www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 6 Issue 2 Feb. 2017, Page No. 20333-20349
Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i2.32

Optimized Chemical Bond Pattern Search Using Tree Indexing Technique

Sathya.S, Rajendran.N
Email : ssathyajai@gmail.com

Abstract

Probabilistic classifiers provide outputs to interpret conditional probabilities and distribution of classes given as input sample. User use sequence patterns to search for chemical formula and chemical names. Indexing of sequence patterns involving chemical formula identifies and index appearances of certain patterns for efficient search and retrieval. However, identifying chemical formulae has been a fundamental problem with increasing presence of formulae in any sequences. This is addressed through feature subset selection and indexing method in this work, called as Chemical Structured Bond Tree-based Indexing (CSBT-I). The algorithms in CSBT-I method are analyzed for different sequence patterns improving the chemical bond indexing accuracy using Bond Tree-based Structure and Sequential feature selection algorithm than existing methods. Bond tree based structure is created as a temporary indexing structure for particular requirement and purpose of indexing, therefore reducing the tree structure computation time. After creation of tree-based structure for several sequential patterns, indexing is performed using Bond Indexed Sequence, where several sequential patterns are analyzed to improve the search performance about chemical information. Chemical information indexing using multi level index pruning with the aid of sequential feature selection algorithm identifies and selects frequent and selective chemical molecular information as features to index, therefore reducing the chemical bond indexing time. Finally, to support user provided search queries that require a match between the chemical names used as a keyword, all possible sub-formulae of formulae that appear in any sequence are indexed. This in turn prunes the indices significantly without compromising the quality of the returned results in a significant manner.

Keywords: Chemical information, Probabilistic classifier, conditional probabilities, Chemical Structure, Bond Tree

1. Introduction

Modern chemistry is regarded as an effort of justifying the fastness and reactivity of compounds in chemical bonding. However, from the point of view of heuristic, recurrence of structural patterns could also be interpreted based on the distances and geometric measurements between atomic nuclei. This therefore helps in analyzing and recognizing the patterns for efficient prediction of presence of atoms, molecules during a chemical reaction.

Based on structural information, Probabilistic Analysis of Molecular Motifs (PAMM) [1] designed and classified structural patterns using fuzzy as an entity. This in turn identified the atomic patterns automatically and also helped in recognizing complex structural patterns. With the increasing complexity of real world scenarios, delivering temporal structures in a sequence was considered to be hard. To address this situation, Temporal Skeletonization [2] was designed that summarized temporal correlations in the form of graph resulting in avoiding curse of dimensionality. However, in case of multi determinant state functions, problem still remains unaddressed. To this, a maximum separation criterion was introduced in [4] to facilitate evaluation of bond orders.

Molecular graph can be used in chemical graph theory to evaluate the chemical, physical properties. In [5], second atom bond connectivity index was analyzed using hexagonal chain model. In [6], an efficient indexing method was designed with the aid of Non-ordered Discrete Data Spaces (NDDS) to perform similarity search in a significant manner. However, optimization remained an issue to be addressed. In [7], a tree structured formation for optimization using memetic algorithm was designed based on nature-inspired tree-based evolutionary operators. However with the objective of improving accuracy, Conditional Random Fields and Support Vector Machines [8] were evolved to enhance search performance.

Based on the aforementioned methods and technique, Chemical Structured Bond Tree-based Indexing (CSBT-I) method is designed and the main contributions of this paper are summarized as follows:

- A new Bond Tree-based structure using geometric concepts is presented which improves the tree construction time on the basis of the molecules present in the sequence patterns. The proposed method follows normalization for corresponding chemical bonding with the aid of adjacency matrix, therefore improving the connectivity between any pair of atoms in the molecule.
- By efficient construction of balanced tree, the redundancy of occurrence of an atom or molecule is reduced in a sequence pattern based on the tree representative of the molecule.
- Searching performance is improved by designing Bond Indexed Sequence model in an efficiency manner that not only performs efficient pruning but also does not compromise the quality of tree structure.

Rest of the paper is organized as follows. Relevant work in this area is presented in section 2. Section 3 introduces the relevant concepts and notations used for Chemical Structured Bond Tree-based Indexing (CSBT-I) method. Section 4 reports our experimental results with detailed discussions. Concluding remarks follow in the last section 5.

2. Related works

Many indexing methods have been presented regarding chemical information. Some of them are atom bond indexing [9], using molecular interactions [10], to name a few. An efficient sorting model using in-plane and out-of-plane was designed in [11] to improve search performances. Mathematical properties of

ABC was provided in [12] for finding extreme values of chemical trees. Another efficient indexing tree that supports ensemble models on data streams was designed in [13] by utilizing Ensemble tree and R tree.

The database community in current years has evolved as a milestone with the growing amount of research on uncertain data modeling, due to its significance in different areas including, data cleaning, radio frequency identification (RFID) networks, chemical formation and so on. In [14], a novel indexing structure named U-Quadtree was evolved to improve search efficiency. LIGht weight Hash Tree (LIGHT) [15] was evolved as a model to improve response time and bandwidth consumption that also served as an efficient query search model. To support geometric document search, rank-based search algorithm was designed in [16].

