Applied Mathematics and Computation 259 (2015) 242–250

Contents lists available at ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc

A comparative analysis of the Tanimoto index and graph edit distance for measuring the topological similarity of trees

Matthias Dehmer [a,*], Kurt Varmuza [b]

[a] Department of Computer Science, Universität der Bundeswehr München, Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany
[b] Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Wiedner Hauptstrasse 7/105, A-1040 Vienna, Austria

ARTICLE INFO

Keywords:
Graph similarity
Graph similarity measures
Graph distance measures
Quantitative graph theory
Graphs
Chemical graphs

ABSTRACT

In this paper we explore interrelations between the Tanimoto index and the graph edit distance (GED) for measuring the topological similarity/distance of graphs. To do so, we discuss cumulative similarity/distance distributions of these measures and other data analysis methods. Also we explore properties of the Tanimoto index by using sets of chemical alkane trees and exhaustively generated ordinary trees as subgraphs. In particular, we discuss numerical results for exploring the approximation behavior of the Tanimoto index by GED.

1. Introduction

Measuring the similarity or distance between networks has been an intricate problem, see, e.g., [17,22,40,57,66]. This problem arises when one needs to compare structural representations quantitatively. In fact, the structural comparison of graphs by using quantitative similarity or distance measures has been performed in various scientific disciplines. In the following, we list some important examples thereof. In chemistry, molecular structures have been compared by employing fingerprint-based methods and other graph similarity approaches for performing similarity searching and drug discovery [44,40,51,64]. In biology, similarity calculations of several types of biological networks inferred from data have been used for classifying biological domains and organisms [22,45,46]. Web-based document structures have been classified structurally based on applying graph similarity measures for solving problems in web structure mining [17,12]. Finally we mention an interesting application of graph similarity in quantitative finance [21]. Emmert-Streib and Dehmer [21] used comparative graph measures for exploring and predicting financial market crashes by using networks which have been inferred from stock data.

We emphasize that measuring the similarity or distance between graphs always depends on a specific model. A classical one to determine graph similarity is based on graph isomorphism [57,66]. This era led to various graph similarity measures and graph metrics which have been investigated extensively and applied in disciplines such as mathematical psychology and social network analysis, see, e.g., [57,66,56,59,33]. So-called inexact methods for matching graphs have been developed as well. A prominent example thereof is the *graph edit distance* (GED) [6,8] and is based on graph edit operations (see Section 2.2.3). Other (inexact) techniques are based on using graph grammars [27,28], string alignments [17,50], and graph kernels [32,25]. Another string-based technique is based on deriving feature vectors of graphs and, then, to employ existing

[*] Corresponding author.
E-mail addresses: matthias.dehmer@unibw.at (M. Dehmer), kvarmuza@email.tuwien.ac.at (K. Varmuza).

http://dx.doi.org/10.1016/j.amc.2015.02.042

similarity measures such as the Jaccard and Tversky coefficient [40,64]. A special case thereof is the Tanimoto index by using binary vector comparison where the vector elements indicate the absence or presence of pre-defined subgraphs [65,63,62].

The main contribution of this paper is twofold. First, we investigate the Tanimoto index on exhaustively generated alkane trees with exhaustively generated trees as subgraphs. The aim is to study the behavior of this widely used measure on more general graphs and, therefore, to generalize earlier findings, e.g., [40,62]. To the best of our knowledge, the Tanimoto index has mainly been used in chemoinformatics [26]. By analyzing this quantity in a more general context, these findings could be translated to other problems/disciplines. Second, we study interrelations of the Tanimoto index and GED. Note that GED has been often used as benchmark measure as it is applicable to any type of graph and has a simple and clear interpretation. From studying these interrelations (see Section 3.3), we draw conclusions on the comparability of these measures.

2. Methods

2.1. Graph-theoretical preliminaries

We start with some graph-theoretical preliminaries we need for our analysis, see [16,31,61]. In this paper, we restrict this study to (chemical) alkane trees but the method (see Section 2.2 and Section 3.1) can be applied to arbitrary graphs as well.

Definition 2.1. $G = (V, E)$, $|V| < \infty$ is a finite undirected graphs. V is the set of vertices and E the set of edges, $E \subseteq \binom{V}{2}$.

Definition 2.2. We shall write $|V| := i \in \mathbb{N}$ for the cardinality of the vertex set. The cardinality of the edge set is denoted by $|E|$. We shall write $N(G)$ and $|E(G)|$ to emphasize to which graph we are referring to.

Definition 2.3. A finite tree $T = (V, E)$, $|V| < \infty$ is a connected and acyclic graph.

Definition 2.4. $\mathcal{T}(i)$ denotes the set of undirected, exhaustively generated trees where each tree has i vertices. All trees are connected and pairwise non-isomorphic.

Example 2.1. It holds $\mathcal{T}(i) := \{T_1^i, T_2^i, \ldots, T_{|\mathcal{T}(i)|}^i\}$. We use the program `geng` contained in the package `Nauty` due to McKay [41] for calculating all connected and non-isomorphic trees exhaustively. For example, we yield:

$$\begin{aligned}
&|\mathcal{T}(2)| = 1,\\
&|\mathcal{T}(3)| = 1,\\
&|\mathcal{T}(4)| = 2,\\
&|\mathcal{T}(5)| = 3,\\
&|\mathcal{T}(6)| = 6,\\
&|\mathcal{T}(7)| = 11,\\
&|\mathcal{T}(8)| = 23,\\
&|\mathcal{T}(9)| = 47,\\
&|\mathcal{T}(10)| = 106.
\end{aligned}$$

Definition 2.5. Alkane trees are connected and acyclic graphs in which the degree of a vertex is maximally four.

