helma [11] in 2001 presented the application of feature mining techniques to the developmental therapeutics program's AIDS antiviral screen database. Kuramochi and Karypis [12] in 2001 presented a computationally efficient algorithm for finding all frequent subgraphs in large graph databases. They evaluated the performance of the algorithm by experiments with synthetic datasets as

Manuscript received June 4, 2011.

Vijender Singh, Department of Computer Science and Engineering, Thapar University, Patiala (Punjab), India, Mobile No. +91-9255074702, (e-mail: vijender_bhar@hotmail.com).

Dr. Deepak Garg, Professor, Department of Computer Science and Engineering, Thapar University, Patiala (Punjab), India (e-mail: deep108@yahoo.com).

well as a chemical compound dataset. Pei, Han, Mortazavi-Asl and Pinto [13] in 2001 proposed a novel sequential pattern mining method called PrefixSpan that is prefix-projected Sequential pattern mining.

Asai, Abe and Kawasoe [14] in 2002 discovered the efficient substructure from large semi structure data and patterns by labeled ordered trees and studied the problem of discovering all frequent tree-like patterns that have at least a minimum support in a given collection of semi-structured data. They presented an efficient pattern mining algorithm FREQT for discovering all frequent tree patterns from a large collection of labeled ordered tree. Borgelt and Berthold [15] in 2002 presented an algorithm to find fragments in a set of molecules that help to discriminate between different classes for instance, activity in a drug discovery context. Yan and Han [16] in 2002 investigated new approaches for frequent graph-based pattern mining in graph datasets and proposed a novel algorithm called gSpan. gSpan is a graph-based substructure pattern mining. This discovered frequent substructures without candidate generation. Zaki [17] in 2002 presented TREEMINER algorithm to discover all frequent subtrees in a forest, using a new data structure called scope-list.

Deshpande, Kuramochi and Karypis [18] in 2002 proposed the technique for classifying chemical compounds. These techniques can be broadly categorized into two groups. The first group consists of techniques that rely mainly on various global properties of the chemical compounds, such as molecular weight, ionization potential, inter-atomic distance etc. for capturing the structural properties of the compounds. Since this information is not relational, existing classification techniques can be easily used on these datasets. However the absence of actual structural information limits the accuracy of such classified. The second group of techniques directly analyzes the structure of the chemical compounds to identify patterns that can be used for classification [8; 5; 9; 11]. One of the earliest studies in discovering substructures was carried out by Dehaspe et al. [8] in 1998 which Inductive Logic Programming (ILP) techniques were used though this approach is quite powerful it is not designed to scale to large graph databases hence may not be able to handle large chemical compound databases. A number of recent approaches focused on analyzing the graph representation of the chemical compounds, to identify frequently occurring patterns, and use these patterns to aid in the classification. Wang et al. [5] in 1997 developed an algorithm to find frequently occurring blocks in the geometric representation of protein molecules and showed that these blocks can be used for classification. Inikuchi et al. [9] developed an algorithm to find all frequently occurring induced sub graphs and presented some evidence that such subgraph can be used to features for future classification.

Cooper and Frieze [19] in 2003 described a general model of a random graph process whose proportional degree sequence obeys a power law. These law recently observed in graph associated with www. Dzeroski [20] in 2003 introduced the Multi-Relational Data Mining. Getoor [21] in 2003 studied on link mining. Link among the objects may demonstrated certain patterns which can be helpful for many data mining tasks and are usually hard to capture with traditional statistical models. Link mining is promising new area where relational learning meets statistical modeling. Huan, wang and Prince [22] in 2003 proposed a novel subgraph mining algorithm: FFSM, which employs a vertical search scheme within an algebraic graph framework. They have developed to reduce the number of redundant candidates proposed. Their studied on synthetic and real datasets demonstrates that FFSM achieves a substantial performance gain over the current start-of-the art subgraph mining algorithm gSpan. Washio and Motoda [23] in 2003 introduced the theoretical basis of graph based data mining and surveys the state of the art of graph-based data mining.