In chemical graph theory, many invariant polynomials and topological indices exists for a molecular graph. The topological index on the other hand provides numerical value for correlation of chemical structure with diversified physical and chemical properties. To this, simple connected molecular graphs were constructed in [17]. In [18], topological index of certain molecular structures from mathematical standpoint was presented. A detailed comparison between geometric arithmetic index and atom bond connectivity index to analyze molecular structure descriptors was presented in [19]. Concept of bond flexibility index was analyzed in [20].

On the basis of the analyses, Chemical Structured Bond Tree-based Indexing is constructed called, CSBT-I, which comprises of two main parts, construction of chemical bond tree and indexing through multi level indexing. The former provides functionality to perform efficient construction of tree for the sequence patterns on the basis of molecular formula and the latter shows a summary of each interaction pattern through indexing and therefore emphasizing on the efficient search.

3. Methodology

In this section a method called Chemical Structured Bond Tree-based Indexing (CSBT-I) for different sequence patterns are analyzed for the sequences in which chemical bonds are organized in the training samples. The novelty of the CSBT-I method in this paper lies in analyzing different sequential patterns, that are capable of sequencing the chemical bonds organized in the training samples more efficiently than any other conventional method of indexing. This is achieved by creating temporary index structure and improving search performance of chemical information through bond indexing. To do this, a Bond Tree-based structure is created as a temporary indexing structure. Figure 1 illustrates the block diagram of Chemical Structured Bond Tree-based Indexing.

DOI: 10.18535/ijecs/v6i2.32

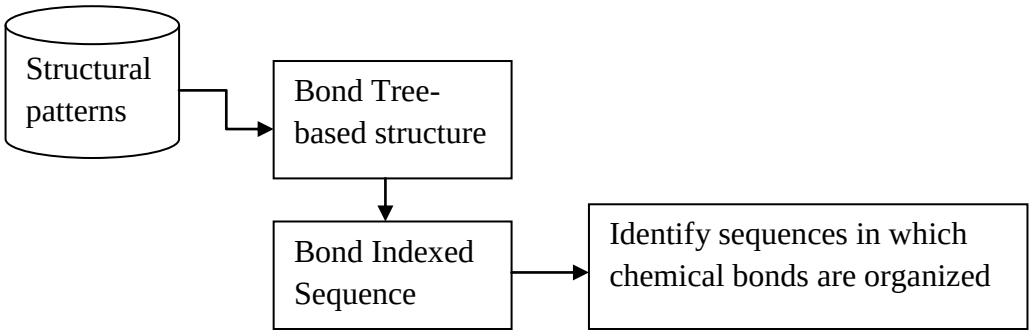


Figure 1 Block diagram of Chemical Structured Bond Tree-based Indexing

As illustrated in the block diagram, the Chemical Structured Bond Tree-based Indexing (CSBT-I) method consists of two parts. The first part involved in the construction of Bond Tree-based structure uses atom-bond connectivity with the core objective of optimizing different sequential patterns. Followed by this, the second part involved is the design of Bond Indexed Sequences which is performed through sequential feature selection algorithm and multi level indexing. The elaborate description of CSBT-I method is given in the following sections with the aid of preliminaries.

3.1 Preliminaries

Let us consider a four element tuple connected graph (i.e. molecular graph that form aggregated atoms) '$G = (V, E, \alpha, \beta)$', where '$V$' represents the set of vertices, '$E \subseteq V * V$' represents the set of undirected edges, with maximum vertex degree of at most 4 is said to be a 'molecular graph'. Here 'α' is the set of vertex and edge labels whereas 'β' maps vertices to the edge labels respectively. The graphical representation resembles a structural formula of a certain molecule. Then, the bond order and atom-bond connectivity [3] graph is as given below,

$$BO = \frac{N_b - N_{ab}}{2} \quad (1)$$

From (1), the bond order 'BO', is the average ratio of difference between the number of bonding electrons 'N_b', and number of anti bonding electrons 'N_{ab}' respectively. With the bond order, the atom-bond connectivity [4] for constructing tree-based structure is obtained as given below.

$$ABC\ (G) or\ ABC\ (pq) = \sqrt{\frac{d(V) + d(E) - 2}{d(V)d(E)}} \quad (2)$$

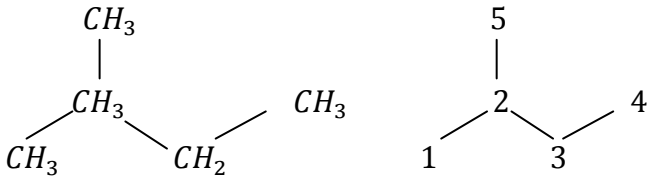
From (2), the atom-bond connectivity 'ABC' of a graph 'G' is the ratio of square root of summation of distance between the vertices '$d(V)$' and edges '$d(E)$' to the product of distance between vertices and edges '$d(V)d(E)$' respectively.

