CHALLENGES CONCERNING WEB DATA MINING

Wolfgang Gaul

Institute for Decision Theory and Operations Research

University of Karlsruhe
Germany

wolfgang.gaul@wiwi.uni-karlsruhe.de

Summary. For many WEB mining tasks very large sets of WEB data have to be analyzed. We give an overview concerning interesting WEB mining applications, sketch selected data analysis techniques that are appropriate for WEB data mining, and describe some new algorithms that allow to derive new solutions for WEB mining problems. Additional challenges concern the provision of results of WEB mining tasks, e.g., delivery and personalization. We will conclude with some hints for further research in WEB data mining.

1 MOTIVATION

1.1 From Data Analysis to WEB Mining

Data Analysis is a well-established area where scientific disciplines as, e.g., (in alphabetical order) computer science, operations research, optimization, and statistics intersect with many application fields (see, e.g., Studies in Classification, Data Analysis, and Knowledge Organization, a series which started in the beginning of the nineties with more than 30 volumes up to the middle of 2005). In the middle of the nineties a new label "Data Mining" was introduced into the discussion of how to analyze data, to draw attention to the special problems that arise when one tries to tackle very large data sets. Known data analysis techniques were checked for their ability to "mine" huge mountains of data and collected in so-called data mining software tools. The only new family of algorithms developed in the spirit of data mining was named "association rules" (see, e.g., [Gau98], [GS99], [GSch99]). An additional challenge concerning the analysis of large data was the everlasting growth of the WEB in terms of, e.g., amount of information, size of the net, and number of users. Of course, all data analysis techniques suited to handle voluminous data sets are candidates for solving WEB mining tasks. As the WEB provides means to

2006/1 (Invited Paper)
in Rizzi, A./Vichi, M. (Eds.), COMPSTAT 2006 –
Proceedings in Computational Statistics, Physica, 40
476

establish contacts to its users in many ways, questions concerning, e.g., delivery and personalization (that will be explained in the next subsection) are of importance. Figure 1 aims at summarizing the underlying situation in terms of an INPUT/OUTPUT description which would allow to classify WEB mining tasks by the input needed and the output provided (see, e.g., [GGHS02] for an input-output description of recommender systems). Sometimes, output of a first algorithm can be used as input for application of a second one (see, e.g., [GWB94]).

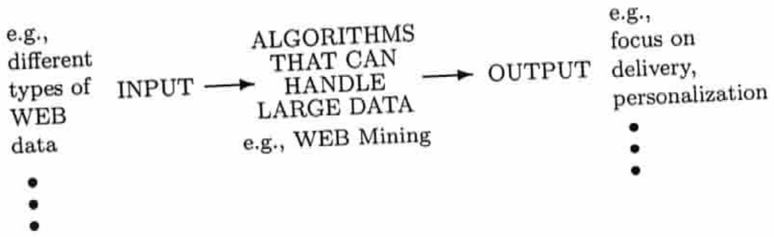
e.g., different types of WEB data
•
•
•
INPUT
ALGORITHMS THAT CAN HANDLE LARGE DATA
e.g., WEB Mining
OUTPUT
e.g., focus on delivery, personalization
•
•
•

Fig. 1. Input/Output Characterization of WEB Mining

1.2 Selected Topics with Respect to WEB Mining

A discussion about challenges concerning WEB data mining can be started from and focus on different points of view:

WEB Data

Descriptions of the different kinds of data available via the WEB can be used as an obvious starting point for considerations of how to extract interesting information. Here, a distinction into WEB usage data (e.g., server log files that document access information, i.e., which parts of WEB sites have been visited by users), WEB content data (e.g., selected (cuts of) text documents that contain certain content and have been demanded by site visitors), and WEB structure data (e.g., design of WEB link graphs and most frequently used IN and OUT links or – more generally – interesting subgraphs and counts of subgraph statistics described by user navigation) is usual.

