International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:10, 2007

Towards Clustering of Web-based Document Structures

Matthias Dehmer, Frank Emmert Streib, Jürgen Kilian and Andreas Zulauf

***Abstract*— Methods for organizing web data into groups in order to analyze web-based hypertext data and facilitate data availability are very important in terms of the number of documents available online. Thereby, the task of clustering web-based document structures has many applications, e.g., improving information retrieval on the web, better understanding of user navigation behavior, improving web users requests servicing, and increasing web information accessibility. In this paper we investigate a new approach for clustering web-based hypertexts on the basis of their graph structures. The hypertexts will be represented as so called generalized trees which are more general than usual directed rooted trees, e.g., DOM-Trees. As a important preprocessing step we measure the structural similarity between the generalized trees on the basis of a similarity measure d. Then, we apply agglomerative clustering to the obtained similarity matrix in order to create clusters of hypertext graph patterns representing navigation structures. In the present paper we will run our approach on a data set of hypertext structures and obtain good results in Web Structure Mining. Furthermore we outline the application of our approach in Web Usage Mining as future work.**

***Keywords*— Clustering methods, graph-based patterns, graph similarity, hypertext structures, web structure mining**

I. Introduction

Clustering is known as the task of organizing objects into certain groups (clusters) on the basis of perceived similarities in such a way that similar objects are in the same group and dissimilar web objects are in different groups. Often clustering methods [2], [9], [11] are used in order to classify complex objects in terms of exploratory data analyis or in many areas of sciences. Normally, at the beginning of the clustering process the number of the resulting clusters and the cluster distribution is unknown. Clustering methods are *unsupervised*, because the goal is to find an *optimal cluster solution* [2] without *a teacher*. For example Fig. (1) shows several cluster solutions. The two well kown major groups of clustering methods are partitioning [11] and hierarchical clustering [11], e.g, *agglomerative* clustering [11]. In this paper we will use agglomerative clustering for creating clusters of graph-based document structures. More formally, we can describe this task as follows: Let $D := \{d_1, d_2, \ldots, d_n\}$, $\mathbb{N} \ni n > 1$ be the set of documents in a certain representation, e.g., graph-based patterns. A cluster solution C_{fin}, $|C_{\text{fin}}| = k$ is now a disjoint decomposition of D, that is $C_{fin} := \{C_i \subseteq D | \, 1 \leq i \leq k\}$.

Matthias Dehmer is with the Technische Universität Darmstadt, 64289 Darmstadt, Germany, e-mail: dehmer@informatik.tu-darmstadt.de. Frank Emmert-Streib is with the Stowers Institute for Medical Research, 1000 E. 50th Street, Kansas City, MO 64110, USA, e-mail: fes@stowers-institute.org. Jürgen Kilian is with the Technische Universität Darmstadt, 64289 Darmstadt, Germany, e-mail: kilian@noteserver.org. Andreas Zulauf is with the Technische Universität Darmstadt, 64289 Darmstadt, Germany, e-mail: zulauf@rbg.informatik.tu-darmstadt.de.

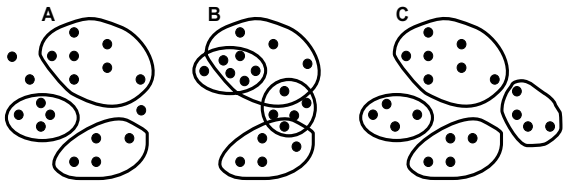

Fig. 1. A: Disjoint cluster solution but not partitioned. Futhermore A contains not groupable objects. B: Overpaping cluster solution. C: partitioned cluster solution.

Thereby the clusters C_i are construct in such a way that the elements $d \in C_i$ have a high similarity value to each other on the basis of a similarity measure $s : D \times D \longrightarrow [0, 1]$. In contrast to this the elements $d, \tilde{d}$ with $d \in C_i \wedge \tilde{d} \in C_j, i \neq j$ should have a low similarity to each other.