DOI: 10.18535/ijecs/v6i2.32

3.2 Bond Tree-based structure

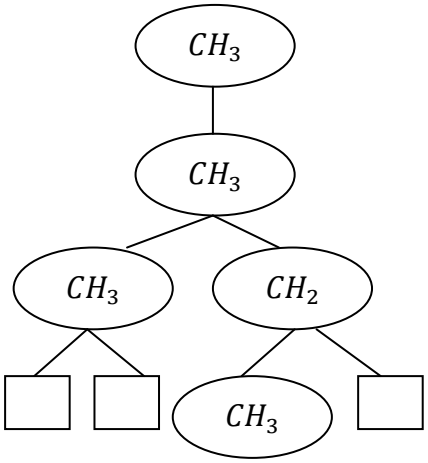
The Bond Tree-based structure in CSBT-I method uses geometric concepts to analyze and optimize different sequence patterns. In order to control the contribution of each chemical bonding for different sequential patterns in the geometric formation, normalization is applied. To perform normalization, the edge length of each chemical bonding is normalized by the domain size of the corresponding chemical bonding with the aid of adjacency matrices and is as given below.

Let us consider a molecular representation of isopentane as shown in figure 2(a), then the graphical representation is as given in figure 2(b) with the adjacency matrix representation as shown in figure 2(c) for which the bond Tree-based structure of isopentane is provided in figure 2(d).



(a) Molecule (b) Graph

$$\begin{vmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{vmatrix}$$ (c) Adjacency matrix



(d)

Figure 2 (a) Molecular representation of isopentane (b) Graphical representation of isopentane (c) Adjacency matrix representation of isopentane (d) Bond Tree-based structure of isopentane

As shown in the figure 2(c), in the adjacency matrix of isopentane, rows and columns corresponds to the atom numbering, off-diagonal elements bond ordering pointed by row and column. With all chemical bonds known a priori, the connectivity between any pair of atoms in the molecule is embedded into a Bond Tree-based Structure. This one-time embedding then allows all subsequent operators, under the guidance of

the constructed tree structure, is then used as a basis for searching and indexing individuals. The pseudo code of Bond Tree-based Structure is provided in figure 3.

Input: Training samples Dataset 'DS', sequence pattern 'SP'
Output: Optimized pattern generation
1: Begin 1: For each sequence pattern 'SP' and allocate it as root node 2: Establish its adjacent sequence patterns and allocate them as its child node 3: Allocate all the child nodes to 'Q' 4: Assign empty set 'S' 5: While 'Q' not empty 6: Repeat 7: Identify neighboring sequence patterns of 'a' 8: Allocate them as the children of 'a' 9: Until (children of 'a' = parent of 'a') 10: If 'a' has at least one child 11: Put 'a' in 'S' 12: End if 13: End while 14: End for 15: End

Figure 3 Pseudo code of the construction of Bond Tree-based Structure

Inspired by the molecular patterns by machine learning [2], in this work, a Bond Tree-based Structure is presented. To start with, different sequence patterns that form as the input are analyzed and are made available for further processing. Based on the tree representative of the molecule (i.e. sequence patterns) and the corresponding empty set thereof, the Bond Tree-based Structure then performs a random rotation. In other words, each molecule in the sequential pattern forms an adjacency matrix between that particular atom and its parent. This procedure is repeated for all set of molecules in the sequence pattern or until the initial population is fully populated. In this way, consequently, it is easy to see a molecule as a tree provided all its bond lengths and angles form independency with each other. Finally, the neighboring sequence patterns are identified to form a tree for each molecule that forms as a basis for temporary indexing structure.

3.3 Bond Indexed Sequence

With the resultant Bond Tree-based Structure, a Bond Indexed Sequence is generated to improve search performance about chemical information and their ordering. In order to ensure search performance

that require a partial match between the atoms present in the molecule (i.e. obtained from sequential pattern through tree structure) and user provided atoms, all possible sequence patterns for each molecule has to be indexed. Indexing all possible sequence pattern, results in increase in the storage and memory requirements. Therefore, in this paper, a Bond Indexed Sequence to prune the indices in a significant manner without compromising the quality of the returned results is designed.

In the process of recognizing structural patterns, if non parametric partitioning [1] has to be performed for recognizing molecular patterns for possible sub-terms of a chemical name, the size of the partitioning will be extremely large and constructing such a partitioning model not only results in the increased memory requirements but also compromises the processing time. To address this problem, in this work, an index pruning technique, called, Bond Indexed Sequence that not only reduces the index size but also reduces degradation in the quality of search results is presented.

To construct Bond Indexed Sequence sequential patterns are analyzed and features are selected for indexing. This in turn improves the corresponding search operations and ensures higher rate of similarity search. For example, same chemical molecule may have dissimilar formula strings mentioned in text, e.g., 'tartaric acid' may have been listed as '$C_4H_6O_6$' (basic formula) or '$HO_2CCH(OH)CH(OH)CO_2H$' (structural formula). Then, given a training samples dataset 'DS' of sequences '$Seq_i = Seq_1, Seq_2, \dots, Seq_n$', '$DS_{Seq'}$', the support of subsequence 'Seq'' represents the set of all sequences 'Seq' containing 'Seq''. The sequence is segmented into subsequences, and is mathematically formulated as given below.

$$Seq' = < Seq_1, Seq_2, \dots, Seq_n > \quad (3)$$

$$Seq' \preccurlyeq Seq \quad (4)$$

$$Freq = \left|D_{seq}\right| \quad (5)$$

From (4) and (5), the subsequence 'Seq'' precede the set of all sequences 'Seq', with '$\left|D_{seq}\right|$' representing the number of sequences in 'seq'. Let us consider a dataset 'DS' with sequence pattern '$\{CH_5, CH_3CL_2, CHCL_2\}$', '$DS_{CH_3}$' the support of subsequence 'CH_3' is then given as '$DS_{CH_3} = \{CH_5, CH_3CL_2\}$'. With the above considerations, a sequential feature selection algorithm is designed that extracts all molecules from a given sequence pattern. Followed by this, all possible partial molecules are generated. With the generated possible partial molecules, the proposed method records the occurrences or the frequency is measured with which the indexing is performed. Figure 4 shows the algorithmic representation of sequential feature selection algorithm.