WEB Mining Applications

From an application and problem oriented point of view one would be interested in, e.g., clickstream analysis (reconstruction of WEB user navigational behavior based on site servers' log files that list all HTTP-requests in the order they occur), WEB robot detection (robot access patterns superimpose human WEB visitor data and influence (and can distort) the analysis of WEB user behavior), recommender systems (software that combines external knowledge (e.g., about actual news and trends, etc., from the world "outside" the WEB) with internal information concerning site visitors (e.g., navigational behavior, preferences, etc., from the world "inside" the WEB) to deliver recommendations), text mining (analysis of content in WEB documents with the help of corresponding algorithms), or link graph design (structuring of the site graph according to frequently used pages) to mention just a few.

Data Analysis Techniques

In order to solve WEB mining problems of the just mentioned kind in whatever application areas, appropriate data analysis algorithms are needed. Association rules have already been mentioned. These techniques have been used not only for clickstream analysis (see subsection 2.1.1) and in recommender systems that support WEB navigational behavior but also in connection with decision tree construction for the analysis of WEB documents (see subsection 2.1.2) and compared with, e.g., support vector machine applications. Collaborative filtering is a standard approach in the area of recommender systems which has recently been compared with two-mode clustering (see subsection 2.2.2). Two-mode clustering with missing values (see subsection 2.2.1) is able to handle a typical situation concerning WEB applications. Although SVD (Singular Value Decomposition) has been applied to WEB mining and is well-known in terms of correspondence analysis or dual scaling (see, e.g., [NG88], [NG90]) we will not present it here. However given the target group for this paper, some algorithmic developments will be sketched in the next section.

New or Better Algorithms and Solutions for WEB Mining Tasks

Of course, new or better (compared to known counterparts) algorithms and solutions are of special interest. We will sketch the a-priori algorithm for generalized sequences (which was already used for improvements of recommender system output), DTARtext, a special case of DTAR (Decision Tree Construction by Association Rules) for text classification (which was already applied in electronic newspaper article selection within NEWSREC (NEWS RECommender System)), and two-mode clustering with missing values (by which traditional collaborative filtering could be improved).

A point that deserves special attention is concerned with output characteristics of WEB mining tasks. As one cannot assume that millions of WEB users are so similar concerning their interests that one standardized way of output would fit all needs, the degree of "personalization" with respect to the results of WEB mining is important. Another characteristic output feature is "delivery". Should targeted WEB users get results of WEB mining tasks unasked or only on demand? While questions of this kind touch important challenges concerning WEB mining, they will not be considered in the following where we focus on algorithmic aspects. Table 1 gives a summary of what has been discussed up to now. Of course, there is much more that could be listed in a frame like the one depicted by Table 1 and it is not all that has been tackled at the Institute for Decision Theory and Operations Research, Karlsruhe. Due to page restrictions for this paper most of the references cited are related to own contributions. Readers are asked to check these references

NEW or Better Solutions	WEB Mining Applications	WEB Data	Data Analysis Techniques	NEW or Better Algorithms
⋮ RS Output Improvements ⋮ NEWS REC ⋮ Output Characteristics Delivery & Personalization	USAGE Mining Clickstream Analysis WEB Robot Detection Recommender Systems (RS) Visualization of RS Results ⋮ CONTENT Mining TEXT Mining Online Visibility ⋮ STRUCTURE Mining Link Graph Design ⋮	USAGE WEB Navigation Data Conversion Rates eMetrics Buying Behavior Information ⋮ CONTENT WEB Documents ⋮ Evaluation Data STRUCTURE WEB Link Graphs IN Links OUT Links ⋮	Positioning MDS ⋮ Segmentation (Two-Mode) Clustering ⋮ Association Rules Collaborative Filtering Decision Trees Neural Nets Support Vector Machines SVD (Singular value Decomposition) ⋮	⋮ A priori Algorithm for Generalized Sequences ⋮ (TM) Clustering with Missing Values ⋮ DTARtext ⋮

Table 1. Selected Topics in WEB Mining

for background literature. For additional research see also, e.g., the chapter on "Electronic Data and Web" in [WG05].

2 Challenges Concerning Algorithmic Aspects

2.1 Association Rules as Starting Point

In order to describe techniques based on association rules the following notation is helpful:

R set of elements,

L list of (sub)sets ((sub)sequences, generalized (sub)sequences) composed of elements of R ($R \cup \{*\}$ where $*$ is called wildcard and explained in the next subsection),

$\preceq$ relation that denotes "... is subset of ..." ("... is subsequence of ...", "... is generalized subsequence of ..."),

$sup_L(c) = \frac{|\{l \in L \,|\, c \preceq l\}|}{|L|}$, where $|\cdot|$ denotes number of elements, is called support of $c \preceq l \in L$.

