WMHAS Model for Improvement Document Classification

IOAN POP
Department of Computer Science
Lucian Blaga University
5-7 Ioan Ratiu Str, Sibiu
ROMANIA

Abstract: The web adaptation and webpages personalization for the needs of a specific user is today's trend of web technologies. In this paper we present an algebraic model for the adaptive classification methods with robust, efficient and simple to use features for the document classification algorithms. The algebraic model, named WMHAS, is a general framework for using data mining tasks and especially for combining classification document algorithms.

Key-Words: algebraic model, document classification, adaptive website.

1 Introduction

A uniform view on different operators in Web Mining (WM) is an important point in supporting WM processes in a adaptive hypermedia system. Such a WM framework requires a comprehensive data model and a sufficient set of operators supporting different kinds of pattern and rules as well as operations.

In hypertext, or hypermedia, the user has many ways to navigate between different information objects. The adaptation refers to the fact that the application changes its behavior based on the context in which it is used. In this paper we will always refer to this context as a user model. However, the context may model not just aspects of the user of the application, but also information about the place where it is being used, the time, the device used for interaction, or any other contextual aspect, like the weather, the recent news, etc. An adaptive application can change the information it shows, depending on the user model. It can change many aspects of this information, like the media used, the length of the presentation, the difficulty, style, etc. An adaptive hypermedia application, being hypermedia, may also change the links it offers to the user, or the presentation of these links. Also, the adaptive web applications can have many benefits by using of this framework that assures interoperability between different web mining operations such as the algorithms for document classification techniques: naive Bayes classifier, tf-idf (term frequency – inverse document frequency), latent semantic indexing, support vector machines, artificial neural network, kNN (k-nearest neighbor), decision trees, Concept Mining, approaches based on natural language processing [3].

The problem of adaptation at the user model is solved by an algebraic model called *Web Mining Heterogeneous Algebraically Structures hierarchies* (shortly WMHAS). Our model helps mining website usage, content and structure, performs different mining tasks, using as input the website's access logs, its structure and the content of its pages.

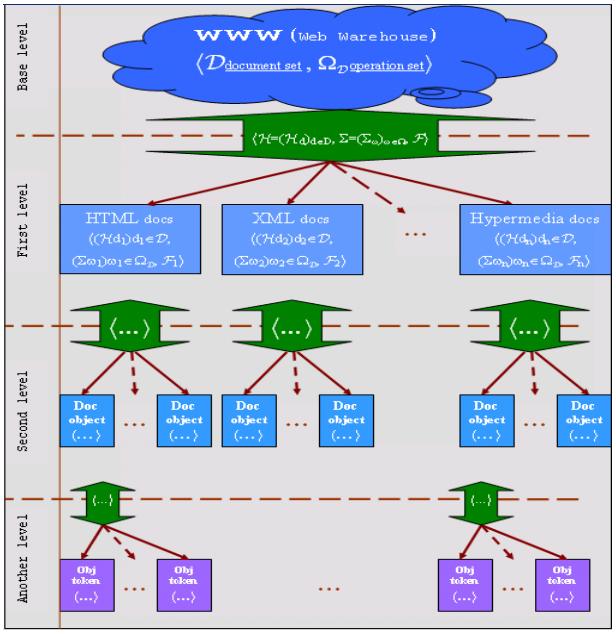

Fig. 1: Simple schemata of the heterogeneous algebraically structures hierarchy for the Web space [2].

The model is based on heterogeneous algebraically structure hierarchies. Our goals are to provide a modular, flexible, extensible, and scalable testbed. Nevertheless, we believe that our approach will allow faster analysis and hence, more results.

To prove the model's characteristics we use a Web repository. In the Web space, Web data set is modelled like more levels of Web repository, where each level represents a universal algebra of

the Web objects and operations (procedures, methods, functions, routines) which acts on these objects. In Figure 1 is presented a simple schema of the Heterogeneous Algebraically Structures hierarchy for the Web space.

The Hypertext, as a virtual object system managed by the Web, gives the possibility of linking documents that are related by words and phrases. This heterogeneous of the data needs a model that best reflects reality, and this can by made by an algebraic model which we will later name Web Mining Heterogeneous Algebraic Structure hierarchy - WMHAS.

The levelled algebraic modelling has the advantage that it can faithfully render the virtual reality of the hyperspace. This mechanism models the heterogeneous set of documents from the web space, which are represented under various forms and can be processed with Web Mining tools using all the existing data mining techniques. The Web Mining techniques will gradually improve along with the web processing technologies. [2].