Input: Training samples Dataset ' DS ', sequence pattern ' SP ', Molecules '$M = mol_1, mol_2, \dots, mol_n$'
Output: Optimized search performance
1: Begin 2: For each sequences 3: Repeat 4: For all molecules 'M' in a sequence pattern 'SP' 5: If $Seq' \preccurlyeq Seq$ 6: Extract all molecules from a given sequence pattern using (3) 7: Measure frequency of occurrences using (5) 8: Compare probability of each molecule in sequence pattern to user specified chemical information using (6) 9: Perform multi level indexing 10: End if 11: End for 12: Until (all sequence patterns are processed) 13: End for 14: End

Figure 4 Sequential feature selection algorithm

Analogously as shown in the figure 4, Sequential feature selection algorithm compares the probability of each molecule in a sequence pattern with the user specified chemical information. This in turn gives more flexibility to different sequencing being used which is mathematically formulated as given below.

$$Prob = 1 - \frac{Prob\left(Seq/_n\right) * F}{Prob\left(Seq/_n\right) + Prob\left(Seq/_M\right)} \quad (6)$$

DOI: 10.18535/ijecs/v6i2.32

From (6), '$Prob\left(Seq/_{n}\right)$' specifies the total summed probability of sequence the method derived by Bond Tree-based Structure algorithm, with '$Prob\left(Seq/_{M}\right)$' specifying the probability of sequence that obtains a random zero value over Markov model. This therefore ensures and recognizes sequences even in the presence of sequencing errors. Finally, 'F' represents the fraction of '$Prob\left(Seq/_{n}\right)$' corresponding to the most likely occurrence of molecule in a sequence, if present '1' or '0' otherwise. With the occurrence of molecule in a sequence, a multi level indexing is applied to improve the search performance. Figure 5 illustrates the Bond multi level indexing followed in the proposed work.

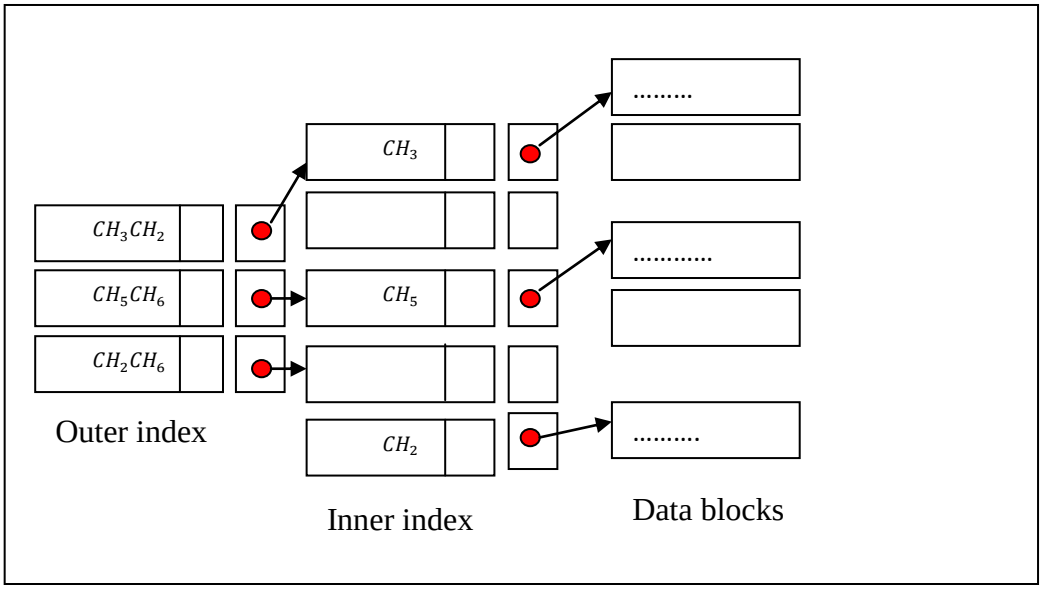


Figure 5 Structure of Bond multi level indexing

As illustrated in figure 5, the Bond index records contains of the search-key values and data pointers. With the increasing size of the molecule, analogously, the size of sequence pattern also grows. In order to speed up the search operations, the proposed work uses bond multi level indexing in order to keep large size index in memory. As shown in the figure, the bond multi level indexing applied in the proposed work is to reduce the part of the index that we have to continue to search. It consists of an inner index, outer index and data blocks with the single level ordered indexes in inner indices, the new index to the inner index called as the outer index and so on. The bond multi level indexing process is continued until all entries of a specific sequence in molecular formula fit in a single block, ensuring optimized searching.