$minsup \in (0,1]$ is the abbreviation for minimal support and a lower bound for checking the frequency of a candidate subset (subsequence, generalized subsequence) c with respect to L. c with $sup_L(c) \geq minsup$ is called "frequent" and the aim of association rule techniques is to find all frequent subsets (subsequences, generalized subsequences) of elements from L. In early papers (see, e.g., [AS94]) L was a list of subsets (notice that in a list identical subsets can appear). A wellknown application for the determination of frequent sets (from the marketing area) is market basket mining where L describes a list of market baskets and the problem is to find frequent subsets of products bought together. For emphasis on marketing and market research see, e.g., [GGHS02] and [Gau04].

2.1.1 Clickstream Analysis

Instead of checking products collected in market baskets, in WEB mining one can look for frequently visited pages of a site. Now, WEB navigation sequences or clickstreams have to be considered. For this application the ordering of the elements in a sequence (successive requests of resources are stored in site server log files from which clickstreams can be reconstructed, see, e.g., [GST00]) is important. While sequence mining dates back to [AS95], the presentation of clickstreams via sequences is not sufficient in WEB usage mining as identical navigational behavior of WEB users does not appear frequently enough. If, however, one uses generalized sequences (see [GST00], [GST02a]), i.e., sequences in which subsequences are replaced by wildcards – symbolized by the $*$ notation – clickstream mining can be used for recommendations that support navigational behavior.

A crucial point within the algorithmic description of association rules is how subsets (subsequences, generalized subsequences) are selected as candidates c for frequency checking. Notice that for a candidate c to be frequent all subsets (subsequences, generalized subsequences) b with $b \preceq c$ have to be frequent. This implies the following pruning rule:

$$b \preceq c \quad \wedge \quad b \text{ is not frequent} \quad \Rightarrow \quad c \text{ is not frequent.}$$

For (generalized) sequences we need some notation:
Let $x = (x_1, x_2, \ldots, x_n)$ be a sequence of length $|x| = n$.
A pair of sequences $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_m)$ is overlapping on $k \in \mathbb{N}_0$ elements if the last k elements of x are equal to the first k elements of y ($k \leq \min\{m, n\}$). If x, y are k-overlapping, we denote

$$x +_k y = (x_1, \ldots, x_{n-k}, y_1, \ldots, y_m) = (x_1, \ldots, x_n, y_{k+1}, \ldots, y_m)$$

as k-telescoped concatenation of x and y (Note, that arbitrary sequences are always 0-overlapping and the 0-telescoped concatenation of these sequences is just their arrangement one behind the other.) and use

$$X +_k Y = \{x +_k y \mid x \in X,\ y \in Y \text{ are overlapping on } k \text{ elements}\}$$

as corresponding notation for sets.
Let F_n be the set of frequent subsequences of length n. Then, the candidate set for subsequences of length $n+1$ is easy to construct as

$$C^{n+1}_{sequences} = F_n +_{n-1} F_n.$$

In a generalized sequence $x = (x_1, \ldots, x_i, \ldots, x_n)$ elements x_i can be wildcards $*$ with the meaning that $*$ replaces a subsequence of x that is not interesting for further examination. Here, the following restriction with respect to the positioning of wildcards has to be obeyed

$$x_i = * \quad \Rightarrow \quad i \notin \{1, n\} \ \wedge \ x_{i-1} \neq *,\ x_{i+1} \neq *$$

as it does not make sense to start or end a sequence with a wildcard as well as to place wildcards next to each other. Now, the candidate set for generalized subsequences of length $n+1$ can be constructed as

$$C^{n+1}_{\substack{generalized\\ sequences}} = F_n +_{n-1} F_n \cup \ldots$$

$$\boxed{\begin{array}{l} \ldots \cup \{(x, *, y) \mid x, y \in F_1\} \\ \quad \cup \{x +_{n-2} y \mid x \in F_n,\ x_2 = *,\ y \in F_{n-1},\ n > 2\} \\ \quad \cup \{x +_{n-2} y \mid x \in F_{n-1},\ y \in F_n,\ y_{n-1} = *,\ n > 2\} \\ \quad \cup \{x +_{n-3} y \mid x \in F_{n-1},\ x_2 = *,\ y \in F_{n-1},\ y_{n-2} = *,\ n > 2\} \end{array}}$$

where additional to $F_n +_{n-1} F_n$ in all sets in the rectangular frame the positioning of wildcards is essential.

