Measuring the Structural Similarity of Web-based Documents: A novel Approach

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

Abstract— Most known methods for measuring the structural similarity of document structures are based on, e.g., tag measures, path metrics and tree measures in terms of their DOM-Trees. Other methods measure the similarity in the framework of the well known vector space model. In contrast to these we present a new approach to measuring the structural similarity of web-based documents represented by so called generalized trees which are more general than DOM-Trees which represent only directed rooted trees. We will design a new similarity measure for graphs representing web-based hypertext structures.

Our similarity measure is mainly based on a novel representation of structures. Then we derive similarity values from the alignments of the structural properties of the graph. The similarity of two graphs is measured by applying it to a string similarity problem. We demonstrate that our similarity measure captures important structural information by applying it to two different test sets consisting of graphs representing web-based documents.

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

I. INTRODUCTION

The application of Data Mining methods [3] on web-based hypertext data is referred to as Web Mining [3]. In view of the vast amount of information available online Web Mining is an important and fruitful research field. Thereby Web Mining can be divided into three major fields (see Fig. (1)): Web Structure Mining, Web Usage Mining and Web Content Mining. In the area of Web Structure Mining we investigate graph based patterns by extracting web-based hypertext structures for measuring the structural similarity of those patterns. In this sense the paper sheds light on the task of the automatic analysis of web genre data. To measure the similarity between graphs is a challenging problem, especially for graphs of large orders the computation of similarity measures is quite involved [17], [23]. Many similarity measures of graphs are based on isomorphic relations and subgraph isomorphism [17], [23], 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 [18]. Apart from measures that are based on isomorphic relations, there are methods in order to measure the structural similarity of trees. For example SELKOW [15] generalized the well known edit distance measure [9], [16] to trees.

Algorithmic improvements of tree alignment techniques for applications in Bioinformatics were proposed by, e.g., JIANG et al. [10] and ZHANG et al. [22]. 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 in 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.

Definition 1.1: Let $\mathcal{H} = (V, E)$ be an directed rooted tree. Let $m_V : V \longrightarrow A_V$ be a vertex labeling function and $A_V$ denote a vertex alphabet. The vertex set $V$ is defined by

$$V := \{v_0, v_1, v_2, \ldots, v_{\sigma_1}, v_{\sigma_2}, v_{\sigma_3}, \ldots, v_h, v_{h+1}, v_{h+2}, \ldots, v_{h+\sigma_h}\}$$

where $v_{ij}$ 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 $E := E_1 \cup E_2 \cup E_3 \cup E_4$ is defined as [13]:

International Scholarly and Scientific Research & Innovation 1(10) 2007  3070

Abstract— Most known methods for measuring the structural similarity of document structures are based on, e.g., tag measures, path metrics and tree measures in terms of their DOM-Trees. Other methods measure the similarity in the framework of the well known vector space model. In contrast to these we present a new approach to measuring the structural similarity of web-based documents represented by so called generalized trees which are more general than DOM-Trees which represent only directed rooted trees. We will design a new similarity measure for graphs representing web-based hypertext structures. Our similarity measure is mainly based on a novel representation of a graph as strings of linear integers, whose components represent structural properties of the graph. The similarity of two graphs is then defined as the optimal alignment 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. We demonstrate that our similarity measure captures important structural information by applying it to two different test sets consisting of graphs representing web-based documents.

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

I. INTRODUCTION

The application of Data Mining methods [3] on web-based hypertext data is referred to as Web Mining [3]. In view of the vast amount of information available online Web Mining is an important and fruitful research field. Thereby Web Mining can be divided into three major fields (see Fig. (1)): Web Structure Mining, Web Usage Mining and Web Content Mining. In the area of Web Structure Mining we investigate graph based patterns by extracting web-based hypertext structures for measuring the structural similarity of those patterns. In this sense the paper sheds light on the task of the automatic analysis of web genre data. To measure the similarity between graphs is a challenging problem, especially for graphs of large orders the computation of similarity measures is quite involved [17], [23]. Many similarity measures of graphs are based on isomorphic relations and subgraph isomorphism [17], [23], 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 [18]. Apart from measures that are based on isomorphic relations, there are methods in order to measure the structural similarity of trees. For example SELKOW [15] generalized the well known edit distance measure [9], [16] to trees.