Yan, Han and Afshar [24] in 2003 proposed an alternative but equally powerful solution: instead of mining the complete set of frequent subsequences, they mined frequent closed subsequences only that are those contained no super-sequence with the same support. They also introduced an efficient algorithm, called CloSpan. (CloSpan is stand for closed sequential pattern mining.) This outperforms the previous work by one order of magnitude. Moreover a deep a deep understanding of efficient sequential pattern mining methods may also have strong implications on the development of efficient methods for mining frequent subtrees, lattices, subgraphs, and other structured patterns in large databases. The sequential pattern mining algorithms developed so far have good performance in databases consisting of short frequent sequences. Yan and Han [25] in 2003 proposed to mine close frequent graph patterns. A graph g is closed in a database if there exists no proper subgraph of g that has the same support as g. A closed graph pattern mining algorithm, CloseGraph, is developed by exploring several interesting looping methods. Their performance studied shown that CloseGraph not only dramatically reduces unnecessary subgrapgs to be generated but also substantially increases the efficiency of mining, especially in the presence of large graph patterns. Yin and Han [26] in 2003 developed a new classification approach is called CPAR (CPAR is classification based on predictive Association Rules). Based on their study performance study, CPAR achieved high accuracy and efficiency, which have many useful features. CPAR represents a new approach towards efficient and high quality classification. It is interesting to further enhance the efficiency and scalability of this approach and compare it with other well-established classification schemes. Moreover, the strength of the derived predictive rules also motivates us to perform an in-depth study on alternative approaches towards effective association rule mining.

Huan, Wang, Prins and Yang [27] in 2004 developed a new algorithm that mines only maximal frequent subgraphs, that is subgraph that are not a part of any other frequent subgraphs. Their algorithm can achieve a five-fold speed up over the current state-of-the-art subgraph mining algorithms. Their mining method is based on a novel graph mining framework in which they first mine all frequent tree patterns from a graph database and then construct maximal frequent subgraphs from trees. Huan, Wang Bandyopadhyay Snoeyink, and Prins [28] in 2004 applied a novel subgraph mining algorithm mining algorithm to three related graph representation of the sequence and proximity characteristics

of a proteins amino acid residues. The subgraph mining algorithm is used to discover spatial motifs that can be used to discriminate among protein in different families found in the SCOP database. Koyuturk, Grama, Szpankowski, [29] in 2004 presented an innovative new algorithm for detecting frequently occurring patterns and modules in biological network. They show experimentally that their algorithm can extracted from the KEGG database within seconds. The proposed model and algorithm are applicable to a variety of biological networks either directly or with minor modification. Meinl, Borgelt and Berthold [30] in 2004 shown that is possible to mine meaningful, discriminative molecular fragments from large databases. Using an existing algorithm that employs a depth-first strategy and a sophisticated ordering scheme allows avoiding costly re-embeddings throughout the candidate growth process, which in turn enables us to find also larger fragments. Yin, Han, Yang and Yu [31] in 2004 developed a CrossMine, an efficient and scalable approach for multi-relational classification. Several novel methods are developed in CrossMine, including 1. tuple ID propagation, which performs semantics-preserving virtual join to achieve high efficiency on databases with complex schemas. 2. a selective sampling method which makes it highly scalable with respect to the number of tuples in the databases. Both theoretical backgrounds and implementation techniques of CrossMine are introduced. Yan, Yu and Han [32] in 2004discovered the issues of indexing graphs and proposed a novel solution by applying a graph mining technique. Different from the existing path-based methods, our approach, called gIndex, makes use of frequent substructure as the basic indexing feature. Frequent substructures are ideal candidates since they explore the intrinsic characteristics of the data and are relatively stable to database updates. gIndex has 10 times smaller index size, but achieves 3-10 times better performance in comparison with a typical path-based method.

