Application of a Similarity Measure for Graphs to Web-based Document Structures

Matthias Dehmer, Frank Emmert Streib, Alexander Mehler, Jürgen Kilian and Max Mühlhäuser

Abstract — Due to the tremendous amount of information provided by the World Wide Web (WWW) developing methods for mining the structure of web-based documents is of considerable interest. In this paper we present a similarity measure for graphs representing web-based hypertext structures. Our similarity measure is mainly based on a novel representation of a graph as linear integer strings, whose components represent structural properties of the graph. The similarity of two graphs is then defined as the optimal alignment of the underlying property strings. In this paper we apply the well known technique of sequence alignments for solving a novel and challenging problem: Measuring the structural similarity of generalized trees. In other words: We first transform our graphs considered as high dimensional objects in linear structures. Then we derive similarity values from the alignments of the property strings in order to measure the structural similarity of generalized trees. Hence, we transform a graph similarity problem to a string similarity problem for developing an efficient graph similarity measure. We demonstrate that our similarity measure captures important structural information by applying it to two different test sets consisting of graphs representing web-based document structures.

Keywords — Graph similarity, hierarchical and directed graphs, hypertext, generalized trees, web structure mining.

I. INTRODUCTION

Measuring the similarity between graphs is a challenging problem, especially for graphs of large orders the computation of similarity measures is quite involved [15], [20]. Many similarity measures of graphs are based on isomorphic relations and subgraph isomorphism [15], [20], respectively. Because it is well known that the subgraph isomorphism problem is NP-complete, the complexity of the underlying graph similarity measure is considered to be unacceptable for practical use [16]. Apart from measures that are based on isomorphic relations, there are methods in order to measure the structural similarity of trees. For example SELKOW [13] generalized the well known string edit distance [8], [14] to trees. Algorithmic improvements of tree alignment techniques for applications in Bioinformatics were proposed by, e.g., JIANG et al. [9] and ZHANG et al. [19]. But for the most graph similarity problems these methods are not suitable because the topology of the underlying graphs is more complex than the topology of trees. For this reason, we will present a new method for measuring the similarity of labeled/unlabeled, hierarchical and directed graphs. The term hierarchical means that there is an underlying directed rooted tree which forms a tree hierarchy. The main idea of our new method consists of two steps: (i) derivation of property strings from a graph and (ii) evaluating the alignment of the corresponding strings representing graphs. This is uniquely possible, because we restrict ourself to hierarchical graphs. In the following we call these graphs generalized trees. This graph class has been first introduced by MEHLER et al. [11] and is defined by

Definition 1.1: Let \( \mathcal{H} = (\hat{V}, E_\hat{h}) \) be an directed rooted tree. Let \( \nu_{\hat{h}} : \hat{V} \rightarrow A_{\hat{h}} \) be a vertex labeling function and \( A_{\hat{h}} \) denote a vertex alphabet. The vertex set \( \hat{V} \) is defined by

\[
\hat{V} := \{ \upsilon_{0,1}, \upsilon_{1,1}, \upsilon_{1,2}, \ldots, \upsilon_{1,\sigma_1}, \upsilon_{2,1}, \upsilon_{2,2}, \ldots, \upsilon_{2,\sigma_2}, \ldots, \upsilon_{h,1}, \upsilon_{h,2}, \ldots, \upsilon_{h,\sigma_h} \}
\]

where \( \upsilon_{i,j} \) denotes the \( j \)-th vertex on the \( i \)-th level, \( 0 \leq i \leq h \), \( 1 \leq j \leq \sigma_i \), \( h \) denotes the depth of \( \mathcal{H} \) and \( \sigma_i \) is the number of vertices on level \( i \). The edge set \( \hat{E} := \hat{E}_1 \cup \hat{E}_2 \cup \hat{E}_3 \cup \hat{E}_4 \) is defined as [11]:

- \( (\hat{E}_1) \) forms the edge set of the underlying directed rooted tree \( \mathcal{H} \).
- \( (\hat{E}_2) \) Up-edges associate analogously nodes of the tree hierarchy with one of their (dominating) predecessor nodes.
- \( (\hat{E}_3) \) Down-edges associate nodes of the tree hierarchy with one of their (dominant) successor nodes in terms of that tree hierarchy.
- \( (\hat{E}_4) \) Cross-edges associate nodes of the tree hierarchy, none of which is an (immediate) predecessor of the other in terms of the tree hierarchy.