In this paper we describe an unsupervised learning approach in order to create clusters of web-based document structures on the basis of a graph-based representation model. This task that has many applications, e.g., *Web Mining* is challenging because we have to construct a meaningful similarity measure for measuring the similarity of graph-based patterns. We represent the graph patterns as *generalized trees* which have been introduced in [7]. As an example Fig. (2) shows a generalized tree together with his edge structure.

Since, we will compare graphs structurally we are looking for a similarity measure which is meaningful enough in order to apply clustering algorithms to the obtained similarity matrices. In [8] we introduced the concept of graph theoretic *indices* [3], [16] for measuring structural properties of hypertexts. The characteristic property of an index is that the described structural property, e.g., structural similarity of hypertext graph patterns, is mapped on a normalised measured value. Especially, for the comparison of hypertext graphs there are simple indices like *Multiplicity* defined in [16]. 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 obivous that Multiplicity is not suitable for a comparison of the overall structure of a graph. In contrast to those simple indices we propose in Section (II) a parametric model for measuring the structural similarity of web-based document structures representing generalized trees.

For creating clusters of web-based document structures representing generalized trees we first measure the structural similarity of our generalized trees described in Section (II). Second, we apply agglomerative clustering to the obtained similarity matrices. Our new approach for creating clusters of web-based document structures consists of two steps:

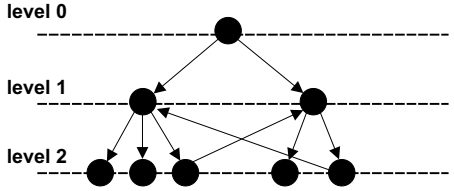

Fig. 2. A generalized tree together with his edge structure. A generalized tree is a hierarchical and directed graph. The term hierarchical means that for each genrealized tree there is an underlying directed rooted tree.

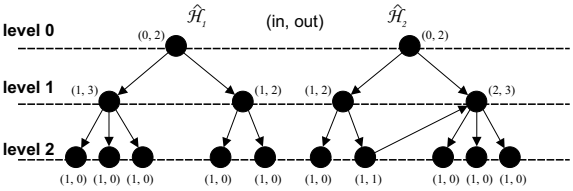

Fig. 3. Aligning level induced out-degree and in-degree sequences with respect to a cost function

1) Application of a new method [8] for measuring the similarity of generalized trees.
2) Application of a agglomerative clustering in order to generate clusters of graph-based documents.

These approach is generic in the sense that our approach for analyzing hypertext structures structurally can be applied in many fields of science which deal with graph-based instances. In Section (III) we will apply our two-step approach in *Web Structure Mining* [4]. In Section (IV) we will outline the application in *Web Usage Mining* [4] which is a subarea of *Web Mining* [4].

This paper is organized as follows: In the next section (Section (II)) we present our algorithm for measuring the structural similarity of web-based documents representing generalized trees. Section (III) presents the experimental results. In more detail, we apply in Section (III) our new approach to a new data set of hypertext structures created by MEHLER et al. [12]. We find clusters of graph-based instances which may be interpreted by psychological features of hypertexts. We finish our paper in Section (IV) with the conclusions.

II. MEASURING THE SIMILARITY OF WEB-BASED DOCUMENT STRUCTURES

In this section we present the construction of our new method for measuring the structural similarity [8] of unlabeled generalized trees which were first introduced in [12]. The class of generalized trees generalizes usual directed rooted trees in the sense that edges are allowed that jump over more than one graph level. The main idea [8] of our 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]. We call these strings property strings because their components represent structural properties of the generalized trees. From the resulting alignment we obtain 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 classes of alignments: *out-degree* and *in-degree* alignments on a generalized tree level. Since we are examining hierarchical graphs, we consider the out-degree and in-degree sequences (on a level i), induced by the vertex sequences $v_{i,1}, v_{i,2}, \ldots, v_{i,\sigma_i}$ and their edge relations (see Fig. (3)). Now, the more similar with respect to a *cost function* α the out-degree and in-degree sequences on the levels $i, 0 \leq i \leq h$ are, the more similar is the common structure of the generalized trees. Define $w_k^{\hat{\mathcal{H}}^k} := v_{0,1}^{\hat{\mathcal{H}}^k}, k \in \{1,2\}$, and let $\hat{\mathcal{H}}^1$ be a given generalized tree and $v_{i,j}^{\hat{\mathcal{H}}^1}$, $0 \leq i \leq h_1$, $1 \leq j \leq \sigma_i$ (upper index on a level i) denotes the j-th vertex on the i-th level of $\hat{\mathcal{H}}^1$, analogous to $v_{i,j}^{\hat{\mathcal{H}}^2}$ for $\hat{\mathcal{H}}^2$. Then the problem of determining the structural similarity between $\hat{\mathcal{H}}^1$ and $\hat{\mathcal{H}}^2$ is equivalent to determining the optimal alignment of