Algorithmic improvements of tree alignment techniques for applications in Bioinformatics were proposed by, e.g., JIANG et al. [10] and ZHANG et al. [22]. 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 in 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. [13] and is defined by
For the comparison of hypertext graphs there are simple indices we design in Section (IV) our similarity measure mathematically and apply it in Section (V) together with agglomerative clustering to two different data sets, each one containing web-based document structures. This paper finishes in Section (VI) with a summary and conclusions.

II. CONTENT-BASED HYPERTEXT ANALYSIS

An important task of automatically assigning labels to the categorized units, e.g., web pages. MEHLER et al. [13] state the hypothesis that polymorphism [13] and realizational ambiguity [13] are characteristic properties of web-based units. DEHMER et al. [7] and MEHLER et al. [13] performed an experiment in the framework of the well known vector space model [12] which focussed the content-based hypertext categorization on the basis of the data set (conference websites; see Section (V)) due to MEHLER et al. [13]. In order to categorize the plain text of the underlying web pages into a chosen category set a Support Vector Machine classification [7], [13] was performed. We receive low precision [7] and high recall [7] values. The low discriminatory power [13] between the categories obtained from the experiment indicates an extreme multiple categorization. These results sustainably support the hypothesis that polymorphism and realizational ambiguity are prevalent characteristics of web-based units [7], [13]. Hence, the necessity for exploring a new representation model of web-based document structures is given. Therefore in the present paper we investigate a new approach for modelling web-based documents not in the framework on the vector space model - that is to say on the basis on a graph-based representation. In the next Section (III) we give a preview of our new approach by repeating known graph-theoretic measures for examining structural properties of hypertexts.

III. MEASURES FOR THE STRUCTURAL ANALYSIS OF HYPERTEXTS

For motivating our new method for measuring the similarity of hypertext structures representing 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], [19]. The characteristic property of an index is that the described structural property, e.g., connectedness, is mapped on a normalised measured value. For example, BOTAFOGO et al. [2] defined the well known graph theoretic measure Compactness of a directed hypertext graph \(H\) as

\[
C := \frac{(|V|^2 - |V|) \cdot K - \sum_{e=1}^{\left |V \right |} \sum_{f=1}^{\left |V \right |} c_{ef} - (|V|^2 - |V|)}{(|V|^2 - |V|) \cdot K - (|V|^2 - |V|)} \in [0, 1],
\]

which expresses how well connected a hypertext graph is.
\[(c_{ij})_{ij} := \begin{cases} w_{ij} & : \text{if } w_{ij} \text{ exists} \\ \mathcal{K} & : \text{else} \end{cases}\]

denotes the converted distance matrix and \(w_{ij}\) denotes the shortest path from \(v_i\) to \(v_j\). \(\mathcal{K}\) defines the conversion constant, \(|V|\) denotes the number of vertices of \(\mathcal{H}\) and BOTAFOGO et al. [2] set \(\mathcal{K} = |V|\). From definition of \(\mathcal{K}\) 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], [11]. CRUZ et al. [5] presented results about the structural similarity of DOM-Trees [4] based on T DFA (Tag Frequency Distribution Analysis). JOSHI et al. [11] proposed an approach for measuring the structural similarity of web-based documents representing DOM-Trees on the basis of a 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.

IV. SIMILARITY MEASURING OF GENERALIZED TREES

Now, 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 (V)).

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. (3)). 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 cumulation 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 \(v_{i,1}, v_{i,2}, \ldots, v_{i,\sigma}\), and their edge relations (see Fig. (3)). 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 \(r_k^{\mathcal{H}_1} := v_{i,1}^{\mathcal{H}_1}, k \in \{1, 2\}\), and let \(\mathcal{H}_1\) be a given graph and \(v_{i,1}^{\mathcal{H}_1}, 0 \leq i \leq h_1, 1 \leq j \leq \sigma_i\) denote the \(j\)-th vertex on the \(i\)-th level of \(\mathcal{H}_1\), analogous to \(v_{i,2}^{\mathcal{H}_2}\) 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

\[S_1 := r_1^{\mathcal{H}_1} \circ v_{1,1}^{\mathcal{H}_1} \circ v_{1,2}^{\mathcal{H}_1} \cdots v_{h_1,\sigma_{h_1}}, \quad (1)\]

\[S_2 := r_2^{\mathcal{H}_2} \circ v_{1,1}^{\mathcal{H}_2} \circ v_{1,2}^{\mathcal{H}_2} \cdots v_{h_2,\sigma_{h_2}}, \quad (2)\]