Definition 2.6. $\mathcal{C}(i)$ denotes the set of undirected, exhaustively generated alkane trees [1] where each tree has i vertices. These alkane trees are connected and pairwise non-isomorphic.

Definition 2.7. $\mathcal{S}(i)$ is the union of the tree sets $\mathcal{T}(2), \ldots, \mathcal{T}(i)$.

Example 2.2. For instance, we obtain $\mathcal{S}(4) = \left\{T_1^2, T_1^3, T_1^4, T_2^4\right\}$ where $|\mathcal{T}(2)| = 1, |\mathcal{T}(3)| = 1$ and $|\mathcal{T}(4)| = 2$.

2.2. Measuring the topological similarity of graphs

First we briefly survey the most important approaches for characterizing *graph similarity* [66]. This process is often referred to as *graph matching* [8] and has been tackled in various disciplines. As examples, we list research areas such as chemoinformatics, bioinformatics, drug design where such methods have been used to determine the structural similarity of relational structures, see, e.g., [17,22,65,54,57].

2.2.1. Overview

To get an overview, we discuss assets and drawbacks of the most significant methods for measuring the structural similarity of graphs:

- Graph matching based on isomorphic and subgraph isomorphic relations: One of the first techniques to measure the structural similarity of graphs is due to Zelinka [66]. The Zelinka distance [56,66] is often referred to as *exact graph matching* as the graphs are matched exactly by determining graph isomorphism [8]. The measure is suitable to compare graphs structurally which have been inferred deterministically, i.e., without any uncertainty. Hence the concept is not applicable for comparing graphs which are effected by measurement errors (structural noise). A specific example thereof are biological networks which have been inferred from microarray data [20]. However the computational complexity of graph similarity measures such as the Zelinka distance [66] is often insufficient as they rely on the graph isomorphism and subgraph isomorphism problem. The subgraph isomorphism is NP complete, see [24].
- Graph matching based on graph transformations: This method belongs to *inexact graph matching* [8] and has been pioneered by Bunke [6,10,8]. The key concept is based on applying graph edit operations (such as adding/deleting edges or vertices) where the distance between graphs have been defined by the sequence of graph edit operations with minimal transformation costs. The resulting graph distance measure is called graph edit distance (GED) (see also Section 3) and has been proven useful for measuring the structural similarity of labeled and unlabeled graphs, see [6,10,11,14]. The calculation of the graph edit distance for large unlabeled graphs is computationally insufficient as the basic algorithm [49] is NP-complete.
- Graph matching based on graph grammars: Gernert [27,29,28] employed graph grammars for measuring the similarity of graphs. However this method is rather of theoretical interest as the underlying grammar is difficult to construct [27].
- Graph matching based on using machine learning techniques: For instance, so-called graph kernels and supervised machine learning methods have been used for classifying networks [32,25]. Also, Müller et al. [45] employed information-theoretic measures to classify metabolic networks by using support vector machines (SVM) efficiently. A significant advantage of such methods is their efficiency; by using appropriate data sets and kernel functions, the classification performance can be high [32,25,45]. In case training data is not available, supervised machine learning methods (e.g., SVM) are not applicable. Also, special graph kernels may be computationally inefficient, see, e.g., [32,25].
- Statistical graph matching methods: Statistical graph matching techniques have been employed in various disciplines such as image analysis [60], computer- and neuroscience [53,58] and computational linguistics [43]. But in view of the number of approaches which have been developed, such techniques are clearly underrepresented and there is no general theory for interpreting and understanding the existing results mathematically. For instance, statistical tests have been employed to tackle the graph similarity problem [23,60]. Importantly, such methods may be used to measure the similarity between networks which are affected by structural noise, see [23].
- String-based graph matching methods: Dehmer [17,13] developed so-called property string-based methods for determining the structural similarity of hierarchical structures called generalized trees [23,67]. The main idea is based on transforming generalized trees into so-called property strings [17,23]. For each level, one obtains an out-degree property string and in-degree property string. Then one derives similarity scores from the alignments of the property strings which have been used to obtain efficient graph similarity measures [13,17,23]. A generalization thereof to compare large undirected graphs by using a statistical techniques has been proposed by Emmert-Streib et al. [23]. Instead of comparing two graphs G_1 and G_2 by determining isomorphic subgraphs, the tree sets resulting from a generalized tree decomposition [23] of G_1 and G_2 are compared. By employing the method for pairwise calculating generalized trees, three similarity distributions have been inferred and compared based on a chi-square test. Finally, Emmert-Streib et al. [23] defined G_1 and G_2 to be similar iff the three similarity distributions obtained from pairwisely calculating similarity scores of the different trees sets are similar.
- Molecular similarity: Graph similarity measures/methods have been applied to determine the structural similarity of molecular graphs, see [40,55,48,47]. This has been an ongoing research problem in computational and structural chemistry. Particularly in chemical graph theory, many comparative graph measures have been explored for determining the similarity between molecular structures [55]. An attempt to measure the structural similarity between chemical graphs is to represent and compare graphs by vectors where the entries are real numbers [40]. Then, based on binary vector comparison (the vectors indicate the absence or presence of pre-defined subgraphs [63]) and by using the Tanimoto index [65,62], further similarity measures have been obtained [63,62]. Also, Klein [35,37,34,36] explored graph metrics in structural chemistry and employed metrical properties of graphs to tackle the graph similarity problem. Topological indices [4,5,38,39] and the concept of the maximum common subgraph [55] have also been employed to measure the similarity between chemical graphs. A recent line of research is due to Dehmer et al. [15] when using topological indices for defining graph distance measures based on existing distance measures defined over the real numbers [15,52]. They proved inequalities between these measures and demonstrated that the measures possess useful properties to be applied for practical problems.