$$\begin{aligned} S_0^{\hat{\mathcal{H}}^1} &:= w_1^{\hat{\mathcal{H}}^1}, \\ S_1^{\hat{\mathcal{H}}^1} &:= v_{1,1}^{\hat{\mathcal{H}}^1} \circ v_{1,2}^{\hat{\mathcal{H}}^1} \circ \cdots \circ v_{i,\delta_{out}(w_1^{\hat{\mathcal{H}}^1})}^{\hat{\mathcal{H}}^1}, \\ &\vdots \\ S_{h_1}^{\hat{\mathcal{H}}^1} &:= v_{h_1,1}^{\hat{\mathcal{H}}^1} \circ v_{h_1,2}^{\hat{\mathcal{H}}^1} \circ \cdots \circ v_{h_1,\sigma_{h_1}}^{\hat{\mathcal{H}}^1}, \end{aligned}$$

and

$$\begin{aligned} S_0^{\hat{\mathcal{H}}^2} &:= w_2^{\hat{\mathcal{H}}^2}, \\ S_1^{\hat{\mathcal{H}}^2} &:= v_{1,1}^{\hat{\mathcal{H}}^2} \circ v_{1,2}^{\hat{\mathcal{H}}^2} \circ \cdots \circ v_{1,\delta_{out}(w_2^{\hat{\mathcal{H}}^2})}^{\hat{\mathcal{H}}^2}, \\ &\vdots \\ S_{h_2}^{\hat{\mathcal{H}}^2} &:= v_{h_2,1}^{\hat{\mathcal{H}}^2} \circ v_{h_2,2}^{\hat{\mathcal{H}}^2} \circ \cdots \circ v_{h_2,\sigma_{h_2}}^{\hat{\mathcal{H}}^2}, \end{aligned}$$

with respect to a *cost function* α. Thereby we distinguish different types of alignments: vertex-vertex, gap-vertex, vertex-gap. In order to determine the optimal alignment between two given generalized trees, we define the sequences

$$S_1 := w_1^{\hat{\mathcal{H}}^1} \circ v_{1,1}^{\hat{\mathcal{H}}^1} \circ v_{1,2}^{\hat{\mathcal{H}}^1} \circ \cdots \circ v_{h_1,\sigma_{h_1}}^{\hat{\mathcal{H}}^1}, \quad (1)$$

$$S_2 := w_2^{\hat{\mathcal{H}}^2} \circ v_{1,1}^{\hat{\mathcal{H}}^2} \circ v_{1,2}^{\hat{\mathcal{H}}^2} \circ \cdots \circ v_{h_2,\sigma_{h_2}}^{\hat{\mathcal{H}}^2}, \quad (2)$$

where $S_k[i]$ denotes the i-th position of the sequence S_k and it holds $S_1[n] = v_{h_1,\sigma_{h_1}}^{\hat{\mathcal{H}}^1}, S_2[m] = v_{h_2,\sigma_{h_2}}^{\hat{\mathcal{H}}^2}, \mathbb{N} \ni n, m \geq 1$, $S_k[1] = w_k^{\hat{\mathcal{H}}_k}$, $k \in \{1,2\}$. The algorithm with the complexity $O(|\hat{V}_1| \cdot |\hat{V}_2|)$ for finding the optimal alignment of the sequences generates a matrix $(\mathcal{M}(i,j))_{ij}, 0 \leq i \leq n, 0 \leq j \leq m$. Now, we define the optimal alignment on the basis of the following algorithm:

$$\begin{aligned} \mathcal{M}(0,0) &:= 0, \\ \mathcal{M}(i,0) &:= \mathcal{M}(i-1,0) + \alpha(S_1[i], -) \ : \ 1 \leq i \leq n, \\ \mathcal{M}(0,j) &:= \mathcal{M}(0,j-1) + \alpha(-, S_2[j]) \ : \ 1 \leq j \leq m, \end{aligned}$$

and

$$\mathcal{M}(i,j) := \min \begin{cases} \mathcal{M}(i-1,j) + \alpha(S_1[i], -) \\ \mathcal{M}(i,j-1) + \alpha(-, S_2[j]) \\ \mathcal{M}(i-1,j-1) + \alpha(S_1[i], S_2[j]), \end{cases}$$

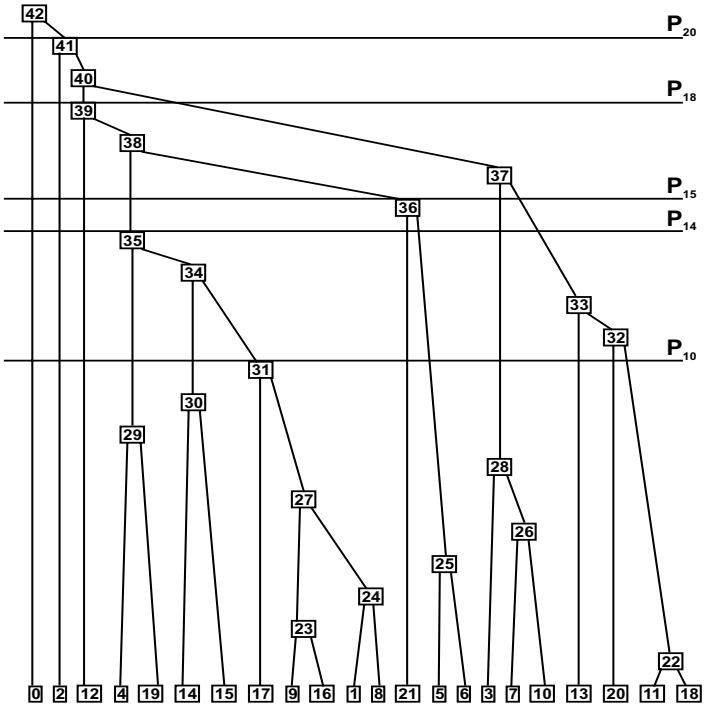

Fig. 4. Solution of the clustering process of T_{small} as dendogram. The 22 graphs are denoted with object numbers.

for $1 \leq i \leq n, 1 \leq j \leq m$. In order to construct the final similarity measure we have to construct two-parametric functions,

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

and

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

$\sigma_k^{out}, \sigma_k^{in} \in \mathbb{R}, k \in \{1, 2\}$, which detect the similarity of an out-degree and in-degree alignment (on a level i). The details about the construction of γ^{out} and γ^{in} can be found in [7]. For integrating our similarity measure d in the class of known similarity measures we express the definition of a *backward similarity measure*. Then we state a theorem which has been proven by DEHMER [7].

Definition 2.1: Starting from a set of units U and a mapping $\phi : U \times U \longrightarrow$ [0,1], we called ϕ a backward similarity measure if it satisfies the conditions $\phi(u, v) = \phi(v, u)$ and $\phi(u, u) \geq \phi(u, v), \forall\, u, v \in U$.