with respect to a cost function \(\alpha\) which evaluates the alignments. Here, we distinguish different types of alignments: (v, v) (vertex-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 \(S_k[i]\) denotes the \(i\)-th position of the sequence \(S_k\) and it holds \(S_k[n] = v_{h_1,\sigma_{h_1}}, S_k[m] = v_{h_2,\sigma_{h_2}}\). Now, we define the optimal alignment on the basis of the following dynamic programming algorithm [9]:

\[M(0,0) := 0, \]

\[M(i,0) := M(i-1,0) + \alpha(S_k[i],-), \quad 1 \leq i \leq n, \]

\[M(0,j) := M(0,j-1) + \alpha(-,S_k[j]), \quad 1 \leq j \leq m, \]

\[M(i,j) := \min \{M(i-1,j) + \alpha(S_k[i],-) \]

\[M(i,j-1) + \alpha(-,S_k[j]) - \alpha(S_k[i],S_k[j]), \quad (3)\]

for \(1 \leq i \leq n, 1 \leq j \leq m\). In order to evaluate the alignments on each level, we defined [6] the functions

\[\gamma^{\text{out}} = \gamma^{\text{out}}(i,\sigma_i^{\text{out}},\sigma_2^{\text{out}}), \]

\[\gamma^{\text{in}} = \gamma^{\text{in}}(i,\sigma_i^{\text{in}},\sigma_2^{\text{in}}), \]

\[\sigma_i^{\text{out}},\sigma_i^{\text{in}} \in \mathbb{R}, k \in \{1, 2\}\] in a natural way and constructed a similarity measure \(d \in [0, 1]\) (on the basis of these functions). \(\gamma^{\text{out}},\gamma^{\text{in}}\) 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 \(\phi : U \times U \rightarrow [0,1]\), we called \(\phi\) a backward similarity measure if it satisfies the conditions

\[\phi(u,v) = \phi(v,u), \forall u, v \in U\]

and

\[\phi(u,v) \geq \phi(u,v), \forall u, v \in U.\]

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

Theorem 4.1: Let \(\mathcal{H}_1, \mathcal{H}_2, 0 \leq i \leq \rho, \rho := \max(h_1, h_2)\),

\[d(\mathcal{H}_1, \mathcal{H}_2) := \prod_{i=0}^{\rho} \gamma^{\text{out}}(i,\sigma_i^{\text{out}},\sigma_2^{\text{out}}), \quad (3)\]

\[\sum_{i=0}^{\rho} \gamma^{\text{in}}(i,\sigma_i^{\text{in}},\sigma_2^{\text{in}}), \quad (4)\]

International Scholarly and Scientific Research & Innovation 1(10) 2007 3072
Transformation into Strings (formal node sequences)

Applying String Alignment Techniques (evaluation of alignments)

Building a Family of Measures $d(\hat{H}_1, \hat{H}_2)$

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

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

By construction we have $d(\hat{H}_1, \hat{H}_2) \in [0, 1]$. 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 than the complexity of methods which deal with isomorphic relations. As a summary Fig. (4) shows the overall procedure for measuring the structural similarity of web-based documents representing generalized trees.

V. 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. [13]. 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 [8]. Our evaluation is based on the two following steps:

1) Examining the cumulative similarity distribution of $T_C$ 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
Fig. 5. The x-axes corresponds to the values of \( d \in [0, 1] \) and the y-axes represents the cumulative similarity distributions for \( D_1-D_5 \).

similarity matrix where the web-based documents are represented by their DOM-Trees [4]. Here, a DOM-Tree is represented by a labeled generalized tree.

**Definition 5.1:** 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:
  \( \sigma^1_{\text{out}} = 1.0, \sigma^1_{\text{in}} = 1.0, \sigma^2_{\text{in}} = 2.0. \)
- \( D_2: \zeta = 0.3 \); parameter settings:
  \( \sigma^1_{\text{out}} = 1.0, \sigma^1_{\text{in}} = 1.0, \sigma^2_{\text{in}} = 1.0, \sigma^2_{\text{out}} = 1.0. \)
- \( D_3: \zeta = 0.5 \); parameter settings:
  \( \sigma^1_{\text{out}} = 1.0, \sigma^1_{\text{in}} = 1.0, \sigma^2_{\text{in}} = 1.0, \sigma^2_{\text{out}} = 1.0. \)
- \( D_4: \zeta = 0.5 \); parameter settings:
  \( \sigma^1_{\text{out}} = 1.0, \sigma^1_{\text{in}} = 1.0, \sigma^2_{\text{in}} = 1.0, \sigma^2_{\text{out}} = 1.0. \)
- \( D_5: \zeta = 0.5 \); parameter settings:
  \( \sigma^1_{\text{out}} = 1.0, \sigma^1_{\text{in}} = 1.0, \sigma^2_{\text{in}} = 1.0, \sigma^2_{\text{out}} = 1.0. \)