4. Experimental settings

DOI: 10.18535/ijecs/v6i2.32

The experimental work is carried out in JAVA language for evaluating the sequence patterns in chemical bonding. The performance of proposed method is evaluated with parameters such as chemical bond density, chemical bond size, tree structure computation time, chemical bond indexing time and chemical bond indexing accuracy against existing state-of-art techniques. The experimental data used for the analysis of proposed and existing methods are extracted from Molecular Description Data Sets (Octane Isomers (O8), PolyAromatic Hydrocarbons (PAH), and PolyChloroBiphenyls (PCB)). The results of the experiments are presented below.

4.1 Scenario 1: Tree structure computation time

Tree structure computation time is measured using the number of chemical bonds (i.e. density) and the tree structure formation time. The mathematical formulation for Chemical bond tree structure computation time is given as below.

$$CT = \sum_{d=1}^{n} CB_d * Time\ (SP_i) \quad (7)$$

From (7), the tree structure computation time 'CT' is measured using the number of chemical bonds or density of chemical bonds 'CB_d' and the time taken to construct tree (i.e. time to construct tree for different sequence patterns provided as input). The computation time is measured in terms of milliseconds.

Table 1 Tabulation for chemical bond tree construction time

Chemical bond density	Chemical bond tree construction time (ms)		
	CSBT-I	PAMM	Temporal Skeletonization
15	5.28	6.19	6.49
30	9.32	10.47	10.85
45	13.21	14.35	15.05
60	18.90	19.12	21.35
75	23.14	23.29	24.89
90	28.90	29.32	31.43
105	33.12	34.29	35.13
120	38.32	39.47	41.99
135	43.89	44.32	46.32
150	49.14	50.30	52.17

DOI: 10.18535/ijecs/v6i2.32

The table given above describe the performance result of the Chemical Structured Bond Tree-based Indexing (CSBT-I) with the existing methods Probability Analysis of Molecular Motifs (PAMM) [1] and Temporal Skeletonization [2]. The chemical bond tree construction time is measured based on the chemical bond density present in the sequence pattern. The value of the proposed CSBT-I method is compared with the existing PAMM [1] and Temporal Skeletonization [2] is illustrated in table 1. Table 1 displays the tree construction time for chemical bond density in the range of 15 and 150. The tree construction time is used to measure the time needed to construct the chemical bonds. The chemical bond density is taken as the input for constructing the tree and the construction time using different methods are obtained.

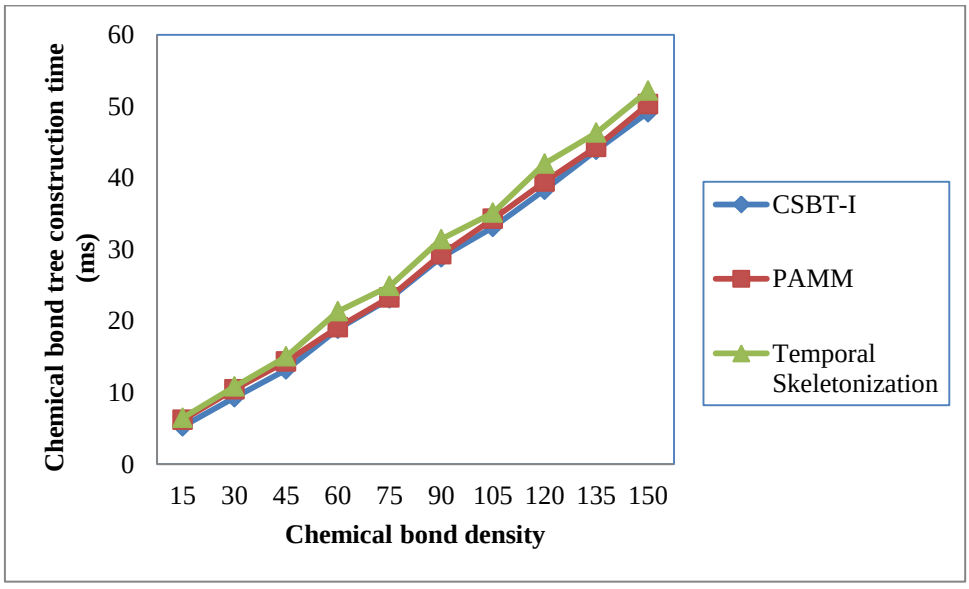


Figure 6 Chemical bond tree construction time versus chemical bond density

Figure 6 shows the chemical bond tree construction time during optimized searching based on the chemical bond density. As shown in the figure, by applying the Bond Tree-based structure using geometric concept, construction time is improved even with the increase in the chemical bond density. In the case of the proposed indexing method, the CSBT-I perform better than the existing PAMM [1] and Temporal Skeletonization [2]. On average, the proposed CSBT-I outperforms the PAMM by 5% and 11% compared to Temporal Skeletonization respectively. In the proposed CSBT-I method, edge length of the chemical bond is normalized using adjacency matrix. With the adjacency matrix values, chemical bond tree is constructed for different sequence patterns. This in turn optimizes the construction of tree and therefore reduces the tree construction time.

Furthermore, the transformation process is applied to coefficients only from the high pass filter and are subsequently provided and are repeated up to the desired level of wavelet computation that further improves the compression ratio for different set of images.