Theorem 2.1: Let $\hat{\mathcal{H}}_1, \hat{\mathcal{H}}_2$, be generalized trees, $0 \leq i \leq \rho$, $\rho := \max(h_1, h_2)$. h_1, h_2 denotes the height of $\hat{\mathcal{H}}_1, \hat{\mathcal{H}}_2$.

$$d(\hat{\mathcal{H}}_1, \hat{\mathcal{H}}_2) := \frac{\prod_{i=0}^{\rho} \gamma^{fin}(i, \sigma_1^{out}, \sigma_2^{out}, \sigma_1^{in}, \sigma_2^{in})}{\frac{\sum_{i=0}^{\rho} \gamma^{fin}(i, \sigma_1^{out}, \sigma_2^{out}, \sigma_1^{in}, \sigma_2^{in})}{\rho+1}} \in [0, 1],$$

$(d(\hat{\mathcal{H}}_1, \hat{\mathcal{H}}_1) = 1)$ is a backward similarity measure, where γ^{fin} is defined as

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

III. Experimental Results of Clustering Web-based Document Structures

In this section we apply our new two-step approach to web-based document structures represented by generalized trees by choosing a subset from the data set T_C due to MEHLER et al. [12]. T_C contains 500 hypertext structures where the hypertexts represents conference websites from engineering and computer science. Starting from conference calendar websites, MEHLER et al. generated the corpus on the basis of a java application that collects the conference links. Based on this link set, MEHLER et al. extracted the websites from the web by `HyGraph` due to GLEIM [10].

Now, the evaluation presented in this section is based on four steps:

1) In order to depict the cluster solution illustratively we create a sub set $T_{\text{small}} \subseteq T_C$ which cardinality is smaller than the cardinality of T_C. Starting from the data set T_C we construct $T_{\text{small}}, |T_{\text{small}}| = 22$ in such a way that the similarity values d_{ii} appear with the same cardinality and covers the interval $[0, 1]$ almost completly.
2) On the basis of T_{small} we calulate the similarity matrix $(d_{ij})_{ij}, 1 \leq i \leq |T_{\text{small}}|, 1 \leq j \leq |T_{\text{small}}|$. We choose the following parameter set[1]

$$\zeta = 0.5; \sigma_{out}^1 = 3.0, \sigma_{out}^2 = 5.0, \sigma_{in}^1 = 3.0, \sigma_{in}^2 = 5.0.$$

3) We apply agglomerative clustering to the computed

[1]This parameter set has been already sucessfully used in [7].

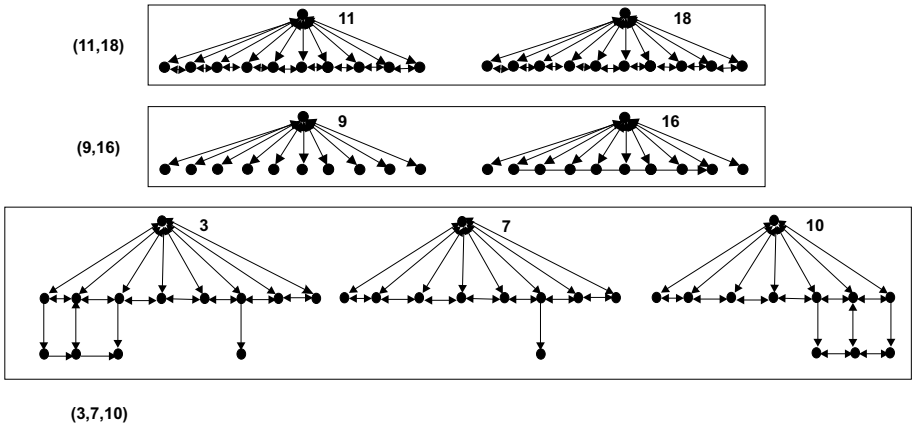

Fig. 5. Web-based hypertext graphs from clusters (11,18), (9,16) und (3,7,10). For measuring the structural simialrity of the respective generalized trees it holds $\zeta = 0.5$; $\sigma^1_{out} = 3.0, \sigma^2_{out} = 5.0, \sigma^1_{in} = 3.0, \sigma^2_{in} = 5.0$.

similarity matrix $(d_{ij})_{ij}$ and average linkage, thus

$$\alpha_{AL}\,(C_i, C_j) := \frac{1}{|C_i||C_j|} \sum_{\tilde{\mathcal{H}} \in C_i} \sum_{\mathcal{H} \in C_j} d(\mathcal{H}, \tilde{\mathcal{H}}).$$

4) We perform a quantitative cluster evaluation.