If \( \hat{E}_2, \hat{E}_3, \hat{E}_4 \neq \emptyset \) then \( \mathcal{H} = (\hat{V}, \hat{E}, m_{\hat{h}}, A_{\hat{h}}) \) denotes a generalized tree. If we set \( A_{\hat{h}} = \emptyset \), the generalized tree is unlabeled (see Fig. (1)).

For the comparison of hypertext graphs there are simple graph theoretic indices (see Section (II)) like Multiplicity defined by WINNE et al. [17]. Multiplicity is defined as the ratio of the edge cut set of two graphs to the number of all possible edges. Because of this definition it is obvious that Multiplicity is not suitable for a comparison of the overall structure of a graph.

In contrast to those simple indices we design in Section (III) a powerful parametric model for measuring the structural similarity of web-based document structures representing generalized trees. We solve this graph similarity problem by first transforming the generalized trees into linear structures. Then we apply a dynamic programming [1] algorithm for determining optimal alignments of the corresponding property
strings. On the basis of the values of the resulting alignments we construct a graph similarity measure \( d \).

This paper is organized as follows: In the next section we repeat some graph theoretic measures for the structural analysis of hypertext structures introduced so far in the context of web-mining. We present in Section (III) our similarity measure mathematically and apply it in Section (IV) together with agglomerative clustering to two different data sets, each one containing web-based document structures. This paper finishes in Section (V) with a summary and conclusions.

II. GRAPH THEORETIC MEASURES FOR STRUCTURAL ANALYSIS OF HYPERTEXT STRUCTURES

In order to motivate our new method for measuring the similarity of generalized trees we first look at known measures for the structural analysis of hypertexts. We will see, that those measures are not suitable for a similarity-based structural analysis of graph-based documents, because they can not capture enough information that allows a meaningful clustering of web-based documents. In the field of structural analysis of hypertexts there are many measures that are called indices describing structural properties of hypertexts [2], [17].

The characteristic property of an index is that the described structural property, e.g., connectedness, is mapped on a normalized measured value. For example, BOTAFOGO et al. [2] defined the well known graph theoretic measure \( \text{Compactness} \) of a directed hypertext graph \( \mathcal{H} \) as

\[
C := \frac{(|V|^2 - |V|) \cdot K - \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} c_{ij}}{(|V|^2 - |V|) \cdot K - (|V|^2 - |V|)} \in [0, 1],
\]

which expresses how well connected a hypertext graph is. Here,

\[
(c_{ij})_{ij} := \begin{cases} 
    w_{ij} & \text{if } w_{ij} \text{ exists} \\
    K & \text{else},
\end{cases}
\]

denotes the converted distance matrix and \( w_{ij} \) denotes the shortest path from \( v_i \) to \( v_j \). \( K \) defines the conversion constant. \(|V|\) denotes the number of vertices of \( \mathcal{H} \) and BOTAFOGO et al. [2] set \( K = |V| \). From definition of \( C \) we conclude: \( C = 1 \) iff \( \mathcal{H} \) is completely connected, \( C = 0 \) iff \( \mathcal{H} = (V, \emptyset) \). Now, assume for two hypertext graphs \( \mathcal{H}_1, \mathcal{H}_2 \) holds \( C_1 \approx C_2 \). It is clear that the graph structures can be noticeable different. Therefore the index \( C \) is not suitable for determining intervals which contain similar hypertext structures. Hence, it is not possible to derive quality features, like "positive navigation behavior" from a certain value \( C \in [0, 1] \).