The concept of generalized subsequences can be further generalized to substructures of higher order, consider, e.g.,

frequent non-continuous subsequences of subsets

as interesting substructures of sequences of sets (see [STG01], [GST02b]).

Figure 2 depicts a situation in which a historical clickstream up to a WEB page p is used for recommendations concerning possible future visits of pages of the underlying site.

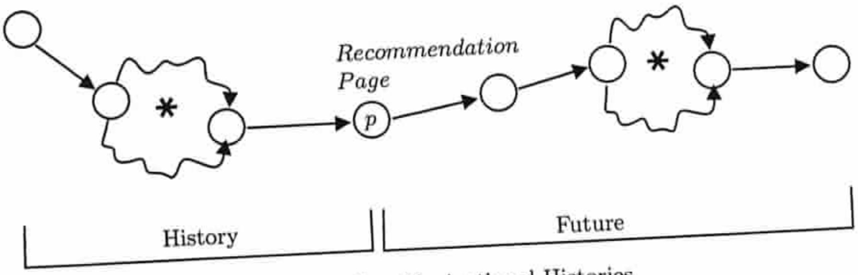

Fig. 2. Recommendations Based on Navigational Histories

2.1.2 WEB document analysis

Besides WEB usage data (by which, e.g., navigational behavior of site visitors is described) the mass of WEB documents asks for support tools (that based on the content of these documents help to classify (and select interesting) information).
The following notation is used in classical text classification:

n number of documents,
d_i document i in text representation, $i = 1, \ldots, n$,
m number of selected distinct words contained in all documents $D = (d_i)$,
w_j unique word, $j = 1, \ldots, m$,
$TF(w_j, d_i)$ number of occurrences of word w_j in document d_i (term frequency),
$BIN(w_j, d_j) =$ 1, if word w_j is contained in document d_i , 0 otherwise (binary),

$$IDF_j = \log\left(\frac{n}{\sum_{i=1}^{n} BIN(w_j, d_i)}\right) \quad \text{(inverse document frequency)},$$

$$R_j = \log(n) + \sum_{i=1}^{n} \frac{TF(w_j, d_i)}{\sum_{g=1}^{n} TF(w_j, d_g)} \log\left(\frac{TF(w_j, d_i)}{\sum_{g=1}^{n} TF(w_j, d_g)}\right)$$

(redundancy).

In order to perform text classification a three-step-preprocessing is needed.
Step 1 (Frequency Transformation):
A document d_i can be represented as document vector $\vec{d_i} = (d_{ij})$ where each component d_{ij} either describes $TF(w_j,d_i)$ (TF-notation), or $\log(1+TF(w_j,d_i))$ (LOG-notation), or $BIN(w_j, d_i)$ (BIN-notation).

Step 2 (Term Weighting):
Each component d_{ij} is multiplied by a weight which can be 1 (NOWEIGHTS-notation), IDF_j (IDF-notation), or R_j (REDUNDANCY-notation).

Step 3 (Normalization):
The document vector $\vec{d_i}$ can be normalized. Normalization can be skipped (NONE-notation), or $\frac{1}{\sum_j d_{ij}}$ (L1-notation) or $\frac{1}{\sqrt{\sum_j d_{ij}^2}}$ (L2-notation) is used.

The kind of preprocessing selected depends on the underlying data and influences the goodness-of-results obtained by different text document analysis procedures.