2.2.2. Tanimoto index

In this paper, we focus on applying the Tanimoto index and GED [8,19,62] for determining graph similarity (see Section 3). First, we define the Tanimoto index for graphs, see [62,65]. Let G and H be connected and non-isomorphic graphs and let $\mathcal{S} \neq \emptyset$ be a set of possible induced subgraphs of both G or H. By using $\mathcal{S}$, we convert G and H to binary vectors

$$v_G := (d_1^G, d_2^G, \dots, d_{|v_G|}^G) \quad \text{and} \quad v_H := (d_1^H, d_2^H, \dots, d_{|v_H|}^H) \quad \text{iff}\, d_i^G, d_1^H \in \{0, 1\}. \tag{1}$$

We define $d_i^G = 1$ if the i-th subgraph of $\mathcal{S}$ is present in G and $d_i^G = 0$ otherwise. The same holds for d_i^H analogously. By employing the Tanimoto index t, we finally obtain the following statement.

Theorem 2.1. *The Tanimoto index for two connected graphs G and H*

$$t(G,H) := \frac{\sum_{i=1}^{|\mathcal{S}|} \text{AND}\,(d_i^G, d_i^H)}{\sum_{i=1}^{|\mathcal{S}|} \text{OR}\,(d_i^G, d_i^H)}, \tag{2}$$

is a graph similarity measure.

Proof. Following Bock [3], a similarity measure for real numbers must fulfill three properties (i)-(iii). Let $\mathcal{G}$ be a class of graphs, so we have here $t : \mathcal{G} \times \mathcal{G} \longrightarrow [0, 1]$ and the measure is defined by Eq. 2. So, if (i)-(iii) holds, Eq. 2 is a graph similarity measure. In the following, we show (i)-(iii) explicitly. (i) The first condition is

$$t(G,H) > 0. \tag{3}$$

This holds iff $v_G, v_H \neq (0, 0, \dots, 0)$ and $\sum_{i=1}^{|\mathcal{S}|} \text{AND}\,(d_i^G, d_i^H) \geqslant 1$; the latter condition is fulfilled as $\mathcal{S}$ containing potential induced subgraphs of G and H is not empty by definition and the subgraphs are chosen appropriately. (ii) The symmetry property

$$t(G,H) = t(H,G), \tag{4}$$

is also fulfilled as AND and OR are symmetric. (iii) The third property is the inequality

$$t(G,H) \leqslant t(G,G) = 1. \tag{5}$$

If $G = H$, then $v_G = v_H$. Thus by using the definitions of AND and OR, the equation $t(G,G) = 1$ is evident. Otherwise we have the situation

$$\sum_{i=1}^{|\mathcal{S}|} \text{OR}\,(d_i^G, d_i^H) > \sum_{i=1}^{|\mathcal{S}|} \text{AND}\,(d_i^G, d_i^H), \tag{6}$$

and, hence, $t(G,H) < 1$. □

Note that the Tanimoto index (sometimes called Jaccard similarity index) is widely used and has been successfully applied in chemistry [40], e.g., for searches for similar chemical structures. Because of the high diversity of chemical structures a large number of subgraphs is necessary and consequently for most chemical structures many vector elements are zero; therefore the Tanimoto index is better suited than the Hamming distance. For an overview on distances and similarity measures, see [18].

2.2.3. Graph edit distance

The GED is an important measure for measuring the similarity of graphs based on so-called graph edit operations [6]. As it is easily interpretable, GED has often been used as a benchmark method [6,9,14] when determining the similarity of graphs. The mentioned graph edit operations are insertions or deletions of edges/vertices or relabelings of vertices along with certain edit costs associated with these operations. Following Bunke [6,22], a sequence of edit operations that transforms a graph G into H by producing minimal transformation costs as an optimal inexact match [6]. By assuming that $m_1, m_2, \dots, m_n$ are the theoretically possible transformations which transform G to H, the optimal inexact match m' has been defined by [6,22]:

$$c(m') = \min\{c(m_i) | 1 \leqslant i \leqslant n\}. \tag{7}$$

$c(m_i)$ is the cost of m_i. Finally, the graph edit distance of two given graphs is the minimum cost associated with a sequence of edit operations. The optimal error-correcting graph isomorphism has finally been defined by the resulting isomorphism after obtaining the optimal sequence of edit operations [6,22]. Bunke [6] proved that the final measure is a graph metric.

Theorem 2.2. *Let* $GED(H, G)$ *be the costs for determining the optimal inexact match between* H *and* G. $\mathrm{GED}(H, G)$ *is a graph metric.*

We emphasize that various methods for determining the matching costs efficiently have been investigated [6,7,49]. In this paper, we use the normalized version of GED (interval [0,1]), see [19].

3. Results

In this section, we interpret the results when applying the Tanimoto index and GED to sets of exhaustively generated alkane trees. It would be useful if it turns out that the Tanimoto index t could serve as an approximation of GED as computing t is more efficient for small graphs. That's also one of the reasons why we have chosen relatively small graph classes, see Section 3.1. In the following, we start with some computational preliminaries.

3.1. Computational preliminaries

The reason why we have chosen chemical alkane trees (see Definition 2.5) is twofold. The first reason is the Tanimoto index has often been used in chemical graph theory and chemoinformatics for measuring the similarity between chemical structures [65,40,62]; for instance, the subgraphs chosen in [62] have a chemical meaning. The second reason is that the underlying graph classes should have a reasonable size in terms of the number of pairwise comparisons.