In contrast to this, the unsupervised learning approach we suggest extends and improves the concept of graph theoretic indices mentioned significantly. Our new approach for analyzing web-based document structures consists of two steps:

1) Developing a method for measuring the similarity of generalized trees. Up to now there are no contributions in the area of the structural analysis of hypertext data that present results in terms of measuring the similarity of document structures by comparing the overall graph structure.

2) Application of multivariate data analysis methods, e.g., clustering methods.

There are already known approaches determining the similarity of web-based documents, e.g., [5], [10]. CRUZ et al. [5] presented results about the structural similarity of DOM-trees [4] based on TDFA (Tag Frequency Distribution Analysis). JOSHI et al. [10] proposed an approach for measuring the structural similarity of web-based documents representing DOM-trees on the basis of the bag of path model. However, our similarity measure is completely different from all these contributions, in the sense that it captures more structural information of a graph.

III. SIMILARITY MEASURING OF GENERALIZED TREES

In this section we design a new similarity measure \( d \) for measuring the structural similarity of generalized trees. In the following we consider the unlabeled case, because the transition to the labeled case is possible by minor modifications (see Section (IV)).

The main idea of this similarity measure is based on the derivation of property strings for each generalized tree and then to align the property strings representing our generalized trees by a dynamic programming technique [1] (see Fig. (2)). From the resulting alignment one obtains a value of the scoring function, which is minimized during the alignment process. The similarity of two generalized trees will be expressed by a cumulative of local similarity functions which weighs two types of alignments: out-degree and in-degree alignments on a generalized tree level. Since we are examining hierarchical graphs, we take a closer look at the out-degree and in-degree sequences (on a level \( i \)), induced by the vertex sequences \( v_{1,i}, v_{2,i}, \ldots, v_{n,i} \), and their edge relations (see Fig. (2)). Now, the more similar with respect to a cost function \( \alpha \) the out-degree and in-degree alignments on the levels \( i \) are, the more similar is the common structure of the graphs. Define \( r_{j}^{\mathcal{H}_1} := v_{0,1}^{\mathcal{H}_1}, k \in \{1, 2\} \), and let \( \mathcal{H}_1 \) be a given graph and \( v_{h_1}, v_{h_2} \), \( 0 \leq i \leq h_1, 1 \leq j \leq \sigma_i \) denote the \( j \)-th vertex on the \( i \)-th level of \( \mathcal{H}_1 \), analogous to \( r_{j}^{\mathcal{H}_2} \) for \( \mathcal{H}_2 \). Then the problem of determining the structural similarity between \( \mathcal{H}_1 \) and \( \mathcal{H}_2 \) is equivalent to computing the optimal alignment of

\[
S_1 := r_{1}^{\mathcal{H}_1} \circ v_{1,1}^{\mathcal{H}_1} \circ v_{1,2}^{\mathcal{H}_1} \circ \cdots \circ v_{h_1,\sigma_1}^{\mathcal{H}_1}, \quad (1)
\]

\[
S_2 := r_{2}^{\mathcal{H}_2} \circ v_{1,1}^{\mathcal{H}_2} \circ v_{1,2}^{\mathcal{H}_2} \circ \cdots \circ v_{h_2,\sigma_2}^{\mathcal{H}_2}, \quad (2)
\]

with respect to a cost function \( \alpha \) which evaluates the alignments. Here, we distinguish different types of alignments:
For two generalized trees with their property strings. As an example the property string in terms of out-degrees of $H_1$ on level $i$ equals $\varnothing \circ 2^i$. The symbol $\varnothing$ denotes usual string concatenation.

$(v, v)$ (vertext-vertex), $(-, v)$ (gap-vertex), and $(v, -)$ (vertex-gap). In order to determine the optimal alignment between two given graphs, we consider the Sequences $(1)$, $(2)$ where $\bar{S}_k[i]$ denotes the $i$-th position of the sequence $\bar{S}_k$ and it holds $\bar{S}_1[n] = v_{\bar{S}_1}^1, \sigma_{\bar{S}_1}, \bar{S}_2[m] = v_{\bar{S}_2}^m, \sigma_{\bar{S}_2}$. In $\mathbb{N} \ni m, n \geq 1, \bar{S}_k[1] = v_{\bar{S}_k}^1, k \in \{1, 2\}$. The algorithm with the complexity $O(|\bar{V}_1| \cdot |\bar{V}_2|)$, for finding the optimal alignment of $\bar{S}_1$ and $\bar{S}_2$ generates a matrix $M(i, j)_r, 0 \leq i \leq n, 0 \leq j \leq m$. Now, we define the optimal alignment on the basis of the following dynamic programming algorithm [8]:

$$
M(0, 0) \equiv 0, \\
M(i, 0) \equiv M(i - 1, 0) + \alpha(S_1[i], -) : 1 \leq i \leq n, \\
M(0, j) \equiv M(0, j - 1) + \alpha(-, S_2[j]) : 1 \leq j \leq m,
$$

and

$$
\gamma^{out} = \gamma^{out}(i, \sigma_1^{out}, \sigma_2^{out}) \\
\gamma^{in} = \gamma^{in}(i, \sigma_1^{in}, \sigma_2^{in}),
$$

$\sigma_1^{out}, \sigma_1^{in} \in \mathbb{R}, k \in \{1, 2\}$ in a natural way and constructed a similarity measure $d \in [0, 1]$ (on the basis of these functions). $\gamma^{out}, \gamma^{in}$ are two-parametric functions, which detect the similarity of an outdegree and indegree alignment (on a level $i$). Finally, if we assume a set of units $U$ and a mapping $\phi : U \times U \longrightarrow [0, 1]$, we called $\phi$ a backward similarity measure if it satisfies the conditions

$$
\phi(u, v) = \phi(v, u), \forall u, v \in U \\
\phi(u, u) \geq \phi(u, v), \forall u, v \in U.
$$

Now we state the key result which has been proven in [6] for measuring the similarity for generalized trees.

**Theorem 3.1:** Let $\bar{H}_1, \bar{H}_2$, $0 \leq i \leq \rho$, $\rho := \max(h_1, h_2)$.

$$
d(\bar{H}_1, \bar{H}_2) := \frac{\prod_{\rho=1}^\rho \gamma^{fin}(i, \sigma_1^{out}, \sigma_2^{out}, \sigma_1^{in}, \sigma_2^{in})}{\sum_{\rho=1}^\rho \gamma^{fin}(i, \sigma_1^{out}, \sigma_2^{out}, \sigma_1^{in}, \sigma_2^{in})},
$$

is a backward similarity measure, where $\gamma^{fin}$ is defined as

$$
\gamma^{fin} = \gamma^{fin}(i, \sigma_1^{out}, \sigma_2^{out}, \sigma_1^{in}, \sigma_2^{in}) \\
= \zeta \cdot \gamma^{out} + (1 - \zeta) \cdot \gamma^{in}. \quad \zeta \in [0, 1].
$$

By construction we have $d(\bar{H}_1, \bar{H}_2) \in [0, 1]$. As a summary we note that our algorithm measures the similarity of two generalized trees by applying the technique of sequence alignments to outdegree and indegree sequences (on a level $i$). These alignments have both global and local significance. On the one hand, the sequence alignments will be implemented in a global sense, to compute the optimal alignment between the sequences $S_1$ and $S_2$. On the other hand, the alignments will be evaluated on the levels of the generalized trees by the function $\gamma^{fin}$. We note that the presented algorithm is suitable for the comparison of large generalized trees, because its complexity is enormously better then the complexity of methods which deal with isomorphic relations.

**IV. EXPERIMENTAL RESULTS**

We applied our method to web-based document structures representing generalized trees and, hence, to a problem from web structure mining. To perform the evaluation, we used a corpus $T_C$ which contains 500 conference websites from mathematics and computer science due to MEHLER et al. [11]. Starting from conference calendar websites, MEHLER et al. created the corpus with a java application that collects the conference links. Based on this set of links, we extracted the websites from the web by HyGraph [7]. Our evaluation is based on the two following steps:

1) **Examining the cumulative similarity distribution** on the basis of website structures, that is we represent a conference website $w \in T_C$ as an unlabeled generalized tree. This examination is impossible without a meaningful similarity measure which covers the complex graph structure of two graph-based documents. Thereby, the interpretation of the cumulative similarity distribution leads us to various applications.