As SVM (Support Vector Machines) were found to be well suited for text classification and outperformed other methods like naive Bayes classifiers or C4.5 we started with SVM light ([Joa99]) with a linear kernel and also checked other techniques.

w	$label_1 = (+)$	$label_2 = (-)$	
documents of son_1	N_{11}	N_{12}	$N_{1\bullet}$
documents of son_2	N_{21}	N_{22}	$N_{2\bullet}$
	$N_{\bullet 1}$	$N_{\bullet 2}$	$N_{\bullet\bullet}$

Fig. 3. Contingency Table with respect to w

Additionally, we developed DTARtext (a special case of DTAR, see [SG00]) based on the following idea:
All documents are labeled positive ($label_1 = (+)$) or negative ($label_2 = (-)$), i.e., from the list D we derive $\vec{D} = (\vec{d_i})$ and, finally, the list of labeled documents (where l symbolizes the label) $\vec{D}^l = ((\vec{d_i}, l))$ which is used for DTARtext. All documents $\vec{D}_f^l$ of a father node of the underlying decision tree which contain a frequent subset of words w are collected in the documents of son_1, the rest of documents belongs to son_2. Figure 3 shows the contingency table of this situation, Fig. 4 depicts the corresponding part of the binary decision tree.
As split criterion one can use, e.g., the GINI-measure

$$\text{GINI}\,(w) = \sum_{i=1}^{2} \frac{N_{i\bullet}}{N_{\bullet\bullet}} \sum_{j=1}^{2} \left(\frac{N_{ij}}{N_{i\bullet}}\right)^2 - \sum_{j=1}^{2} \left(\frac{N_{\bullet j}}{N_{\bullet\bullet}}\right)^2$$

which can be rewritten in terms of support ($\approx sup$) and confidence ($\approx conf$) values determined with the help of an a-priori algorithm for sets (with a restriction with respect to $|w|$) where

$$conf_L(a,b) = \frac{sup_L(a \cup b)}{sup_L(a)}\ , \quad a, b,\ a \cup b \preceq l \in L\,.$$

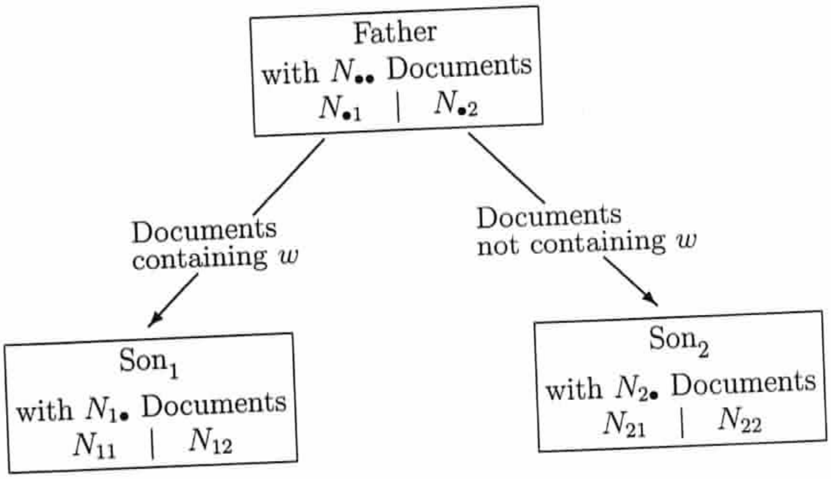
Father
with $N_{\bullet\bullet}$ Documents
$N_{\bullet 1}$ | $N_{\bullet 2}$
Documents containing w
Documents not containing w
Son$_1$
with $N_{1\bullet}$ Documents
N_{11} | N_{12}
Son$_2$
with $N_{2\bullet}$ Documents
N_{21} | N_{22}

Fig. 4. Part of Binary Decision Tree corresponding to Fig. 3

We get

$$\begin{aligned}\mathrm{GINI}\,(w) = {} & sup_{\vec{D}^l_f}(w) \sum_{j=1}^{2} \left(conf_{\vec{D}^l_f}(w, label_j)\right)^2 \\ & + \left(1 - sup_{\vec{D}^l_f}(w)\right) \sum_{j=1}^{2} \left(\frac{sup_{\vec{D}^l_f}(label_j) - sup_{\vec{D}^l_f}(w \cup label_j)}{1 - sup_{\vec{D}^l_f}(w)}\right)^2 \\ & - \sum_{j=1}^{2} \left(sup_{\vec{D}^l_f}(label_j)\right)^2\end{aligned}$$

Results will be presented in a forthcoming paper.