Hu, Yan, Huang, Han and Zhou [33] in 2005 developed a novel algorithm, CODENSE, to efficiently mine frequent coherent dense subgraphs across large number of massive graph on biological networks for function discovery. Li and Tan [34] in 2005 proposed a novel graph mining algorithm to detect the dense neighborhoods in an interaction graph which may correspond to protein complexes. Their algorithm first located local cliques for each graph vertex and then merge the detected local cliques according to their affinity to form maximal dense regions. Yan, Yu and Han [35] in 2005 investigated the issues of substructure similarity search using indexed features in graph databases. By transforming the edge relaxation ratio of a query graph into the maximum allowed missing features, their structural filtering algorithm called Grafil, can filter many graphs without performing pairwise similarity computations. Yin, Han and Yu [36] in 2005 proposed a new approach, called CROSSCLUS, which performs cross-relational clustering with user's guidance. Yan, Zhou and Han [37] in 2005 developed two approaches to handle different mining requests: CLOSECUT, a pattern-growth approach, and SPLAT, a pattern-reduction approach. They applied these methods in biological datasets and found the discovered patterns interesting.

Chakrabarti and Faloutsos [38] in 2006 discussed the problem of detecting abnormalities in a given graph and of generating synthetic but realistic graphs have received considerable attention recently. Both are tightly coupled to the problem of finding the distinguished characteristics of real-world graphs i.e. the patterns that show up frequently in such graphs and can be considered as marks of realism. Karunaratne and Bostrom [39] in 2006 developed a method for learning from structured data are limited with respect to handling large isolated substructures and also imposed constraints on search depth and induced structure length. An approach to learning from structured data using a graph based canonical representation method of structured called finger printing. Krasky, Rohwer, Schroeder and Selzen [40] in 2006 discussed on a combined bioinformatics and chemoinformatics approach for the development of new ant parasitic drugs. Meinl, [41] in 2006 solved the problem Parallel molecular mining Worlein, Urzova, Fischer, and Philippsen which are used in chemoinformatics to find common molecular fragments a database of molecules represented as two-dimensional graphs. In ParMol package they have implemented four of the most popular frequent subgraph miners using a common infrastructure: MoFa, gspan, FFSM and Gaston. They also added additional functionality to some of the algorithms like parallel search, mining directed graphs and mining in one big graph instead of a graph database. Meinl, Worlein, Fischer, and Philippsen [42] in 2006 presented the thread-based parallel versions of MoFa and gSpan that achieve speedup up to 11 on a shared memory SMP system using 12 processors. They discussed the design space of the parallelization, the results, and the obstacles, that are caused by the irregular search space and by the current state of Java technology. Tsuda and Kudo [43] in 2006 proposed a graph clustering method that selects informative patterns at the same time. Their task is analogous to feature selection for vectors however the difference is that the features are not explicitly listed. This method is fully probabilistic adopting a binomial mixer model defined on a very high dimensional vector indicating the presence or absence of all possible patterns. Wegner, Frohlich, Mielenz and Zell [44] in 2006 presented a classification method which is based on a coordinate-free chemical space. Thus it does not depend on descriptor values commonly used coordinated-based chemical space methods.

Merkwirth and Ogorzalek [45] in 2007 described a method for construction of specific types of neural networks composed structures directly linked to the structure of the molecule under consideration. Each molecule can be represented by a unique neural connectivity problem (graph) which can be programmed on to a cellular neural network. A composite network can further successfully perform classification and regression on real-world chemical data-set. Dong, Gilbert, Guha, Heiland, Kim, Pierce, Fox, and Wild [46] in 2007 developed an infrastructure of chemoinformatics web service that simplifies that access to this information and the computational techniques that can be applied to it. They described this infrastructure and give some examples of its uses and then discuss their plans to use it as a platform for chemoinformatics application development in the future. Rhodes, Boyer, Kreulex, Chen, and Ordonez [47] in 2007

developed a system that allow user to explore the US patent corps using Molecular information. Their system contains three main technologies. In this system a user may go to a web page, draw a molecule search for related Intellectual property (IP) and analyzed the results. Bogdanov [48] in 2008 studied on Graph searching, indexing, mining and modeling for Bioinformatics, chemoinformatics and Social network. Fahim et al. [49] in 2008 proposed a method which is based on shifting the center of the large cluster toward the small cluster and recompiling the membership of small cluster points, the experimental results reveal that the proposed algorithm produce satisfactory results. Godeck and Lewis [50] in 2008 stated that QSAR models can play a vital role in both the opening phase and the endgame of lead optimization. In the opening phase there is often a large quantity of data from high throughput screening (HTS) and potential leads need to be selected from several distinct chemotypes.