2 Algebraic Model for Web Mining – WMHAS

The WMHAS model, that serves to the Web Mining process, can be thought of on more layers by the representation of a universal algebra $(\mathcal{D}, \Omega)$ associated with the fitting layer, where $\mathcal{D}$ is a base set, (maybe the collection of web documents, the set of objects from a HTML document, the set of tokens from a document, etc.). The Ω is the operation set that applies to the elements in $\mathcal{D}$ (generally operations on files, on objects from the DOM structure (Document Object Model), operations as processing methods for the web mining, operations from the Text Mining domain, operations for transforming data to integrated information, operations based on the Fuzzy logics, etc.).

Such a theory must be organized as an algebraic hierarchycal structure, in which each level of the hierarchy has a certain binding level with the initial application. By passing from a given level of this hierarchy to a higher level, the suitable algebraic theory is characterized by the increasing binding rank with the initial application; in other words by passing from a level of this hierarchy to a higher level, we increase the accuracy with which the obtained theory models the considered initial application and vice versa by passing from a level of the hierarchy to a lower hierarchic level. The rank of binding with the application decreases, which means that the accuracy with which the theory obtained is modelling the initial application decreases.

To fulfil our main goals, that is, adaptivity, extensibility, and scalability we decided to use an algebraic approach for the model, coupled with the support of efficient internal algorithms. The WMHAS model components are documents, logs, objects of the documents, tokens, operations, as shown below. The set of objects is modified by operations performed on the objects. Like shown in Figure 1, the web space is represented as a multilevel hierarchy where each level is an algebraic structure expressed by formally pair $\langle \mathcal{D}, \Omega_{\mathcal{D}} \rangle$. $\mathcal{D}$ is the object set from web space. $\Omega_{\mathcal{D}}$ is the operation set like tasks, methods, functions and routines which acts on the $\mathcal{D}$ object set. As we show in Figure 3, the operations can be both inter-level operation (i.e. function for transforming data to integration information) and intra-level operation which assure the interoperability between levels.

The *document* object will be denoted with *d*, and a document collection with $\mathcal{D}$. Here a document collection $\mathcal{D}$ does not constitute a site in terms of the Internet architecture, but $\mathcal{D}$ is a set of the hypermedia documents from the web space.

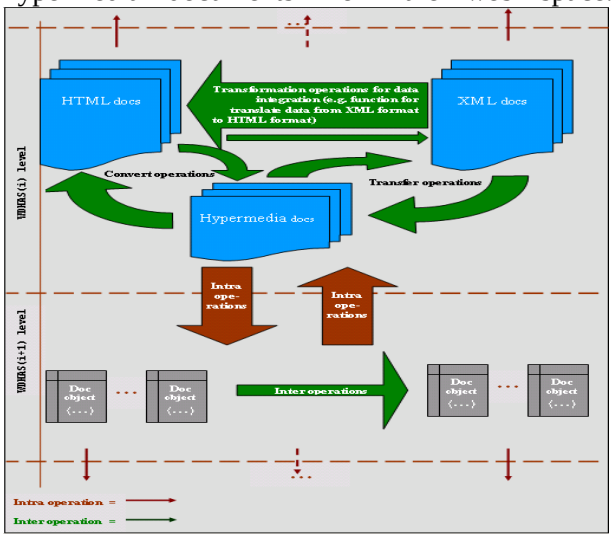

Fig. 2: The mechanism of operation deed in WMHAS model [1].

In this model hypermedia means all the documents from the Web space, documents that are created with different web techniques, which respect the accepted standards by Web service.

We will use the following convention for denotation: the algebraic structure WMHAS(*i*) as a *documents' base* from the web space in relation with the algebraic structure WMHAS(*i*+1) from the same hierarchy, which we simply denote as

$$\mathcal{B} = \langle \mathcal{D}, \Omega_{\mathcal{D}} \rangle, \quad (1)$$

where $\mathcal{D}$ is the support set of the WMHAS(i), and $\Omega_{\mathcal{D}}$, is the operation set that acts on the web documents which define the $\mathcal{B}$ structure. The

structure (1) behaves as a universal algebra. Practically, in the web environment the $\Omega_{\mathcal{D}}$ operations are those that manipulate the documents with the help of the networks' operating systems such as: creating, browsing, deleting, destroying, copying, moving, sorting, modifying, filtering, interrogation and so on. In these circumstances the structure of the high level WMHAS (i+1) from the same hierarchy is specified based on $\mathcal{B}$ under the form of a triplet:

$$\mathcal{H} = \langle \mathcal{D} = (\mathcal{D}_d)_{d\in\mathcal{D}}, \Sigma = (\Sigma_\omega)_{\omega\in\Omega\mathcal{D}}, F \rangle \qquad (2)$$

where we made following denotations: $\mathcal{D} = (\mathcal{D}_d)_{d\in\mathcal{D}}$ is a set family indexed by the $\mathcal{B}$ base support. That is, each d element ($d\in\mathcal{D}$) of the base support corresponds to a set familly of documents.