4.2 Scenario 2: Chemical bond indexing accuracy

The chemical bond indexing accuracy is the measure to determine the rate of probability of the structures of chemical bonds being properly indexed so that resulting in the improved search performance. The indexing accuracy is mathematically formulated as given below.

$$IA = \sum_{i=1}^{n}((Chemical\ bonds\ properly\ indexed)/CB_i) * 100 \qquad (8)$$

From (8), the indexing accuracy 'IA' is measured with chemical bonds properly indexed to the total chemical bonds 'CB_i' used in the experimentation. It is measured in terms of percentage and higher the indexing accuracy results in the improvement of the search performance and therefore the efficiency of the method is said to be proven. The chemical bond indexing accuracy with varying chemical bond density with the existing methods PAMM, Temporal Skeletonization and CSBT-I is illustrated in table 2.

Table 2 Tabulation for chemical bond indexing accuracy

Chemical bond density	Chemical bond indexing accuracy (%)		
	CSBT-I	PAMM	Temporal Skeletonization
15	82.15	71.37	63.40
30	88.13	78.32	67.35
45	91.53	83.48	79.44
60	92.32	85.27	81.23
75	85.14	78.09	74.05
90	87.10	80.05	76.01
105	90.32	83.27	79.23
120	85.56	77.51	73.47
135	90.14	83.09	79.04
150	93.33	86.28	82.24

The chemical bond indexing accuracy is used to evaluate the rate at which the indexing is performed and the accurate indexed evolved which further helps to measure the search performance for chemical information. The actual chemical bond density is taken to measure the chemical bonds that are properly indexed. Next, the chemical bond indexing accuracy for measuring search performance using different methods is obtained. Finally, the ratio of chemical bonds that are properly indexed and the chemical bond density gives the chemical bond indexing accuracy.

DOI: 10.18535/ijecs/v6i2.32

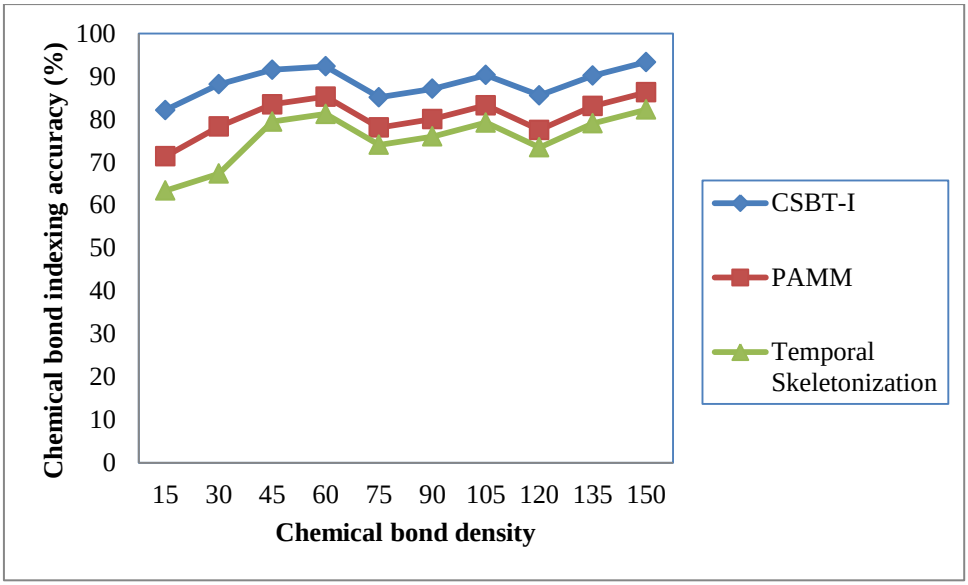


Figure 7 Chemical bond indexing accuracy versus chemical bond density

As shown in the figure 7, the chemical bond indexing accuracy is measured using chemical bond density in the range of 15 to 150 and performed for 10 simulation runs. Compared to the existing methods, the proposed CSBT-I method involves highest indexing accuracy even when chemical bond density increases. In the proposed CSBT-I method, bond indexed sequence is performed for varying chemical bond density a sequential feature selection algorithm is performed. Here the sequence are segmented into subsequences using the precede formation. With the resultant value obtained, the sequential feature selection algorithm extracts all molecules from a given sequence pattern that has dual advantages. They are the index size is reduced and in time reduces the degradation in search quality. The indexing process is finally performed only if the support of subsequence 'Seq'' represents the set of all sequences 'Seq' that helps in maximizing the indexing accuracy by 9% compared to PAMM and 15% compared to Temporal Skeletonization respectively.

4.3 Scenario 3: Chemical bond indexing time

Chemical bond indexing time is the time taken for indexing with respect to the number of chemical bonds (i.e. density). The mathematical formulation for Chemical bond indexing time is as given below.

$$IT = \sum_{d=1}^{n} CB_d * Time\ (Multi\ level\ indexing) \qquad (9)$$

From (9), the chemical bond indexing time 'IT' is measured using the number of chemical bonds or density of chemical bonds 'CB_d' and the time taken to perform multi level indexing. The indexing time is measured in terms of milliseconds. Table 3 provides an insight into the chemical bond indexing time using the proposed CSBT-I method and the existing methods, PAMM and Temporal Skeletonization respectively.