Guha and Schurer [51] in 2008 investigated various aspects of developing computational models to predict cell toxicity based on cell proliferation screening data generated in the MLSCN. By capturing feature-based information in that data set, such predictive models would be useful in evaluating cell based screening results in general and could be used as an aid to identify and eliminate potentially undesired compounds.

Hubler, Kriegel, Borgwardt, Ghahramani and Metropolis [52] in 2008 presented metropolis algorithm for sampling a representative small subgraph from the original large graph with representative describing the requirement that the sample shall preserve crucial graph properties of the original graph. Karunaratne and Bostrom [53] in 2008 presented a case study in chemoinformatics in which various types of background knowledge are encoded in graphs that are given as input to a graph learner. In this paper shown that the type of background knowledge encoded indeed has an effect on the predictive performance. Lam and Chan [54] in 2008 studied on graph data mining algorithm are increasingly applied to biological graph data set. In this paper they proposed graph mining algorithm MIGDAC (Mining graph data for classification) that applies on graph theory and an interestingness measure to discover interesting sub graphs which can be both characterized and easily distinguished from other classes. Maji, and Mehta [55] in 2008 proposed a novel a measure of similarity between labeled graphs which has applications to structured data analysis for example chemical informatics, web document clustering etc. their metric on graphs exploits vertex context similarity and computes an over all matching score in polynomial time in the size of the graphs using a network flow formulation of the problem.

Tsuda and Kurihara [56] in 2008 proposed a nonparametric Bayesian method for clustering graph and selecting salient patterns at the same time. Variation inference is adopted here because sampling is not applicable due to extremely high dimensionality. Schietget, Costa, Ramon, and Raedt [57] in 2009 proposed a direct efficient and simple approach for generation of interesting graph pattern. They computed maximum common subgraph from randomly selected pairs of examples and directly use them as features. Jiang, Coenen and Zito [58] in 2010 examined a number of edge weighting schemes; and suggested three strategies for controlling candidate set generation. Spjuth, Willighagen, Guha, Eklund and Wikberg [59] in 2010 Studied on toward interoperable and reproducible QSAR analysis: Exchange of data sets. QSAR is widely used method to relate chemical structures to responses or properties based experimental observations. Much effort has been made to evaluate and validate the statistical modeling QSAR, but these analyses treat the dataset as fixed. An overlooked but highly important issue is the validation of the setup of the dataset, which comprises addition of chemical structure as well as selection of descriptors and software implementations prior to calculations. This process is hampered by the lack of standard and exchange formats in the field, making it virtually impossible to reduce and validate analyses and drastically constrain collaboration and re-use of data.

Yang, Parthasarthy and Sadayappan[60] in 2010 presented a novel approach to data representation for computing this kernel, particularly targeting sparce matrices representing power-law graphs. They shown their representation scheme, coupled with a novel tiling algorithm, can yield significant benefits over the current state of the art GPU and CPU efforts on a number of core data mining algorithms such as PageRank, HITS and Random Walk with Restart.

A graph transaction is represented by adjacency matrices and the frequent patterns appearing in matrices are mined through the extended algorithm.

These are modeled as attribute graph in which each vertex represents an atom and each edge a bond between atoms. Each vertex carries attribute that indicates the atom type.

III. Conclusion and Future Scope

The main challenge in the development of the algorithm is how to split up the discovery process into several phases efficiently. The algorithm should behave like a specialized free tree miner when faced with free tree databases, but should also be able to deal with graphs databases efficiently.Existing algorithm for frequent pattern mining become very costly in time and space as the pattern sizes and network number increase. Currently no efficient algorithm is available for mining recurrent patterns across large collection of genome wide network.There are various domains like chemoinformatics bioinformatics etc. where no efficient algorithms are available, for example, for mining recurrent patterns across large collection of genome-wide networks.