2) Application of agglomerative clustering to the obtained similarity matrix where the web-based documents are represented by their DOM-structure [4]. Here, a DOM-tree is represented by a labeled generalized tree.

**Definition 4.1:** In terms of $T_C$ we define data classes $D_1-D_5$, which are manifest by the following parameter spectra:

- $D_1$: $\zeta = 1.0$ (solely alignments of Kernel-edges); parameter settings:
  $\sigma_1^{out} = 1.0, \sigma_2^{out} = 2.0, \sigma_1^{in} = 1.0, \sigma_2^{in} = 2.0$.
- $D_2$: $\zeta = 0.3$; parameter settings:
  $\sigma_1^{out} = 0.3, \sigma_2^{out} = 1.0, \sigma_1^{in} = 1.0, \sigma_2^{in} = 1.0$.
- $D_3$: $\zeta = 0.5$; parameter settings:
  $\sigma_1^{out} = 0.5, \sigma_2^{out} = 1.0, \sigma_1^{in} = 1.0, \sigma_2^{in} = 1.0$.
- $D_4$: $\zeta = 0.5$; parameter settings:
  $\sigma_1^{out} = 3.0, \sigma_2^{out} = 3.0, \sigma_1^{in} = 3.0, \sigma_2^{in} = 3.0$.
- $D_5$: $\zeta = 0.5$; parameter settings:
  $\sigma_1^{out} = 2.0, \sigma_2^{out} = 2.0, \sigma_1^{in} = 2.0, \sigma_2^{in} = 2.0$.

In general the computation of the cumulative similarity distribution of a corpus of graph-based hypertext structures opens new perspectives, e.g.:
Fig. 3. The x-axes corresponds to the values of \( d \in [0, 1] \) and the y-axes represents the cumulative similarity distributions for \( D_1-D_5 \).

- The distribution in terms of structural similarity of graph-based hypertexts is unknown, because there is no efficient graph similarity measure available, which capture enough structural information for measuring the structural similarity. Our graph similarity measure \( d \) together with the cumulative similarity distribution provides a better understanding of web-based hypertexts and their interactions.

- Suppose we have a test corpus \( T \) of graph-based hypertexts from a specific web-genre, e.g., conference websites. MEHLER et al. described in [12] the interesting problem of computing a structural prototype of \( T \) by a graph median. As a preprocessing step for the graph prototyping [12] we can decide, on the basis of the cumulative similarity distribution, how structurally different the web-based document structures are.

Here, we compute the cumulative similarity distribution of \( T_{C} \) based on Definition (4.1) in order to examine the navigation strategies in terms of a web-genre. We assume that a generalized tree reflects all possible navigation paths of a graph-based conference website. The computation and interpretation of the cumulative similarity distribution of \( T_{C} \) leads us to the question how different the navigation strategies within the specific web-genre are. As example, we choose the web-genre of conference websites. In order to discuss the cumulative similarity distribution of \( T_{C} \) (see Fig. (3)) we note that the data classes \( D_1-D_5 \) are manifest by the same corpus \( T_{C} \). We obtain a certain data class by only varying the parameters mentioned in Definition (4.1). Now, by varying the parameters we find the parameter tuple \( (\zeta, \sigma_{out}^{1}, \sigma_{out}^{2}, \sigma_{in}^{1}, \sigma_{in}^{2}) \) which captures enough structural information during the similarity measuring. In the following we notice that the plot of class \( D_1 \) differs in principle from the plots of data classes \( D_2-D_5 \). We recognize that, e.g., 20% of the conference websites have already the similarity value \( d \leq 0.5 \). Unlike 90% of the conference websites in \( D_2 \) have the similarity value \( d \leq 0.5 \). In summary, we conclude from Fig. (3), that the similarity values of the conference websites in \( D_1 \) were significantly higher compared to the conference websites of data classes \( D_2-D_5 \). This is plausible, because the conference websites in \( D_1 \) are treated solely as rooted trees without Cross-edges, Up-edges and Down-edges. Hence, the main part of the conference websites of \( D_1 \) is significantly less structurally different than the websites of the remaining data classes. In terms of \( D_2-D_5 \) the situation is inverted: In consideration of all types of conference websites the main part of the graph-based hypertext structures are structurally dissimilar on the basis of \( d \). The plot of \( D_4 \) equals the plot of \( D_5 \). Finally we notice that for the data classes \( D_2-D_5 \) the main part of all possible navigation strategies are very different within our web-genre. This is reflected by psychological features of hypertext navigation, e.g.:

- Certain strategies of treatment
- Existence of previous knowledge
- Specific user preferences

The next application consists of an evaluation of a set of websites \{\( W_1, W_2 \)\} which are treated as sets of their DOM-structures [4]. Now we have to modify our similarity measure \( d \) because our generalized trees are vertex labeled by HTML-markups. We define in a similar way

\[
\gamma_{in}(i, \sigma_{out}^{1}, \sigma_{out}^{2}, \sigma_{in}^{1}, \sigma_{in}^{2}) := (1 - \zeta_m) \cdot \gamma_{out}(i, \sigma_{out}^{1}, \sigma_{out}^{2}, \sigma_{in}^{1}, \sigma_{in}^{2}) + \zeta_m \cdot \gamma_{out}(i, \sigma_{in}^{1}, \sigma_{in}^{2})
\]

where \( \zeta_m \in [0, 1] \), \( \sigma_{in} \in \mathbb{R} \) and \( \gamma_{out}(i, \sigma_{in}) \) state a cumulative function which detects the similarity of vertex markup alignment on a level \( i \). With \( \zeta_m = 0.5 \) \( \sigma_{out}^{1} = 1.0, \sigma_{out}^{2} = 2.0 \) and \( \sigma_{in}^{1} = 1.0, \sigma_{in}^{2} = 2.0 \) and on the basis of a simple scheme which expresses the similarity relations between vertex labels, we apply an agglomerative clustering method\(^2\) [3] to the obtained similarity matrices \((d_{ij})_{ij}, 1 \leq i \leq |W_1|, 1 \leq j \leq |W_2| \) and \((d_{ij})_{ij}, 1 \leq i \leq |W_2|, 1 \leq j \leq |W_2|\) (\( d_{ij} \) denotes the number of extracted DOM-structures from \( W_1, W_2 \), respectively.

As a result we found cluster \( C_i \) which contain graph-based document structures where the \( C_i \) manifest structure

---


\(^2\)We used Average Linkage [3].

---

**TABLE I**

RESULTS OF PERFORMANCE EVALUATION OF THE CLUSTERING PROCEDURE CONCERNING \( W_1 \).

<table>
<thead>
<tr>
<th>( C_i )</th>
<th>Cluster</th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_1 )</td>
<td>staff web pages</td>
<td>84%</td>
<td>99%</td>
</tr>
<tr>
<td>( C_2 )</td>
<td>lecture announcements</td>
<td>76%</td>
<td>91%</td>
</tr>
<tr>
<td>( C_3 )</td>
<td>summary web pages</td>
<td>63%</td>
<td>92%</td>
</tr>
<tr>
<td>( C_4 )</td>
<td>lecture materials</td>
<td>85%</td>
<td>60%</td>
</tr>
<tr>
<td>( C_5 )</td>
<td>Download pages</td>
<td>72%</td>
<td>84%</td>
</tr>
</tbody>
</table>

**TABLE II**

RESULTS OF PERFORMANCE EVALUATION OF THE CLUSTERING PROCEDURE CONCERNING \( W_2 \).