$\Sigma = (\Sigma_\omega)_{\omega\in\Omega\mathcal{D}}$ is a family of scheme operation, specified by the base operations. Each $\omega\in\Omega_D$ operation with n-arity that belongs to the base operation induces in the specified structure a set of scheme operation defined by following relation:

$$\Sigma_{\mathcal{D}} = \{\sigma = <d_1 d_2 \ldots d_n d> \mid d = \omega(d_1, d_2, \ldots d_n)\} \quad (3)$$

Finally, in the specifying structure WMHAS(i +1) from WMHAS (i) given under the form of a triplet, we used the F symbol which is considered to be the function symbol that associates to each operation scheme, denoted $\sigma\in\Sigma_{\omega,}$ an $\omega\in\Omega_{\mathcal{D}}$ heterogeneous operation specific to WMHAS(i +1). The definition domain, the values domain, the n-arity and the operation symbol, are all obviously established by the σ scheme and the way of action is typical of WMHAS (i +1) and therefore they is established by F. In other words the operation schemes are inherited from the base but the action of the operations is specific to the new defined structure, and therefore can't be inherited from the base. In this manner, if the σ operation scheme is

$\sigma = < n, s_0 s_1 \ldots s_n, d_1 d_2 \ldots d_n d>$, then $F(\sigma)$ is a specified operation in WMHAS (i +1) and in the same time behaves like the function:

$$F(\sigma): \mathcal{D}_{d1} \times \mathcal{D}_{d2} \times \ldots \times \mathcal{D}_{dn} \rightarrow \mathcal{D}_d \qquad (4)$$

So, an object from the $\mathcal{D}$ support set of the $\mathcal{B}$ base is a hypertext document which is itself made of heterogeneous elements (unstructured or semi-structured data built with the help of web technologies such as: HTML, DHTML, XHTML, XML, CSS, scripting languages, development environments etc.). As an informatics object the document from this hypermedia document class is dealt with by the operating systems from the Internet network through operations that are specific to the file operating systems, so the document at this basis level has a very high abstracting level, and the accuracy level is zero. If the document is dealt at the hierarchical level WMHAS (i+1) becomes a $(\mathcal{D}_d)_{d\in\mathcal{D}}$ family of elements over which we can act with a family of $(\Sigma_\omega)_{\omega\in\Omega}$ operations. Here we will not show how the documents were made, even though it is very important, but we can build classification criteria at the document's base level from the documents in the hypermedia documents.

The $(\mathcal{D}_d)_{d\in\mathcal{D}}$ family of elements can be interpreted as a family of elements from the same document or as a family of elements from different documents which form a class of objects on which Web Data Mining operations act.

An $F(\sigma)$ function is a specific operation that applies only to a family of elements from the hypermedia space that can give a result which expresses the measure of similarity with other elements from that class of documents, and can afterwards give a similarity rank between the documents that contain this type of elements. This is how similarity is modelled by this specification mechanism of a Web Data Mining instrument.

Therefore, a combination of many algorithms at same level can improve classification by applying the σ operation scheme ($\sigma = < n, s_0 s_1 \ldots s_n, d_1 d_2 \ldots d_n d>$) of classifiers. From the Web applications designing perspective, a σ operation scheme follow (cuprinde) many degrees of developing: support of interlated events (e.g. clickstream), structure of pages (e.g. HTML, XML), location of processing (e.g. client, server, client-server), etc.

Following kinds of operators are supported by the proposed document classification framework: extracting patterns of special interest, projecting patterns (shrinking the feature set), comparing of models (according patterns and interestingness), merging of models (combining two models) and renaming of attributes [1]-[3].