For this reason, we have chosen the sets of chemical alkane trees $\mathcal{C}(14)$ and $\mathcal{C}(15)$. By applying the program `Molgen` [1], we yield $|\mathcal{C}(14)| = 1858$ and $|\mathcal{C}(15)| = 4347$. Note that for these relatively small graph sets, the number of comparisons we have to perform are 1.725.153 and 9.446.031, respectively. In general, the number of comparisons to be performed equals $N = \frac{|\mathcal{G}|(|\mathcal{G}|-1)}{2}$; $\mathcal{G}$ is the underlying graph class. According to Definition 2.6, these classes contain all alkane trees with 14 and 15 vertices that have been generated exhaustively.

Also, the similarity/distance distributions are estimated by pairwisely calculating the structural similarity/distance between the graphs of a graph class $\mathcal{G}$ and determining the cumulative similarity/distance distributions (see Fig. 3–5). Suppose we apply the comparative graph measures M_1 and M_2 on $\mathcal{G}$ in the just described way. This leads to the value distributions D_{M_1} and D_{M_2}. As pairwisely calculating the structural similarity/distance of the graphs gives $N = \frac{|\mathcal{G}|(|\mathcal{G}|-1)}{2}$ similarity/distance values, we yield $|D_{M_1}| = |D_{M_2}|$. From D_{M_1} and D_{M_2}, the cumulative similarity/distance distributions are obtained.

Further we have used the programming language `R` [2] to compute the comparative graph measures; the Tanimoto index and GED. In order to do so, we first converted the structure information for both classes $\mathcal{C}(14)$ and $\mathcal{C}(15)$ originally present in Molfile format [30] to `graphNEL` objects. Then, the computation of the comparative graph measures has been performed on these `graphNEL` objects representing the alkane trees.

3.2. Distributions of the Tanimoto index by varying the subgraph set

We start our analysis by examining value distributions of the Tanimoto index by varying the set $\mathcal{S}(i)$. That means we investigate the influence of the chosen set of subgraphs $\mathcal{S}(i)$ on the Tanimoto index (see Eq. 2). To do so, we use the sets $\mathcal{S}(i)$ (see Definition 2.7). Each set $\mathcal{S}(i)$ contains all exhaustively generated non-isomorphic and connected trees up to i vertices, i.e., the classes $\mathcal{T}(2), \ldots, \mathcal{T}(i)$. From a mathematical point of view, it surely makes sense to use these exhaustively generated trees classes as set $\mathcal{S}$ because all those graphs could be possible induced subgraphs and the classes of chemical alkane trees $\mathcal{C}(14)$ and $\mathcal{C}(15)$ have been generated exhaustively too.

In Fig. 1 we show results obtained from the alkane trees of $\mathcal{C}(14)$ exemplarily. Note that X-axis is ranked by the number of comparisons $N = \frac{|\mathcal{G}|(|\mathcal{G}|-1)}{2}$. The case for $\mathcal{S}(4)$ is not shown as it's trivial; The usage of $\mathcal{C}(14)$ and $\mathcal{S}(7), \mathcal{S}(10), \mathcal{S}(14)$ shows Fig. 1.

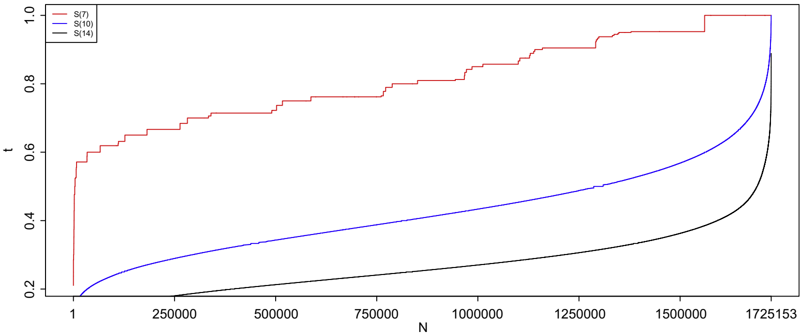

Fig. 1. Similarity values of the Tanimoto index t vs. number of graph comparisons by using $\mathcal{C}(14)$ and $\mathcal{S}(7), \mathcal{S}(10), \mathcal{S}(14)$ as set of subgraphs.

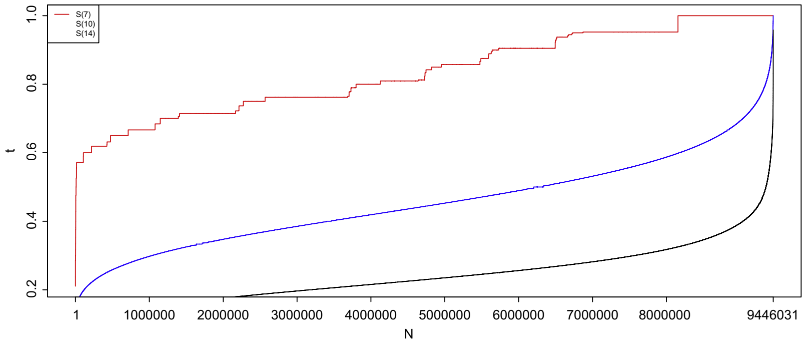

Fig. 2. Similarity values of the Tanimoto index t vs. number of graph comparisons by using $\mathcal{C}(15)$ and $\mathcal{S}(7), \mathcal{S}(10), \mathcal{S}(14)$ as set of subgraphs.

As the size of $\mathcal{S}(7)$ equals 24, not all graphs are subgraphs of the graphs $\in \mathcal{C}(14)$ and, therefore, the coverage of the value domain of t is much better than by using $\mathcal{S}(4)$ (as $t(G,H) = 1$ for almost all graphs). Evidently, this is even more significant when using $\mathcal{S}(10)$ and $\mathcal{S}(14)$, see Fig. 1. Altogether we conclude that all three measures (depending on $\mathcal{S}(i)$) are well defined and seem to be meaningful as they fully cover the value domain. However, this does not mean that the Tanimoto index defined in this form can solve a particular graph classification problem efficiently. Exploring this problem would be beyond of the scope of this paper. As Fig. 2 looks extremely similar, it holds the same relationship.