<table>
<thead>
<tr>
<th>( C_i )</th>
<th>Cluster</th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_1 )</td>
<td>Staff web pages</td>
<td>99%</td>
<td>99%</td>
</tr>
<tr>
<td>( C_2 )</td>
<td>summary web pages</td>
<td>99%</td>
<td>99%</td>
</tr>
<tr>
<td>( C_3 )</td>
<td>seminar announcements</td>
<td>75%</td>
<td>99%</td>
</tr>
<tr>
<td>( C_4 )</td>
<td>lecture announcements</td>
<td>93%</td>
<td>93%</td>
</tr>
<tr>
<td>( C_5 )</td>
<td>techn. doc. pages</td>
<td>50%</td>
<td>99%</td>
</tr>
</tbody>
</table>
types, e.g., staff web pages, and special structured template web pages. We evaluated the performance of clustering with the well know measures recall = \[ \frac{|M \cap g|}{|M|} \] and precision = \[ \frac{|M \cap g|}{|g|} \] [3]. Here, \( M \) denotes the number of all relevant documents and \( M_g \) the number of documents found, based on a Cluster \( C \). The results in the Tables (I), (II) show clearly that the agglomerative clustering method created type classes, which contain structurally meaningful web pages. The high precision values express that the web pages found in a cluster \( C_i \) are actually relevant and the high recall values state that the web pages, which are relevant for Cluster \( C_i \), were actually found. In comparison with the remaining precision values the cluster \( C_3 \) in Table (II) achieves a lower precision value. However, \( C_3 \) has a high recall value. That means, (i) some web pages found related for \( C_3 \) are not relevant and (ii) all relevant web pages for \( C_3 \) were found. Regarding to recall the situation for \( C_4 \) in Table (I) is reversed. Hence, we have shown that \( d \) is successfully applicable for the structural filtering of web pages.

V. CONCLUSIONS

We started pointing out the problems that graph theoretic indices have in capturing important structural information of graphs. For this reason we designed a new method to measure the structural similarity of generalized trees. Our similarity measure is based on the representation of generalized trees as linear strings. We call these strings property strings, because their components represent structural properties of the generalized trees. The similarity of two generalized trees is then defined as the optimal alignment of the corresponding property strings of the generalized trees. From this definition it is clear, that our similarity measure is also different from measures, which are based on isomorphic relations [15], [20].

We demonstrated in Section (IV) that the similarity measure \( d \) has important applications in web structure mining [3]. We have shown that the cumulative similarity distribution provides useful information about the test corpus \( T_C \). On the basis of Definition (4.1) we answered the question how structurally different the conference websites of \( T_C \) are. In our experiment in Section (IV) the corpus \( T_C \) comprised 500 generalized trees and the average number of nodes per tree was 23. Because our similarity measure is a parametric measure depending on \( (\zeta, \zeta_{in}, \sigma_{out}, \sigma_{in}, \sigma_{out}, \sigma_{in}, \sigma_{out}, \sigma_{in}, \sigma_{out}) \) we are able to emphasize different structure types of generalized trees during the evaluation of the alignment. For example, by setting \( \zeta = 1 \) we consider an unlabeled hypertext structure as a directed rooted tree. More precisely, we align only the out-degree property strings induced by edges from the underlying directed rooted tree. If we set \( \zeta = 0 \), we align the property strings induced by in-degree sequences only. In most of the cases we used \( \zeta = \frac{1}{2} \), which weighs in- and out-degree sequences equally, \( \sigma_{in} = \sigma_{out} = \frac{1}{2} \). In the case of labeled generalized trees we considered in Section (IV) websites as sets of their DOM-trees \( (|W_1| + |W_2| = 270) \). We applied agglomerative clustering to the obtained similarity matrix and found type classes, which contain structurally meaningful web pages. We evaluate the clusters found by comparing them with labels manually assigning to the web pages and find high recall and high precision values. That means web pages found in a cluster \( C_i \) are actually relevant and web pages which are relevant for cluster \( C_i \) were actually found. Hence, our similarity measure \( d \) is applicable to filter web pages according to their structural organization. The interpretation of our results from the filtering process is that similar DOM-trees tend to have a similar content and layout elements. Furthermore, clusters tell us something about the meaning of web pages. Altogether we have found document groups which are comparable on the basis of their structure types.

REFERENCES