Due to increasing size and complexity of patterns in computer sciences the need for efficient graph mining algorithm is increasing. Still there is a scope of improvement in graph mining algorithm; the improvement can be in speed or sensitivity.

References

[1] J. R. Ullmann, “An algorithm for subgraph isomorphism”. J. ACM, 23, 1976, pp. 31-42.

[2] R. Agrawal, R. Srikant, “Fast Algorithms for mining association rules. In the proc. Of the 20th Int. conf. on very large databases (VLDB), 1994.

[3] D. J. Cook and L. B. Holder, “Substructure discovery using minimum description length and background knowledge” Journal of Artificial intelligence Research, 1, 1994, 231-255.

[4] Holder, L. B. Holder, Cook, D. J. Cook, Djoko, S. Djoko,

"Substructure Discovery in the SUBDUE system", In Proc. AAAI'94 Workshop knowledge Discovery in Databases (KDD'94), pp 169-180.

[5] X. wang, J.T.L. Wang, D. Shasha, B. Shapiro, S. Dikshitulu I.Rigoutsos, K. Zhang, "Automated discovery of active motifs in three dimensional molecules", In Proc. Of the 3rd int. conf. on knowledge discovery and data mining, 1997.

[6] H. Blockeel, L.D. Raedt, "Top-down induction of first-order logic decision trees", Artificial Intelligence, 101, 1998, pp. 285-297.

[7] S. Chakrabarti, B. Dom, P. Indyk, "Enhanced hypertext categorization using hyperlinks" ACM, (SIGMOD'98), 1998, pp. 307-318.

[8] L. Dehaspe, H. Toivonen, R. D. King, "Finding frequent substructures in chemical compounds". In 4th Int. conf. on knowledge Discovery and Data mining, 1998.

[9] A. Inokuchi, T. Washio, H. Motoda, "An Apriori-based Algorithm for Mining Frequent substructures from Graph Data. In proc. 2000 European Symp. Principle of Data mining and knowledge Discovery (PKDD'00), 1998, pp. 13-23.

[10] T. Calders, J. Wijsen, "On Monotone mining Languages", In proc. Of international workshop on database programming Languages(DBPL), 2001, pp. 119-132.

[11] S. Kramer, L.D. Raedt, C. Helma, "Molecular feature mining in HIV data", In Proc.ational conf. on of the 7th ACM SIGKDD International conf. on knowledge discovery and data mining, 2001, pp. 136-143

[12] M. Kuramochi, G. Karypis, " Freovequent Subgraph Discovery ", In Proc 2001 Int. conf. Data mining(ICDM'01).

[13] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, "PrefixSpan: Mining Sequential Pattern Growth." In proc. 2001 int. conf. Data Engineering (ICDE'01), 2001, pp. 215-224.

[14] T. Asai, K. Abe, S. Kawasoe, H. Arimura, H. Satamota, and S. Arikawa, "Efficient substructure discovery from large semi-structured data." In proc. 2002 SIAM Int. conf. Data mining(SDM'02), 2002, pp. 158-174.

[15] C. Borgelt and M.R. Berthold, "Mining molecular fragments: Finding relevant substructures of molecules." In Proc. 2002 int. conf. Data Mining (ICDM'02), PP. 211-218.

[16] X. Yan, and J. Han, "gSpan: Graph-Based Substructure Pattern Mining." In Proc. 2002 Int. conf. Data mining, 2002, pp. 721-724.

[17] M.J. Zaki, "Efficiently Mining Frequent trees in a forest." In Proc. 2002 ACM SIGKDD Int. conf. knowledge Discovery and Datamining(KDD'02), 2002, pp. 71-80.

[18] M.Deshpande, M. Kuramochi, G. Karypis, "Automated approaches for classifying structures." In Proc. 2002 workshop on Data mining in Bioinformatics (BIOKDD'02**),** 2002, pp. 11-18.