DOI: 10.18535/ijecs/v6i2.32

Table 3 Tabulation for chemical bond indexing time

Chemical bond density	Chemical bond indexing time (ms)		
	CSBT-I	PAMM	Temporal Skeletonization
15	4.28	4.95	5.33
30	7.14	8.32	9.43
45	10.35	12.14	13.25
60	14.56	16.89	17.90
75	18.32	20.14	21.25
90	22.14	24.33	25.44
105	25.14	27.87	28.98
120	30.33	32.14	33.25
135	34.14	37.78	38.89
150	38.13	41.32	52.43

Chemical bond indexing time is obtained by the product of chemical bond density and the time taken to perform multi level indexing. Lower the chemical bond indexing time, more efficient the method is said to be and is measured in terms of milliseconds (ms). The targeting results of chemical bond indexing time using CSBT-IT with two state-of-the-art methods [1], [2] in table 3 presented for comparison based on the chemical bond density.

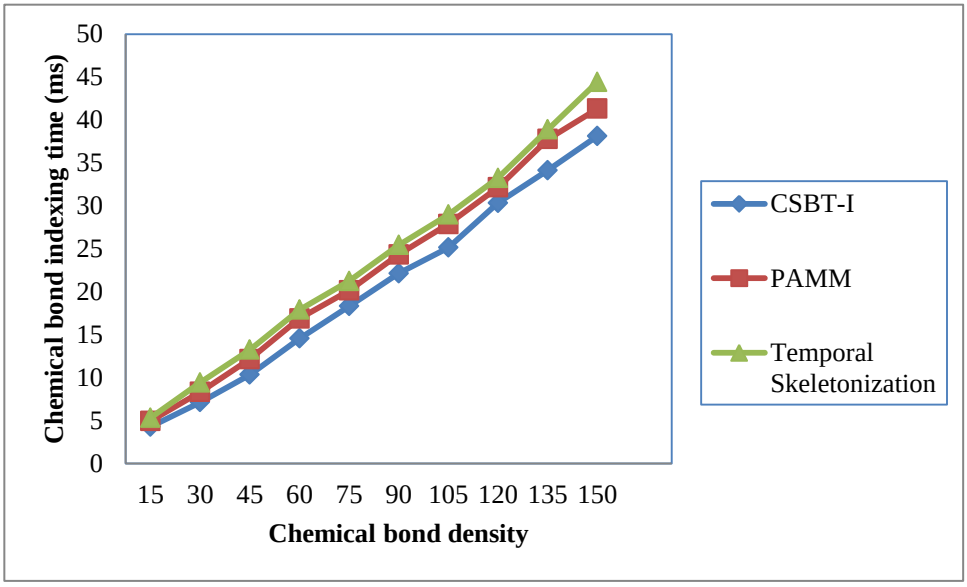


Figure 8 Chemical bond indexing time versus chemical bond density

Figure 8 presents the variation of chemical bond indexing time with respect to chemical bond density. All the results provided in figure 8 confirm that the proposed CSBT-I method significantly outperforms the other two methods, PAAM [1] and Temporal Skeletonization [2]. The chemical bond indexing time is improved in the CSBT-I method using the multilevel indexing technique. With the application of multilevel indexing technique, the sequential feature selection algorithm obtains a measure of probable occurrence of each molecule in a sequence pattern to the user specified chemical information. This in turn ensures higher rate of flexibility towards different sequencing. Followed by this, random zero value over Markov model is applied to measure the probability rate. With the molecule occurrences in a sequence, the CSBT-I method performs multi level indexing is applied. This in turn reduces the chemical bond indexing time using CSBT-I by 12% compared to PAAM. As a result chemical indexing time is reduced in CSBT-I method using multilevel indexing. In addition by applying multilevel indexing in CSBT-I method in turn minimizes the index that we have to continue to search, in turn reducing the chemical bond indexing time by 19% compared to Temporal Skeletonization.

5. Conclusion

Indexing scheme is one of the key issues to be handled for different structural patterns for the sequences in which the chemical bonds are organized. Tree structure is developed for organization of chemical bonds and the ordering of chemical bonds is obtained by chemical bond tree-based structure. This paper presents an emergence of new search performance about chemical information called Chemical Structured Bond Tree-based Indexing (CSBT-I). To overcome the limitations of existing search performance for sequences in which chemical bonds are organized, three parameters such as the chemical bond indexing accuracy, chemical bond indexing time and tree structure computation time is taken into account along with the sequential feature selection algorithm from which it is identified that the proposed CSBT-I method has the improved search performance. Besides, a tree-based structure to reduce the chemical bond tree construction time and multilevel indexing to reduce the chemical bond indexing time is introduced. Performance results revealed that the proposed CSBT-I method provides 12% indexing accuracy by reducing the indexing time by 16% compared to the state-of-the-art methods.

REFERENCES

[1] Piero Gasparotto and Michele Ceriotti, “Recognizing molecular patterns by machine learning: An agnostic structural definition of the hydrogen bond”, The Journal of Chemical Physics, Volume 141, Issue 17, November 2014, Pages 1-13.

[2] Chuanren Liu, Kai Zhang, Hui Xiong , Guofei Jiang, Qiang Yang, “Temporal Skeletonization on Sequential Data:Patterns, Categorization, and Visualization”, IEEE Transactions on Knowledge and Data Engineering, Volume 28, Issue 1, January 2016, Pages 211 – 223.