3.3. Interrelations between the Tanimoto index and GED

In order to explore interrelations between Tanimoto index and GED (see Section 2.2.3), we now depict their cumulative similarity distributions. In general, the Y-axis is the percentage rate of all graphs contained in the underlying set and the X-axis is the value range of both t and GED. Here, we use the normalized version of GED, see also [19].

Before evaluating them, we briefly give some arguments why we have chosen GED as a benchmark measure. The first reason relates to the fact that GED has a clear and simple interpretation as it is based on graph edit operations (see Section 2.2.3). That means, a graph is being transformed into another one by finding the minimal edit costs of the transformation sequence. Second, GED has been used in various disciplines [8,19,22] and, therefore, the properties and behavior of this measure seems to be widely understood. The third reason is of computational nature. As known, the classical version has exponential time complexity [8] and this would cause substantial difficulties when using large graphs.

We start with interpreting Fig. 3. The cumulative similarity distribution of GED looks staircase-shaped indicating that cluster exist whose graph pairs contain different similarity values based on a fixed percentage rate of graphs. For instance, there exist such cluster for approximately 35% of the graphs $\in \mathcal{C}(14)$ whose similarity values lie in value domain interval $[0.35, 0.42]$. In summary, we see that the two cumulative similarity distributions of both t and GED are considerably different. By considering Fig. 3, we observe the cumulative similarity distributions of the Tanimoto index t and GED by using $\mathcal{C}(14)$ and $\mathcal{S}(7)$. Again, we see that the two cumulative similarity distribution are quite different. For example when using GED, approximately 60% of the graphs $\in \mathcal{C}(14)$ possess similarity values $\leqslant 0.42$. By considering the Tanimoto index, approximately 60% of the graphs $\in \mathcal{C}(14)$ possess similarity values $\leqslant 0.85$. The cumulative similarity distributions of the Tanimoto index t and GED by using $\mathcal{C}(14)$ and $\mathcal{S}(10)$ are shown in Fig. 4. This result can be interpreted as a very rough approximation of GED by t based on their cumulative distributions. To compare this with $\in \mathcal{C}(14)$, see Fig. 5. Here the distribution for t is a bit shifted but could

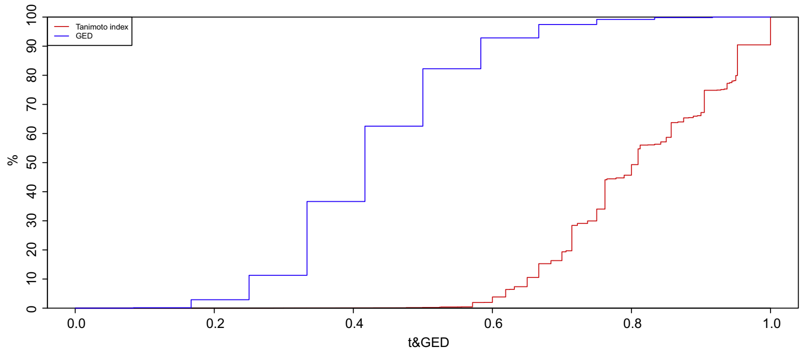

Fig. 3. Cumulative similarity distributions of the Tanimoto index t and GED by using $\mathcal{C}(14)$ and $\mathcal{S}(7)$.

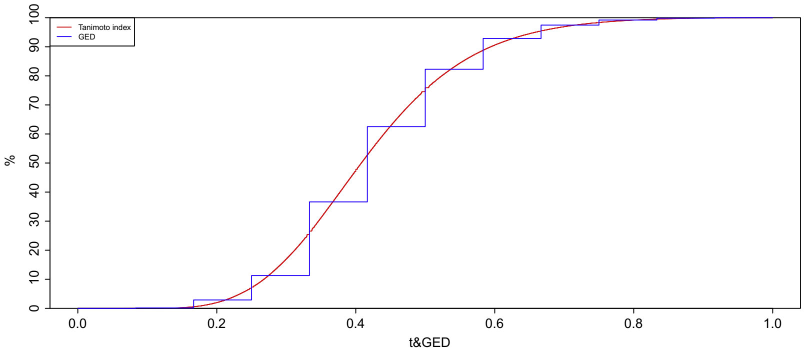

Fig. 4. Cumulative similarity distributions of the Tanimoto index t and GED by using $\mathcal{C}(14)$ and $\mathcal{S}(10)$.

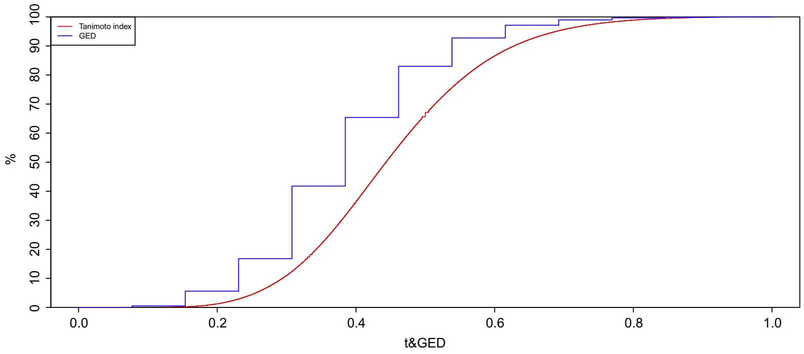

Fig. 5. Cumulative similarity distributions of the Tanimoto index t and GED by using $\mathcal{C}(15)$ and $\mathcal{S}(10)$.

be also seen as a very rough approximation of GED by the Tanimoto index. In contrast, the approximation behavior of the two distributions cannot be seen in Figs. 3 and 4.