In general the computation of the cumulative similarity distribution of a corpus of graph-based hypertext structures opens new perspectives, e.g.:

- 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 [14] the interesting problem of computing a structural prototype of \( T \) by a graph median. As a preprocessing step for the graph prototyping [14] 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 (5.1) 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. (5)) 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 (5.1). Now, by varying the parameters we find the parameter tuple \( (\zeta, \sigma^1_{\text{out}}, \sigma^1_{\text{in}}, \sigma^2_{\text{in}}, \sigma^2_{\text{out}}) \) 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. (5), 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 \( \{W_1, W_2\}^1 \) which are treated as sets of their DOM-Trees [4]. Now we have to modify our similarity measure \( d \)

We define in a similar way $C$ because our generalized trees are vertex labeled by $\text{HTML}$-markups. We define in a similar way

$$\zeta_m(i, \sigma_m) \cdot \gamma_m(i, \sigma_m),$$

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

As a result we found cluster $C_1$ which contain graph-based document structures where the $C_1$ manifest structure types, e.g., staff web pages, and special structured template web pages. We evaluated the performance of clustering with the well know measures $\text{recall} = \frac{|M_r \cap M_e|}{|M_r|}$ and $\text{precision} = \frac{|M_r \cap M_e|}{|M_r|}$ [3]. Here, $M_r$ denotes the number of all relevant documents and $M_e$ 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 $C_1$ are actually relevant and the high recall values state that the web pages, which are relevant for Cluster $C_1$, were actually found. In comparison with the remaining precision values the cluster $C_5$ in Table (II) achieves a lower precision value. However, $C_5$ has a high recall value. That means, (i) some web pages found related for $C_5$ are not relevant and (ii) all relevant web pages for $C_5$ 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.

VI. CONCLUSIONS

The main contributions of this paper are shown in Fig (6). Fig (6) shows the overall idea of our new approach for measuring the structural similarity of web-based documents represented by generalized trees. The first main idea from Fig (6) characterizes the transformation of web-based documents into hierarchical and directed graphs we call generalized trees. Then, by transforming the generalized trees into property strings we applied string alignment techniques for measuring the structural similarity of the generalized trees. Finally, by parameterization, that is the functions which evaluate the alignments have parameters, we obtain a high flexibility of our new graph similarity method. From this it is clear that it is possible to define new graph similarity measures $d$, which can be optimized by parameterization for special graph classes. In the present paper we examined only one similarity measure $d$ presented in Section (IV). In order to motivate our method we stated in Section (II) some results from hypertext catego-
rization. Following these results we did not concentrate on the content-based hypertext analysis but on the structural analysis of our web-documents. We repeated some problems that graph theoretic indices have in capturing important structural information of graphs. Then, we designed a new method to measure the structural similarity of generalized trees. The new 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 applicable to measures, which are based on isomorphic relations [17], [23]. We demonstrated in Section (V) 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 (5.1) we answered the question how structurally different the conference websites of \( T_C \) are. In our experiment in Section (V) 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 \( (\zeta,\zeta^{in},\sigma^{out},\bar{\sigma}^{out},\bar{\sigma}^{in},\bar{\sigma}^{in},\bar{\sigma}^{1},\bar{\sigma}^{2}) \) 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 a 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 weights in- and out-degree sequences equally, \( \gamma^{fin} = \frac{\sigma^{out}}{2} + \frac{\bar{\sigma}^{out}}{2} \). In the case of labeled generalized trees we considered in Section (V) websites as sets of their DOM-Trees \(|W_1| + |W_2| = 270\). We applied agglomerative clustering to the obtained similarity matrix and found type clusters, which contain structurally meaningful web pages. We evaluated the 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 found document groups which are comparable on the basis of their structure types.

REFERENCES