[19] C. Cooper and A. Frieze, "A general model of web graph." 2003.

[20] S. Dzeroski, "Multi-Relational Data mining: An Introduction" SIGKOD Explore. Newsl, 5(1), 2003, pp.1-16.

[21] L. Getoor, "Link Mining: A new data mining challenge." SIGKDD Explorations, (5), 2003, pp.84-89.

[22] J. Huan, W. Wang and J. Prins, "Efficient Mining of frequent Subgraph in the Presence of Isomorphism." In Proc. 2003 int. conf. Data mining (ICDM'03), 2003, pp. 549-552.

[23] T. Washio and H. Motoda, "State of the art of Graph- based data mining." SIGKDD Explorations, (5), 2003, pp. 59-68.

[24] X. Yan, J. Han and R. Afshar, "CloSpan: Mining Closed Sequential patterns in Large Datasets." In Proc. 2003 SIAM Int. conf. Data mining (SDM'03), 2003, pp. 166-177.

[25] X. Yan and J. Han CloseGraphs: Mining Closed Frequent Graph Patterns. In proc. 2003 ACM SIGKDD Int. conf. knowledge Discovery and Data Mining (KDD'03), 2003, pp. 286-295.

[26] X. Yin and J. Han, " CPAR: Classification based on Predictive Association Rules. In Proc. 2003 SIAM Int. conf. Data Mining (SDM'03), 2003, pp. 331-335.

[27] J. Huan, W. Wang, J. Prins and J. Yang,"Spin: mining maximal frequent subgraphs from graph Databases", KDD04 Seattle, Washington, USA, 2004.

[28] J. Huan, W. Wang, D Bandyopadhyay, J. Snoeyink, J. Prins and J. Tropsha," A. Mining Sapitial Motifs from Protein Structure Graphs. In Proc. 8th int. conf. Research in computational Molecular Biology (RECOMB), 2004, pp. 308-315.

[29] M. Koyuturk, A. Grama, and W. Szpankowski. "An Efficient algorithm for detecting frequent subgraphs in biological networks." Bioinformatics, (20), 2004, pp. i200-i207.

[30] T. Meinl, C. Borgelt and M.R. Berthold, " Discriminative closed fragment mining and perfect extensions in MoFa." 2004.

[31] X. Yin, J. Han, J.Yang and P.S. Yu, "CrossMine: Efficient Classification Across Multiple Database Relations. In Proc. 2004 int. conf. Data Engineering (ICDE'04), 2004, pp. 399-410.

[32] X. Yan, P.S. Yu, and J. Han. " Graph Indexing: A Frequent Structure-based approach." In Proc. 2004 ACM SIGKDD Int. conf. management of Data, 2004, pp.335-346.

[33] H. Hu, X. Yan, Y. Huang, J. Han and X. J. Zhou, "Mining coherent dense subgraphs across massive biological networks for functional discovery." In proc. 2005 Int. conf. Intelligent system for molecular Biology (ISMB'05), 2005, pp. 213-221.

[34] X.L. Li, S.H. Tan, "Interaction graph mining for protein complexes using local clique merging." Genome Informatics, 16(2),2005, pp. 260-269.

[35] X. Yan, P.S. Yu, J. Han, "Substructure similarity search in graph databases." In proc. 2005 ACM-SIGMOD Int. conf. Management of Data (SIGMOD'05), 2005, pp. 766-777.

[36] X. Yin, J. Han, P.S. Yu, "Cross-Relational Clustering with User's Guidance. In Proc. 2005 ACM SIGKDD Int. conf. knowledge Discovery and Data mining (KDD'05), 2005, 344-353.

[37] D. Chakrabarti, C. Faloutsos, "Graph mining: Laws, Generators, and Algorithm." ACM computing survey, 38(2), 2006, pp. 1-69.

[38] S. Maji, S. Mehta, "A Netflow distance between labeled graphs applications in chemoinformatics." www.cs.berkeley.edu., 2008.