3 WMHAS and Document Classification

The Web mining process can be divided into three categories: content mining, usage mining, and structure mining. Web content mining is an automatic process that extracts patterns from on-line information, such as the HTML files, images, or E-mails, and it already goes beyond only keyword extraction or some simple statistics of words and phrases in documents. Web structure mining is a research field focused on using the analysis of the link structure of the web, and one of its purposes is to identify more preferable documents. The intuition is that a hyperlink from document d_1 to document d_2 implies that the author of document d_1 thinks document d_2 contains worthwhile information. Web servers record and accumulate data about user interactions whenever

requests for resources are received. Analyzing the web access logs of different web sites we can help understand the user's behaviour and the web structure, thereby improving the design of this colossal collection of resources.

In recent years many algorithms for different data mining tasks were developed. Overviews of different techniques were introduced in different frameworks and were proposed to support the Web mining process in a uniform manner.

The 3W model and algebra is close to our model. It uses a heterogeneous hierarchy to define a uniform framework and operators. The model consists of many levels in ths hierarchy where a single level assure possibility of combining several operation on same object class (document, token of document, etc.) [1]

The task of document classification is to assign a document to one or more categories, based on its contents. Document classification tasks can be divided into two sorts: supervised document classification where some external mechanism (such as human feedback) provides information on the correct classification for documents, and unsupervised document classification, where the classification must be done entirely without reference to external information. Also, the tasks of Web Mining are to create the models based on document content, clickstream analysis, website structure etc.

The goal of classification is to build a set of models that can correctly predict the class of the different objects. The input to these methods is a set of objects (i.e., training data), the classes which these objects belong to (i.e., dependent variables), and a set of variables describing different characteristics of the objects (i.e., independent variables). Once such a predictive model is built, it can be used to predict the class of the objects for which class information is not known a priori. The key advantage of supervised learning methods over unsupervised methods (for example, clustering) is that by having an explicit knowledge of the classes the different objects belong to, these algorithms can perform an effective feature selection if that leads to better prediction accuracy.

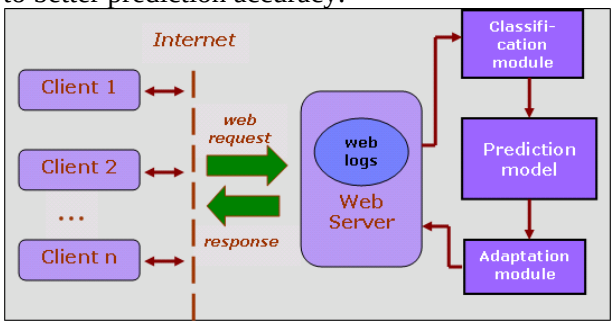

Fig. 3: General architecture of web system with web acces prediction [7].

A possible architecture of the Web Mining platform with adaptive futures is showed in Fig. 3.

In WMHAS model at level i - WMHAS(i) exist $\mathcal{D}$ a document corpus and many operation $\Omega_{\mathcal{D}}$ on the documents. By passing to the i+1 level - WMHAS(i+1), the family of the scheme operation $\Sigma = (\Sigma_{\omega})_{\omega \in \Omega_{\mathcal{D}}}$ is the family of the classification algorithms on the text documents, content documents, web sites, web logs.

Here is a brief overview (Table 1) of the basic classification algorithms for document classification as a family of the Web mining tasks.

Algorithm	Description
k-Nearest Neighbor (KNN): the KNN classifier is an instance-based learning algorithm that is based on a distance function for pairs of observations, such as the Euclidean distance or Cosine.	
KNN using average cosine	1. Select k nearest training documents, where the similarity is measured by the cosine between a given testing document and a training document. 2. Using cosine values of *k* nearest neighbors and frequency of documents of each class *i* in *k* nearest neighbors, compute average cosine value for each class *i*, Avg_Cosine(i). 3. Assign (i.e., classify) the testing document a class label which has largest average cosine.
KNN using combi nation1	This combines objective functions of both classical KNN and KNN using average cosine.
KNN using combi nation 2	1. Select *k* nearest training documents, where the similarity is measured by the cosine between a given testing document and a training document. 2. Using cosine values of *k* nearest neighbors and frequency of documents of each class in *k* nearest neighbors, compute combined objective function for each class i as follows: *Obj*(*i*) = *Pos_Avg_Cosine*(*i*) - *Neg_Avg_Cosine*(*i*), where *Pos_Avg_Cosine*(*i*) (or *Neg_Avg_Cosine*(*i*)) is the average cosine value for the set of the positive (or negative) instances among the *k* nearest neighbors. Here the positive instances mean the documents (in k nearest neighbors) whose class label is *i*. So when computing objective value for class *i*, only class label *i* is considered positive and others are considered negative. 3. Assign (i.e., classify) the testing document a class label which has the largest objective value.
Naive Bayesian (NB): The basic idea is to use the joint probabilities of words and categories to	