2.2 Two-Mode Clustering as Starting Point

In traditional clustering it is assumed that information about relationships between pairs of elements from a common set of a single mode is available and used for building clusters. Many applications, however, consist of two-mode (and even higher mode) data which describe relationships between elements of different modes. In the two-mode case, two-mode clustering is needed, i.e., one wants to cluster the elements of the first mode <u>and</u> the elements of the second mode based on the given two-mode data (see, e.g., [GSch96], [BGS97]).

2.2.1 Consideration of Missing Values in Two-Mode Data

In reality it may appear that relationships between pairs of elements are missing and the question is whether clustering algorithms can be adapted to handle such situations with incomplete information (see, e.g., [GSch94], [GSch96]).

For an algorithmic description the following notation is applied:

$i \in I$	first mode elements,
$j \in J$	second mode elements,
$k \in K$	first mode clusters,
$l \in L$	second mode clusters,
$S = (s_{ij})$	observed two-mode data matrix (with possibly missing values),
$\hat{S} = (\hat{s}_{ij})$	"best fitting" matrix with respect to S,
$P = (p_{ik})$	matrix describing first mode cluster membership with $p_{ik} \begin{cases} 1, & \text{first mode element } i \text{ belongs to first mode cluster } k, \\ 0, & \text{otherweise}, \end{cases}$
$Q = (q_{jl})$	matrix describing second mode cluster membership,
n_k (m_l)	number of first mode (second mode) elements in first mode (second mode) cluster k (l),
$W = (w_{kl})$	matrix of weights.

Now, different algorithmic approaches are possible:
In "fuzzy" procedures one assumes $p_{ik} \in [0,1]$ (and $q_{jl} \in [0,1]$) instead of zero-one cluster memberships.
In "overlapping" procedures the condition $\sum_{i \in I} p_{ik} = 1$ $(\sum_{j \in J} q_{jl} = 1)$ is replaced by $\sum_{i \in I} p_{ik} \geq 1$ $(\sum_{j \in J} q_{jl} \geq 1)$ for all $k \in K$ $(l \in L)$.
"Best fitting" is normally understood as optimizing the least squares criterion $\sum_{i \in I} \sum_{j \in J} (s_{ij} - \hat{s}_{ij})^2$ with $\sum_{k \in K} \sum_{l \in L} p_{ik}\, w_{kl}\, q_{jl}$.
In the "non-overlapping, no missing values" special case this criterion is minimized by $w_{kl} = \frac{1}{n_k\, m_l} \sum_{i \in I} \sum_{j \in J} p_{ik}\, s_{ij}\, q_{jl}$ $, k \in K,\ l \in L,$ for given matrices P and Q. Here, in order to find best arrangements of first mode clusters and second mode clusters, alterations of the matrices P and Q have to be checked.
The "missing values" case will be tackled in a forthcoming paper.

In WEB mining, one can imagine that WEB visitors (elements of a first mode) evaluate WEB documents (elements of a second mode). As the number of interesting WEB documents can become large, WEB visitors cannot check all documents, thus, the two-mode matrix S of observed relationships will have missing values.

As new WEB documents appear from time to time and also new WEB visitors show up over time a situation as described by Fig. 5 has to be evaluated. In case I (old) WEB visitors are clustered based on the existing (non-missing) evaluations with respect to (old) WEB documents, in case II new WEB documents have to be taken into consideration, in case III new WEB visitors are clustered together with old visitors, i.e., cluster memberships are reanalyzed.

Case IV describes the most difficult problem when evaluations of new WEB visitors with respect to new WEB documents are considered.