In the following we describe the quantitative cluster evaluation. Now, one can ask if there are partitions which can be considered as possible abort levels. In order to determine those levels in the resulting dendogram we used the abort criterion of RIEGER [15]:

1) The cluster distance on the j-th agglomeration step is denoted by α^j_{AL}.
2) One generate the values $\eta_i = |\alpha^j_{AL} - \alpha^{j+1}_{AL}|, i = 1, 2, \ldots, m-2$, where for each agglomeration step j denotes a vertex number (denotes the vertices in the resulting dendogram) $m = |T_{\text{small}}| + j$.
3) Now, for each new agglomeration step $j+1$ one can compute the respective cluster distances, where $j = 2, 3 \ldots, m-1$. On the basis of $\bar{\eta} = \frac{1}{m-2} \sum_{i=1}^{m-2} \eta_i$ and the standart deviation

$$\sigma = \sqrt{\sum_{i=1}^{m-2} (\eta_i - \bar{\eta})^2},$$

one can define the lower bound $\theta = \bar{\eta} + \frac{\sigma}{2}$.
4) All levels with $\eta_i \geq \theta$ are possible abort levels.

Fig. (4) shows the result of the clustering process of T_{small} together with the possible abort levels. Now, we combine the criterion of RIEGER with a criterion of *cluster homogeneity*. We call a cluster C_i homogeneously if the objects $o \in C_i$ are very similar to each other with respect to our similarity measure d. In order to measure the cluster homogeneity we state [2]

$$h(C_i) := \frac{1}{|C_i| \cdot (|C_i| - 1)} \sum_{\mu \in I_{C_i}} \sum_{\nu \in I_{C_i}} d_{\mu\nu} \in [0, 1],$$

where I_{C_i} denotes the corresponding index set. Because of the major feature of an agglomerative clustering method the value of cluster homogeneity decreases from the root node up to the leaf nodes in the resulting dendogram. Therefore we choose as a final abort criterion the calculated abort level stated above in combination with the highest remaining sum

$$\sum_{i=1}^{k} h(C_i),$$

of a partition $P = (C_1, C_2, \ldots C_k)$. According to this we choose in Fig. (4) the partition P_{10}: On one hand the criterion of RIEGER was satisfied. On the other hand P_{10} is the highest remaining value $\sum_{i=1}^{k} h(C_i)$ on P_{10}.

In order to get an impression how detailed the measure d reflects the structural similarity of our generalized trees within the obtained clusters we now look at Fig. (5). From our resulting partition P_{10} we chose exemplary the clusters with the object numbers $(11, 18), (9, 16)$ and $(3, 7, 10)$. For interpreting the results of the cluster solution we assume that a generalized tree reflects all possible navigation paths of a graph-based website of our web-genre under consideration, that is conference websites. According to the major feature of agglomerative clustering the cluster from the first agglomeration step contains the most similar graphs on the basis of the similarity measure d: In that case they are identical. The navigation pattern graphs from cluster $(9, 16)$ were generated in the second agglomeration step. They merely differ by one Across-edge [7] on the first generalized tree level. Otherwise the graph orders (number of nodes) and the edge sets of both graphs are indentical. Cluster $(3, 7, 10)$ was generated in a advanced agglomeration step. First, the cluster $(7, 10)$ was produced. In a further agglomeration step the graph with object number 3 was added. The graph structure of graphs 3, 7 is up to the first generalized tree level identical. Structural differences of the graph based navigation patterns appear at the second level. Compared to graph 3 and graph 7 graph 10 possesses the same graph structure up to the first level, but by one node reduced. In order to interpret the higher clusters in the dendogram hierarchy it holds: The higher a cluster in the dendogram hierarchy is the more structurally unsimilar are the graphs in this cluster to each other. Altogther, our main result of our experiments presented in this section is the detection of groups