[39] A. Krasky, A. Rower, J. Schroeder, P.M. Selzen,"A combined bioinformatics and chemoinformatics approach for the development of new antiparasitic drugs." Elsevier, genomics, 2006, pp.1-8.

[40] T. Meinl, M. Worlein, O. Urzova, , I. Fischer, M. Philippsen, "The ParMol package for frequent subgraph mining. Electronics communication of ESST 1, 2006.

[41] T. Meinl, M. Worlein, I. Fischer, M. Philippsen "Mining Molecular datasets on Symmetric Multiprocessor systems. 2006.

[42] Tsuda, K. Kudo, T. Clustering Graphs by weighted substructure mining. Proc. Of 23rd Int. conf. on machine learning, ACM, 148:953-960, 2006.

[43] J. K. Wegner, H. Frohlich, H.M. Mielenz, A. Zell, " Data and graph mining in chemical space for ADME and activity data sets." Wiley-VCH, 25(3),2006, 205-206.

[44] C. Merkwith, M. Ogorzalek, "Applying CNN to chemoinformatics." IEEE Xplore, 2007, 2918-2921.

[45] X. **D**ong, K.E. Gilbert, R. Guha, R. Heiland, J. Kim, M.E. Pierce, G.C. Fox, D.J. Wild, "Web Service Infrastructure for chemoinformatics." J. Chem. Inf. Model, 47, 2007, pp. 1303-1307.

[46] J. Rhodes, S. Boyer, J. Kreulex, Y. Chen, P. Ordonez, "Mining patents using molecular similarity search." Pacific symp. On Biocomputing, 12, 2007, pp. 304-315.

[47] P. Bogdanov, "Graph searching, indexing, mining and modeling for Bioinformatics, cheminformatics and Social network." 2008.

[48] A.M. Fahim, G. Saake, A.M. Salem, F. A. Torkey, M.A. Ramadan, "K-mens for spherical clusters with large variance in sizes." WorldAcademy of science, Engineering & Tech., 45, 2008, pp. 177-182.

[49] P. Godeck, R.A. Lewis, "Exploiting QSAR models in lead optimization." Curr. Opin. Drug Discov Devel, 11(4), 2008, pp. 569-575.

[50] R. Guha, S.C. Schurer, "Utilizing high throughput screening data for predictive toxicology models: Protocols and Application to MLSCN assays." J. comput. Aided Mol Des 22(6-7), 2008, pp. 367-384.

[51] Hubler, C. Kriegel, H.P. Borgwardt, K. Ghahramani, Z. Metropolis Algorithms for Representative Subgraph sampling. IEEEXplore, 2008, pp. 283-292.

[52] T.Karunaratne, H. Bostrom, "Using background knowledge for graph based learning: a case study in chemoinformatics." Springer, Artificial Inteligence, (6), 2008, pp. 151-153.

[53] W.W.M.Lam, K.C.C. Chan, "A Graph mining algorithm for classifying chemical compounds." IEEE Int. conf. on Bioinformatics and Biomedicine, 2008.

[54] S. Maji, S. Mehta, "A Netflow distance between labeled graphs applications in chemoinformatics." www.cs.berkeley.edu., 2008.

[55] K. Tsuda, K. Kurihara, "Graph Mining with variational Dirichlet process mixture models." SIAM, 432-442, 2008.

[56] L. Schietget, F. Costa, J. Ramon, L.D. Raedt, "Maximum common subgraph mining: A Fast and effective Approach towards feature generation." In Proc. At SRL-MLG-ILP, Leven, Belgium, 2009.

[57] C. Jiang, F. Coenen, M. Zito, M. "Frequent Sub-graph mining on Edge Weighted Graphs." 2010.

[58] O. Spjuth, E.L. Willighagen, R. Guha, M. Eklund, J.E.S. Wikberg, "Toward interoperable and reproducible QSAR analysis: Exchange of data sets." Journal of cheminformatics, 2:5, 2010.

[59] X. Yang, S. Parthasarthy, P. Sadayappan, " Fast Mining Algorithms of Graph data on GPUs." ACM, KDD-LDMTA'10, 2010.