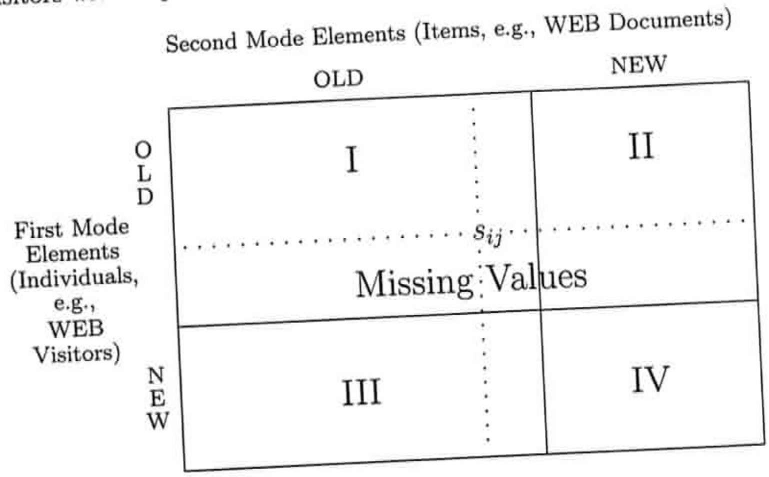
Second Mode Elements (Items, e.g., WEB Documents)
OLD
NEW
First Mode Elements (Individuals, e.g., WEB Visitors)
OLD
NEW
I
II
s_{ij}
Missing Values
III
IV

Fig. 5. Two-Mode Matrix with Missing Values

2.2.2 Comparisons of Fuzzy Two-Mode Clustering to Collaborative Filtering

Collaborative filtering is a technique that is well-known in WEB mining to determine recommendations for recommender systems and works on two mode data with missing values.

With J_j as set of items (e.g., WEB documents) that individual (e.g., a WEB visitor) i has rated, one gets for an actual individual a

$$COR\,(a,i) = \frac{\sum\limits_{j \in J_a \cap J_i} (s_{aj} - \bar{s}_a)(s_{ij} - \bar{s}_i)}{\sqrt{\sum\limits_{j \in J_a} (s_{aj} - \bar{s}_a)^2 \sum\limits_{j \in J_i} (s_{ij} - \bar{s}_i)^2}}$$

with $\bar{s}_i = \frac{1}{|\,J_i\,|} \sum\limits_{j \in J_i} s_{ij}$

and the CF (Collaborative Filtering) estimate

$$\hat{s}_{aj}^{CF} = \bar{s}_a + \frac{\sum\limits_{i \in I_j} COR\,(a,i)(s_{ij} - \bar{s}_i)}{\sum\limits_{i \in I_j} |\,COR\,(a,i)\,|}$$

where I_j denotes the set of individuals that have provided ratings for item j.

Now, $\hat{s}_{aj}^{CF}$ and $\hat{s}_{aj}^{FTMC}$ (with FTMC as abbreviation of Fuzzy Two-Mode Clustering) can be compared and first evaluations [SchG05] have shown that

Two-Mode Clustering gives better results (depending on the numbers K and L of the underlying first mode and second mode clusters) than collaborative filtering.

3 CONCLUSIONS AND FURTHER RESEARCH

Starting with an overview concerning selected topics with respect to WEB mining, this paper – due to page restrictions – emphasized only some aspects of mathematical modeling and algorithmic descriptions of WEB data analysis. As "association rules" is the label for a family of algorithms developed in the spirit of data mining these techniques were used as starting point for generalizations (to describe navigational behavior of internet users) and as underlying methodology for decision tree construction (to analyze WEB documents). As "two-mode data with missing values" is the type of information used for collaborative filtering (to calculate recommendations in the area of recommender systems), (fuzzy) two-mode clustering was introduced to show how known data analysis techniques can be favorably applied to solve this kind of WEB mining tasks. Further research concerning personal recommendations with respect to electronic newspaper article selection is just under consideration.

References

[AS94] Agrawal, R., Srikant, R., (1994), Fast Algorithms for Mining Association Rules. In Bocca, J.B., Jarke, M., Zaniolo, C. (Eds.), Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94), Santiago de Chile, Morgan Kaufmann, 487–499.

[AS95] Agrawal, R., Srikant, R., (1995), Mining Sequential Patterns. In Yu, P.S., Chen, A.L.P. (Eds.), Proceedings of the 11th International Conference on Data Engineering, Taipei, Taiwan, IEEE Computer Society, 3–14.

[BGS97] Baier, D., Gaul, W., Schader, M. (1997), Two Mode Overlapping Clustering With Applications to Simultaneous Benefit Segmentation and Market Structuring. In Klar, R., Opitz, O. (Eds.), Classification and Knowledge Organization, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, 557–566.