We get a different picture when exploring interrelations between t and GED by using a different data analysis technique. By considering Figs. 6–8, we see that there is a rough linear relationship between t and GED for $\mathcal{C}(14)$. Again, these plots are staircase-shaped that induces the existence of clusters containing graph pairs with $\mathrm{GED}(G,H) \in [a,b]$, $0 < a, b < 1$ for a fixed value $t(G,H) \in [0,1]$. For instance, a large cluster can be seen in Fig. 6 for $\mathcal{C}(14)$ and $\mathcal{S}(7)$ for $t(G,H) \approx 1$. As the results for $\mathcal{C}(15)$ are extremely similar, they are not shown. In summary we conclude from these considerations that the curves can be interpreted as a rough linear approximation of t by GED.

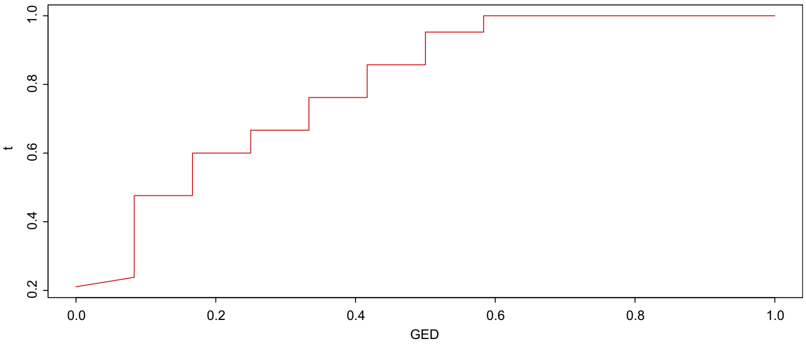

Fig. 6. Tanimoto index t vs. GED for $\mathcal{C}(14)$ and $\mathcal{S}(7)$ as set of subgraphs.

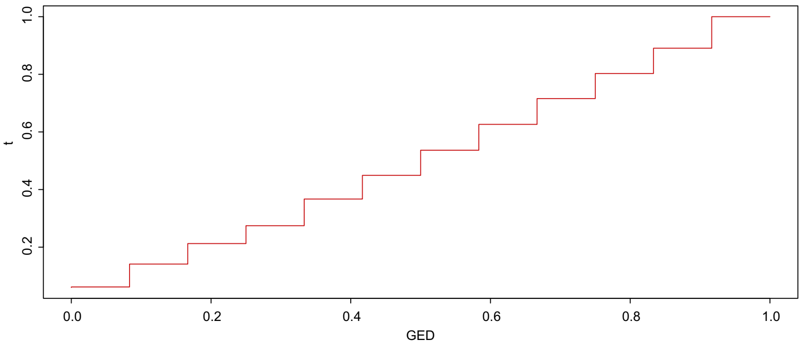

Fig. 7. Tanimoto index t vs. GED for $\mathcal{C}(14)$ and $\mathcal{S}(10)$ as set of subgraphs.

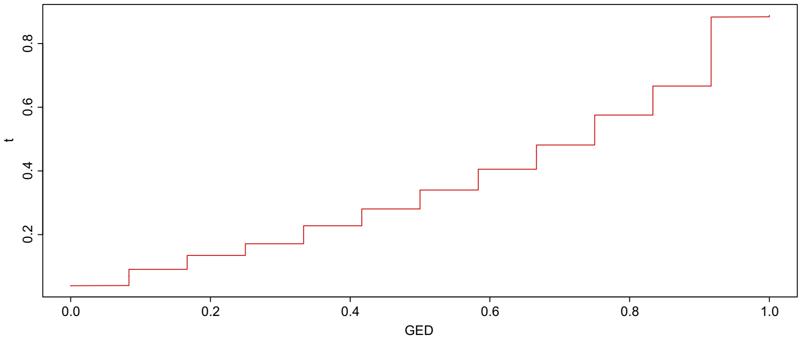

Fig. 8. Tanimoto index t vs. GED for $\mathcal{C}(14)$ and $\mathcal{S}(14)$ as set of subgraphs.

4. Summary and conclusion

In this paper, we explored the Tanimoto index when applied to chemical alkane trees. We investigated some properties of this index by using exhaustively generated trees as subgraphs. This extends earlier related work, see [65,40,62]. As a result, we found that the Tanimoto index is well-defined and possesses useful properties such as a good coverage of its value domain on the alkane trees. A question which is practically more relevant relates to examine the interrelations between the Tanimoto index and GED. As known, GED is NP-complete, see [49]; this may cause considerable drawbacks when calculating GED by using large networks. In this case, we believe that the Tanimoto index can serve as rough estimation as demonstrated in Section 3.3. Also, the computational complexity is polynomial when using a fixed set of subgraphs for calculating Eq. 2.

Future work would involve to repeat the study on general graphs (containing cycles) rather than on trees. It would be of considerable interest if the found interrelations would also hold for general graphs. Graph edit operations as used for the determination of GED exhibit similarities with chemical fragmentations of molecules (chemical bond cleavages). So, it may be fruitful to investigate this task with considering the found close relationship between GED and the Tanimoto index. Also, it would be interesting to prove extremal properties of the Tanimoto index.

Acknowledgments

Matthias Dehmer gratefully acknowledges financial support from the German Federal Ministry of Education and Research, Project RiKoV, Grant No. 13N12304 and from Austrian Science Funds, Project P26142. We thank Zhonglin Wang for help regarding the data analysis and fruitful discussions, and Adalbert Kerber (University Bayreuth) for providing the software Molgen.