[3] E. Estrada, L. Torres, L. Rodr´ıguez, and I. Gutman, "An atombond connectivity index: modelling the enthalpy of formation of alkanes," Indian Journal of Chemistry A, vol. 37, no. 10, pp. 849–855, 1998.

[4] Dariusz W. Szczepanik , Janusz Mrozek, "Ground-state projected covalency index of the chemical bond", Elsevier, Computational and Theoretical Chemistry, Volume 1023, 1 November 2013, Pages 83–87.

[5] Wei Gao andWeifanWang, "Second Atom-Bond Connectivity Index of Special Chemical Molecular Structures", Hindawi Publishing Corporation, Journal of Chemistry, Volume 2014, October 2014, Pages 1-9.

[6] Changqing Chen, Alok Watve, Sakti Pramanik, and Qiang Zhu, "BoND-Tree: An Efficient Indexing Method for Box Queries in Non ordered Discrete Data Spaces", IEEE Transactions on Knowledge and Data Engineering, Volume 25, Issue 11, Novemver 2013, Pages 2629-2643.

[7] M.M. Ellabaan, S.D. Handoko, Y.S. Onga, C.K. Kwoh, S.A. Bahnassy , F.M. Elassawy , HY. Mand, "A tree-structured covalent-bond-driven molecular memetic algorithm for optimization of ring-deficient molecules", Elsevier, Computers & Mathematics with Applications, Volume 64, Issue 12, December 2012, Pages 3792–3804.

[8] Bingjun Sun, Prasenjit Mitra, C. Lee Giles, Karl T. Mueller, "Identifying, Indexing, and Ranking Chemical Formulae and Chemical Names in Digital Documents", Journal ACM Transactions on Information Systems (TOIS), Volume 29 Issue 2, April 2011, Pages 1-43.

[9] Kinkar Ch. Das, "Atom-bond connectivity index of graphs", Elsevier, Discrete Applied Mathematics, Volume 158, Issue 11, 6 June 2010, Pages 1181–1188.

[10] Kota Kasahara and Kengo Kinoshita, "GIANT: pattern analysis of molecular interactions in 3D structures of protein–small ligand complexes", BMC Bioinformatics 2014, January 2014, Pages 1-6.

[11] Olga V. Sizova, Leonid V. Skripnikov, Alexander Yu. Sokolov, "Symmetry decomposition of quantum chemical bond orders", Elsevier, Journal of Molecular Structure: Theochem, Volume 870, Issues 1–3, 15 December 2008, Pages 1–9.

[12] Boris Furtula, Ante Graovac , Damir Vuki£evi, "Atom-bond connectivity index of trees", Elsevier, Discrete Applied Mathematics, Volume 157, Issue 13, 6 July 2009, Pages 2828–2835.

[13] Peng Zhang, Chuan Zhou, Peng Wang, Byron J. Gao, Xingquan Zhu, and Li Guo, "E-Tree: An Efficient Indexing Structure for Ensemble Models on Data Streams", IEEE Transactions on Knowledge and Data Engineering, Volume 27, Issue 2, February 2015, Pages 461-474.

[14] Ying Zhang, Wenjie Zhang, Qianlu Lin, Xuemin Lin, "Effectively Indexing the Multi-Dimensional Uncertain Objects for Range Searching", IEEE Transactions on Knowledge and Data Engineering, Volume 26, Issue 3, March 2014, Pages 608-622.

[15] Yuzhe Tang, Shuigeng Zhou, and Jianliang Xu, "LIGHT: A Query-Efficient Yet Low-Maintenance Indexing Scheme over DHTs", IEEE Transactions on Knowledge and Data Engineering, Volume 22, Issue 1, January 2010, Pages 59-75.

[16] Zhisheng Li, Ken C.K. Lee, Baihua Zheng, Wang-Chien Lee, Dik Lun Lee, and Xufa Wang, "IR-Tree: An Efficient Index for Geographic Document Search", IEEE Transactions on Knowledge and Data Engineering, Volume 23, Issue 4, January 2011, Pages 585-599.

[17] Mohammad Reza Farahani, Wei Gao, M. R. Rajesh Kanna, R. Pradeep Kumar, and Jia-Bao Liu, "General RandiT, Sum-Connectivity, Hyper-Zagreb and Harmonic Indices, and Harmonic Polynomial of Molecular Graphs", Hindawi Publishing Corporation, Advances in Physical Chemistry, Volume 2016, August 2016, Pages 1-7.

[18] Wei Gao, Weifan Wang, Muhammad Kamran Jamil, and Mohammad Reza Farahani, "Electron Energy Studying of Molecular Structures via Forgotten Topological Index Computation", Hindawi Publishing Corporation, Journal of Chemistry, Volume 2016, July 2016, Pages 1-8.

[19] Zahid Raza, Akhlaq Ahmad Bhatti, And Akbar Ali, "More on Comparison Between First Geometric-Arithmetic Index and Atom-Bond Connectivity Index", Combinatorics (math.CO), June 2015, Pages 1-10.

[20] Claus-Wilhelm von der Lieth, Klaus Stumpf-Nothof, and Ursula Prior, "A Bond Flexibility Index Derived from the Constitution of Molecules", J. Chem. Inf. Comput. Sci., 1996, Volume 36, Issue 4, Pages 711–716.