[Gau98] Gaul, W. (1998), Data Mining: A New Label for an Old Problem?, Proceedings of Workshop on Data Mining and Knowledge Discovery in Business Applications, Osaka, Japan.

[Gau04] Gaul, W. (2004), Market Research and the Rise Of the Web: The Challenge. In Wind, Yoram (Jerry), Green, Paul E. (Eds.), Market Research and Modeling: Progress and Prospects: A Tribute to Paul E. Green, International Series in Quantitative Marketing, Kluwer, 103–113.

[GGHS02] Gaul, W., Geyer-Schulz, A., Hahsler, M., Schmidt-Thieme, L. (2002), eMarketing mittels Recommendersystemen, Marketing ZFP, 24, 47–55.

[GS99] Gaul, W., Säuberlich, F. (1999), Classification and Positioning of Data Mining Tools. In Gaul, W., Locarek-Junge, H. (Eds.), Classification in the Information Age, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, 145–154.

[GSch94] Gaul, W., Schader, M. (1994), Pyramidal Classification Based on Incomplete Dissimilarity Data, Journal of Classification, 11, 171–193.

[GSch96] Gaul, W., Schader, M. (1996), A New Algorithm for Two-Mode Clustering. In Bock, H.H., Polasek, W. (Eds.), Data Analysis and Information Systems, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, 15–23.

[GSch99] Gaul, W., Schader, M. (1999), Data Mining: A New Label for an Old Problem? In Gaul, W., Schader, M. (Eds.), Mathematische Methoden in den Wirtschaftswissenschaften, Physica, 3–14.

[GST00] Gaul, W., Schmidt-Thieme, L. (2000), Frequent Generalized Subsequences – A Problem From Web Mining. In Gaul, W., Opitz, O., Schader, M. (Eds.), Data Analysis: Scientific Modeling and Practical Application, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, 429–445.

[GST02a] Gaul, W., Schmidt-Thieme, L. (2002a), Recommender Systems Based on User Navigational Behavior in the Internet, Behaviormetrika, 29, 1–22.

[GST02b] Gaul, W., Schmidt-Thieme, L. (2002b), Mining Web Navigation Path Fragments. In Nishisato, S., Baba, Y., Bozdogan, H., Kanefuji, K. (Eds.), Measurement and Multivariate Analysis, Springer, 249–260.

[GWB94] Gaul, W., Wartenberg, F., Baier, D. (1994), Comparing Proposals for the Solution of Data Analysis Problems in a Knowledge-Based System, Annals of Operations Research, 52, 131–150.

[Joa99] Joachims, T. (1999), Making Large Scale SVM Learning Practical. Advances in Kernel Methods. In Schölkopf, B., Burges, C., Smola, A. (Eds.), Support Vector Learning, MIT-Press.

[NG88] Nishisato, S., Gaul, W. (1988), Marketing Data Analysis by Dual Scaling, International Journal of Research in Marketing, 5 (3), 151–170.

[NG90] Nishisato, S., Gaul, W. (1990), An Approach to Marketing Data Analysis. The Forced Classification Procedure of Dual Scaling, Journal of Marketing Research, 27, 354–360.

[SG00] Säuberlich, F., Gaul, W. (2000), Decision Tree Construction by Association Rules. In Decker, R., Gaul, W. (Eds.), Classification and Information Processing at the Turn of the Millennium, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, 245–253.

[SchG05] Schlecht, V., Gaul, W. (2005), Fuzzy Two-Mode Clustering vs. Collaborative Filtering. In Weihs, C., Gaul, W. (Eds.), Classification – the Ubiquitous Challenge, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, 410-417.

[STG01] Schmidt-Thieme, L, Gaul, W. (2001), Frequent Substructures in WEB Usage Data – A Unified Approach, Workshop on Web Mining, 1st SIAM International Conference on Data Mining, Chicago, 15–23.

[WG05] Weihs, C., Gaul, W. (Eds.) (2005), Classification – the Ubiquitous Challenge, Studies in Classification, Data Analysis, and Knowledge Organization, Springer.