References

[1] Molgen isomer generator software. <www.molgen.de>, 2000. Institute of Mathematics II, University of Bayreuth, Germany.
[2] R software, A Language and Environment for Statistical Computing, R Development Core Team, Foundation for Statistical Computing, Vienna, Austria, 2011. www.r-project.org.
[3] H.H. Bock, Automatische Klassifikation. Theoretische und praktische Methoden zur Gruppierung und Strukturierung von Daten, Studia Mathematica, Vandenhoeck & Ruprecht, Göttingen, 1974.
[4] D. Bonchev, Information Theoretic Indices for Characterization of Chemical Structures, Research Studies Press, Chichester, 1983.
[5] D. Bonchev, Information theoretic measures of complexity, in: R. Meyers (Ed.), Encyclopedia of Complexity and System Science, vol. 5, Springer, 2009, pp. 4820–4838.
[6] H. Bunke, What is the distance between graphs?, Bull EATCS 20 (1983) 35–39.
[7] H. Bunke, Error correcting graph matching: on the influence of the underlying cost function, IEEE Trans. Pattern Anal. Mach. Intell. 21 (9) (1999) 911–917.

[8] H. Bunke, Graph matching: theoretical foundations, algorithms, and applications, Proceedings of Vision Interface 2000 (2000) 82–88.
[9] H. Bunke, Recent developments in graph matching. in: 15th International Conference on Pattern Recognition, vol. 2, 2000, pp. 117–124.
[10] H. Bunke, G. Allermann, A Metric on graphs for structural pattern recognition. In EUSIPCO, editor, in: Proc. 2nd European Signal Processing Conference EUSIPCO, 1983, pp. 257–260.
[11] H. Bunke, M. Neuhaus, Graph matching. exact and error-tolerant methods and the automatic learning of edit costs, in: D. Cook, L.B. Holder (Eds.), Mining Graph Data, Wiley-Interscience, 2007, pp. 17–32.
[12] D. Buttler, A short survey of document structure similarity algorithms, in: International Conference on Internet Computing, 2004, pp. 3–9.
[13] M. Dehmer, Strukturelle Analyse web-basierter Dokumente, Multimedia und Telekooperation, Deutscher Universitäts Verlag, Wiesbaden, 2006.
[14] M. Dehmer, F. Emmert-Streib, Comparing large graphs efficiently by margins of feature vectors, Appl. Math. Comput. 188 (2) (2007) 1699–1710.
[15] M. Dehmer, F. Emmert-Streib, Y. Shi, Interrelations of graph distance measures based on topological indices, PLoS ONE 9 (2014) e94985.
[16] M. Dehmer, M. Grabner, K. Varmuza, Information indices with high discriminative power for graphs, PLoS ONE 7 (2012) e31214.
[17] M. Dehmer, A. Mehler, A new method of measuring similarity for a special class of directed graphs, Tatra Mountains Math. Publ. 36 (2007) 39–59.
[18] M.M. Deza, E. Deza, Encyclopedia of Distances, second ed., Springer, 2012.
[19] F. Emmert-Streib, M. Dehmer, Detecting pathological pathways of a complex disease by a comparative analysis of networks, in: F. Emmert-Streib, M. Dehmer (Eds.), Analysis of Microarray Data: A Network-Based Approach, Wiley-VCH, Weinheim, Germany, 2008, p. 285305.
[20] F. Emmert-Streib, M. Dehmer (Eds.), Analysis of Microarray Data: A Network-based Approach, Wiley VCH Publishing, 2010.
[21] F. Emmert-Streib, M. Dehmer, Identifying critical financial networks of the djia: Towards a network based index, Complexity 16 (1) (2010).
[22] F. Emmert-Streib, M. Dehmer, Networks for systems biology: conceptual connection of data and function, IET Syst. Biol. 5 (2011) 185–207.
[23] F. Emmert-Streib, M. Dehmer, J. Kilian, Classification of large graphs by a local tree decomposition, in: H.R. Arabnia et al., (Ed.), Proceedings of DMIN'05, International Conference on Data Mining, Las Vegas, USA, 2006, pp. 200–207.
[24] M.R. Garey, D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness. Series of Books in the Mathematical Sciences, W.H. Freeman, 1979.
[25] T. Gärtner, P.A. Flach, S. Wrobel, On graph kernels: Hardness results and efficient alternatives, in: COLT, 2003, pp. 29–143.
[26] J. Gasteiger, T. Engel, Chemoinformatics – A Textbook, Wiley VCH, Weinheim, Germany, 2003.
[27] D. Gernert, Measuring the similarity of complex structures by means of graph grammars, Bull. EATCS 7 (1979) 3–9.
[28] D. Gernert, Distance or similarity measures which respect the internal structure of the objects, Methods Oper. Res. 43 (1981) 329–335.
[29] D. Gernert, Graph grammars which generate graphs with specified properties, Bull. EATCS 13 (1981) 13–20.
[30] M. Grabner, K. Varmuza, M. Dehmer, Rmol: A toolset for transforming sd/molfile structure information into R objects, Source Code Biol. Med. 7 (2012) 1–4.
[31] F. Harary, Graph Theory, Addison Wesley Publishing Company, Reading, MA, USA, 1969.
[32] T. Horváth, T. Gärtner, S. Wrobel, Cyclic pattern kernels for predictive graph mining, in: Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 158–167.
[33] F. Kaden, Graph similarity and distances, in: Bodendiek, R. Henn (Eds.), Topics in Combinatorics and Graph Theory, Physica-Verlag, 1990, pp. 397–404.
[34] D.J. Klein, Graph geometry, graph metrics and wiener, MATCH Commun. Math. Comput. Chem. 35 (1997) 7–27.