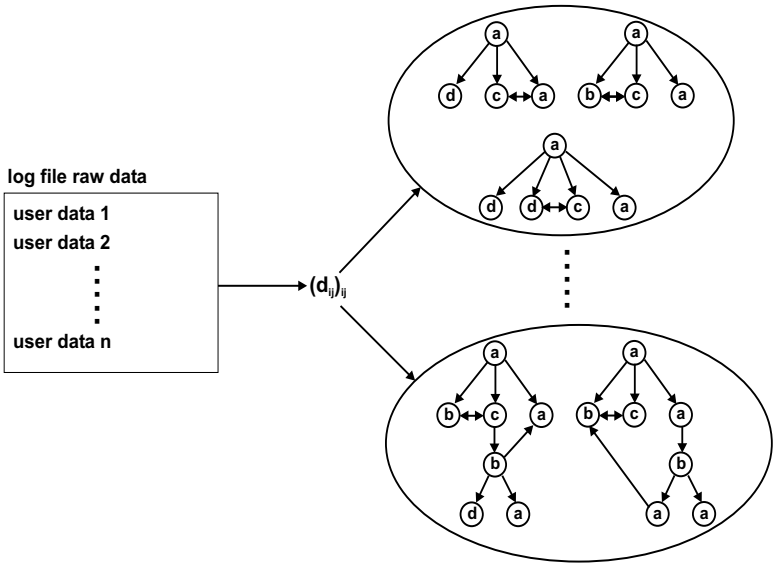

Fig. 6. This picture shows schematically the application of our new approach on the basis of log file data in Web Usage Mining.

of hypertext users represented by graph-based clusters. Each cluster contains navigation patterns representing generalized trees which can be distinguish by similar navigation behavior. From those clusters we can derive psychological features which reflect different navigation strategies. Compared to other methods for analyzing navigation behaviour, e.g., RICHTER et al. [14], our approach is based on a graph class - generalized trees - which is more meaningful than the usual graph models used in hypertext research. Because of the existing edges types [12] Up-Links, Down-Links, Across-Links and Kernel-links[2] generalized trees have a stronger semantical meaning than normal directed graph patterns used for describing hypertext navigation problems. For example RICHTER et al. analyzed in [14] hypertext structures representing directed graph patterns by graph-based indices, e.g. Compactness [3] and *Stratum* [3] mentioned in Section (I). Because of the weakness of those indices [8] RICHTER et al. can not create clusters of graph based patterns, because graph-based indices capture not enough structural information of the underlying graph patterns. The direct consequence of this is a dramatic loss of information. We avoid this information (and structure) loss by using our new graph similarity measure d. These was already succesfullly used in [8] for measuring the structural similarity of generalized trees in order to obtain a structural filtering of web- based documents on the basis of their DOM-Structures [5]. Another advantage of our new clustering approach for web-based documents is that we obtain a high level of abstraction because we receive clusters of graph-based instances which can be interpreted and formalized by graph theory.

IV. CONCLUSIONS AND FUTURE WORK

In this paper we presented a new approach in order to create clusters of graph-based hypertext structures. Instead of simple graph-theoretic indices [16] for measuring the structural similarity of hypertext graph patterns we used our graph similarity measure d introduced in [8]. This similarity measure d is based on the representation of generalized trees as linear integer strings we call property strings. We applied our new approach to a data set of web-based hypertext structures representing conference websites from mathematics and computer science. As an important result we found clusters which elements represent graph-based hypertext navigation patterns. One can interpret this result as similar navigation strategies of hypertext users represented by graph patterns. The main difference of our new method compared to other known approaches for clustering web-based documents, e.g., CRUZ et al. [6] is that we compare graphs structurally as a whole and then we apply clustering algorithms to the obtained similarity matrix. For example CRUZ et al. [6] represent web-based document structures as DOM-Structures [5] representing only usual directed rooted trees. Our graph class - generelized trees - is more general and therefore we can capture more structural information of our hypertext structures. A further advantage of our approach for clustering web-based documents is that we can apply the new method in several research areas which deal with graphs like generalized trees. As a future work we outline the application of our graph-based clustering approach in Web Usage Mining and *E-Learning* [13]. Fig. (6) shows schematically the application in Web Usage Mining:

1) From web server log files we will extract at first the raw user data of website navigation.
2) From this raw user data we will generate graph-based hypertext navigation patterns representing generalized trees.
3) Then we will compute the similarity matrix $(d_{ij})_{ij}$ on the basis of our similarity measure d.
4) Finally we apply clustering algorithms, e.g., agglomerative clustering to the obtained a similarity matrix.

[2]Kernel-Links [12] form the generalized tree hierarchy.

Hence, we will receive clusters of website users. The results of the interpretation of those clusters can be very useful in Web Usage Mining, e.g., for analyzing and optimizing customer behaviour. Finally, we hope that our work can enrich the area of clustering web-based documents - especially in the area of measuring the similarity of graphs which is still a challenging problem.

REFERENCES

[1] R. Bellman, *Dynamic Programming*. Princeton University Press, 1957
[2] H. H. Bock: *Automatische Klassifikation. Theoretische und praktische Methoden zur Gruppierung und Strukturierung von Daten*, Studia Mathematica - Mathematische Lehrbücher, Vandenhoeck & Ruprecht Verlag, 1974
[3] R. A. Botafogo, B. Shneiderman: *Structural analysis of hypertexts: Identifying hierarchies and useful metrics*, ACM Trans. Inf. Syst. 10 (2), 1992, 142-180
[4] S. Chakrabarti: *Mining the Web. Discovering Knowledge from Hypertext Data*, Morgen and Kaufmann Publishers, 2003
[5] S. Chakrabarti: *Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction*, Proc. of the 10th International World Wide Web Conference, Hong Kong, 2001, 211-220
[6] I. F. Cruz, S. Borisov, M. A. Marks, T. R. Webb: *Measuring Structural Similarity Among Web Documents: Preliminary Results* , Lecture Notes In Computer Science, Vol. 1375, 1998
[7] M. Dehmer, *Strukturelle Analyse web-basierter Dokumente*, Ph.D Thesis, Department of Computer Science, Technische Universität Darmstadt, 2005
[8] M. Dehmer, F. Emmert-Streib, A. Mehler, J. Kilian, M. Mühlhäuser, *Application of a similarity measure for graphs to web-based document structures*, International Conference on Data Analysis ICA 2005, in conjuction with the 7-th World Enformatika Conference, Budapest/Hungary
[9] B. S. Everitt, S. Landau, M. Leese: *Cluster Analysis*, Arnold Publishers; 4-th edition, 2001
[10] R. Gleim: *HyGraph – Ein Framework zur Extraktion, Repräsentation und Analyse webbasierter Hypertextstrukturen*, Beiträge zur GLDV-Tagung 2005, Bonn/Germany, 2005
[11] A. K. Jain, R. C. Dubes: *Algorithms for Clustering Data*, Prentice Hall, 1988
[12] A. Mehler, M. Dehmer, R. Gleim: *Towards logical hypertext structure. A graph-theoretic perspective*, Proc. of I2CS'04, Guadalajara/Mexico, Lecture Notes in Computer Science, Berlin-New York: Springer, 2004
[13] M. Mühlhäuser: *eLearning After Four Decades: What About Sustainability*?, Proceedings of ED-MEDIA 2004, 3694-3700
[14] T. Richter, J. Naumann, S. Noller: LOGPAT: *A semi-automatic way to analyze hypertext navigation behavior*, Swiss Journal of Psychology, Vol. 62, 2003, 113-120
[15] B. Rieger: *Unscharfe Semantik*, Peter Lang Verlag, 1989
[16] P. H. Winne., L. Gupta, J. C. Nesbit: *Exploring individual differences in studying strategies using graph theoretic statistics*, The Alberta Journal of Educational Research, Vol. 40, 177-193, 1994