[35] D.J. Klein, Resistance-distance sum rules, Croat. Chem. Acta 75 (2002) 633–649.
[36] D.J. Klein, M. Randić, Resistance distance, J. Math. Chem. 12 (1993) 81–95.
[37] D.J. Klein, H.-Y. Zhu, Distances and volumina for graphs, J. Math. Chem. 23 (1998) 179–195.
[38] X. Li, Y. Shi, A survey on the randić index, MATCH Commun. Math. Comput. Chem. 59 (1) (2008) 127–156.
[39] X. Li, Y. Shi, I. Gutman, Graph Energy, Springer, New York, 2012.
[40] G.M. Maggiora, V. Shanmugasundaram, Molecular similarity measures, in: Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery, Humana Press, Totowa, NJ, USA, 2004, pp. 1–50.
[41] B.D. McKay. Nauty. http://cs.anu.edu.au/~bdm/nauty/, 2010
[43] A. Mehler, P. Wei, A. Lücking, A network model of interpersonal alignment, Entropy 12 (6) (2010) 1440–1483.
[44] O. Mekenyan, D. Bonchev, A.T. Balaban, Unique description of chemical structures based on hierarchically ordered extended connectivities. v. new topological indices, ordering of graphs, and recognition of graph similarity, J. Comput. Chem. 84 (5) (1984) 629–639.
[45] L.A.J. Müller, K.G. Kugler, A. Graber, M. Dehmer, A network-based approach to classify the three domains of life, Biol. Direct 6 (2011) 140–141.
[46] N. Pržulj, Network comparison using graphlet degree distribution, Bioinformatics 23 (2007) e177–e183.
[47] M. Randić, Design of molecules with desired properties. molecular similarity approach to property optimization, in: M.A. Johnson, G. Maggiora (Eds.), Concepts and Applications of Molecular Similarity, Wiley, 1990, pp. 77–145.
[48] M. Randić, C.L. Wilkins, Graph theoretical approach to recognition of structural similarity in molecules, J. Chem. Inf. Comput. Sci. 19 (1979) 31–37.
[49] K. Riesen, H. Bunke, Approximate graph edit distance computation by means of bipartite graph matching, Image Vision Comput. 27 (2009) 950959.
[50] A. Robles-Kelly, R. Hancock. Edit distance from graph spectra. in: Proceedings of the IEEE International Conference on Computer Vision, 2003, pp. 234–241.
[51] I.L. Ruiz, M. Urbano-Cuadrado, M.A. Gómez-Nieto, Advantages of the approximate similarity approach in the QSAR prediction of ligand activities for alzheimer disease detection, World Congr. Eng. (2007) 165–170.
[52] C. Schädler. Die Ermittlung struktureller Ähnlichkeit und struktureller Merkmale bei komplexen Objekten: Ein konnektionistischer Ansatz und seine Anwendungen (Ph.D thesis), Technische Universität Berlin, 1999.
[53] L.B. Shams, M.J. Brady, S. Schaal, Graph matching vs mutual information maximization for object detection, Neural Networks 14 (3) (2001) 345–354.
[54] M.I. Skvortsova, I.I. Baskin, I.V. Stankevich, V.A. Palyulin, N.S. Zefirov. Molecular similarity in structure-property relationship studies. analytical description of the complete set of graph similarity measures, in: International symposium CACR-96. Book of Abstracts, 1996, pp. 16.
[55] M.I. Skvortsova, I.I. Baskin, I.V. Stankevich, V.A. Palyulin, N.S. Zefirov, Molecular similarity. 1. analytical description of the set of graph similarity measures, J. Chem. Inf. Comput. Sci. 38 (1998) 785–790.
[56] F. Sobik. Graphmetriken und Klassifikation strukturierter Objekte. ZKI-Informationen, Akad. Wiss. DDR, 2(82):63–122, 1982.
[57] F. Sobik. Modellierung von Vergleichsprozessen auf der Grundlage von Ähnlichkeitsmaen für Graphen. ZKI-Informationen, Akad. Wiss. DDR, 4:104–144, 1986.
[58] O. Sokolsky, S. Kannan, I. Lee, Simulation-Based Graph Similarity, in: TACAS, Springer, LNCS, 2006, pp. 426–440.
[59] E. Sommerfeld. Modellierung kognitiver strukturtransformationen auf der grundlage von graphtransformationen. ZKI-Informationen, Akad. Wiss. DDR, 4:1–103, 1984.
[60] Ch. Theoharatos, N. Laskaris, G. Economou, S. Fotopoulos. A similarity measure for color image retrieval and indexing based on the multivariate two sample problem. in: Proceedings of EUSIPCO, Vienna, Austria, 2004.
[61] N. Trinajstić, Chemical Graph Theory, C.R.C. Press, Boca Raton, FL, USA, 1992.
[62] K. Varmuza, W. Demuth, M. Karlovits, H. Scsibrany, Binary substructure descriptors for organic compounds, Croat. Chem. Acta 78 (2005) 141–149.
[63] K. Varmuza, H. Scsibrany, Substructure isomorphism matrix, J. Chem. Inf. Comput. Sci. 40 (2000) 308–313.
[64] Y. Wang. Molecular Complexity Effects and Fingerprint-Based Similarity Search Strategies [Ph.D thesis], Mathematisch-NaturwissenschaftlichenFakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn, 2009.
[65] P. Willet, Similarity and clustering in chemical information systems, Research Studies Press, Letchworth, United Kingdom, 1987.
[66] B. Zelinka, On a certain distance between isomorphism classes of graphs, Časopis pro pěst. Mathematiky 100 (1975) 371–373.
[67] A. Mehler, R. Gleim, M. Dehmer, Towards Structure-Sensitive Hypertext Categorization, in: M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, W. Gaul (Eds.), Proceedings of the 29th Annual Conference of the German Classification Society Universität Magdeburg, March 9-11, LNCS, Springer, BerlinNew York, 2005, pp. 406–413.