estimate the probabilities of categories given a document. NB algorithm computes the posterior probability that the document belongs to different classes and assigns it to the class with the highest posterior probability. There are two versions of NB algorithm. One is the multi-variate Bernoulli event model that only takes into account the presence or absence of a particular term, so it doesn't capture the number of occurrence of each word. The other model is the multinomial model that captures the word frequency information in documents.	
Variation of Naive Bayesian	The multinomial model of Naive Bayesian classification algorithm captures the word frequency information in document. So it requires the word frequency that is not weighted and normalized. However we also tried with tfn-scaled word frequency data. So, only one difference from the original multinomial model is that tfn-scaled word frequency is used instead of word frequency(i.e., txx).
Concept Vector-based (CB)	For each set of documents belonging to the same class, we compute its concept vector by summing up all vectors in the class and normalize it by its 2-norm. If there are *c* classes in the training data set, this leads to *c* concept vectors, where each concept vector for each class. The class of a new sample is determined as follow. First, for a given testing document, which was already normalized by 2-norm so that it has unit length, we compute cosine similarity between this given testing document to all *k* concept vectors. Then, based on these similarities, we assign a class label so that it corresponds to the most similar concept vector's label.
Singular Value Decomposition (SVD)-based	It use concept vectors for concept vector based algorithm. 1. For each class *i* of training documents in *k* nearest neighbors, compute first singular vector. 2. Compute cosine similarity between a given testing document and every singular vector. 3. Assign (i.e., classify) the testing document a class label which has largest cosine value.
Hierarchical Concept Vector-based (H_CB): The idea is to make good use of hierarchical structure of data set in top-down manner.	
Combination algorithms: The idea is to reduce the dimensionality of VSM and keep useful information.	
CB_KNN	1. Compute a concept vector for each category using true label information of training documents and then construct concept vector matrix C(*w*-by-*c*), where *c* is the number of categories. 2. Do projection of VSM model A(*w*-by-*d*) using concept vector matrix C(*w*-by-*c*) (i.e., $C^T * A$) 3.Apply KNN with the projected VSM model (i.e., *c*-by-*d* matrix)
Clustering + CB + K-Nearest Cluster algorithm (Cluster_CB_KNC)	1. Do clustering on each class of training data with given number of clusters. 2. Compute a concept vector for each cluster using cluster information (from step 1) and then construct concept vector matrix C(*w*-by-*k*), where *k* is the total number of clusters. 3. Apply KNN to concept vector matrix C(*w*-by-*k*) and VSM model A(*w*-by-*d* matrix). So, the testing document is classified into the class to which the major nearest neighbor cluster belongs.
Clustering + CB + KNN algorithm (Cluster_CB_KNN)	1. Do clustering on each class of training data with given number of clusters. 2. Compute a concept vector for each cluster using cluster information (from step 1) and then construct concept vector matrix C(*w*-by-*k*), where *k* is the total number of clusters. 3. Do projection of VSM model A(*w*-by-*d*) using concept vector matrix C(*w*-by-*k*) (i.e., $C^T * A$) 4. Apply KNN to the projected VSM model (*k*-by-*d* matrix)

Table 1: Important document classification algotithms and possible combinings.

4 WMHAS and Adaptive Webpages

Adaptive Web-based systems are ready to make the jump from single applications to modular distributed frameworks in which multiple applications can share user models and adaptation rules. The challenge for the future is to get research groups to work together to develop standards for exchanging information at the user model and adaptation model level, so that different systems can indeed start to share user modeling and adaptation information.

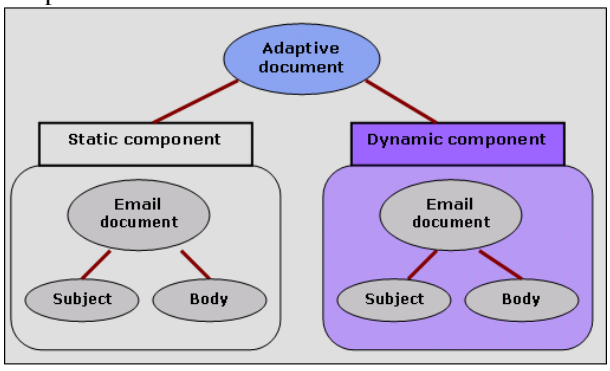

Fig. 4: Example of adaptive document structure. [8].

While new technology is being developed we also need clarity in the legal issues involved in sharing

user modeling information.
A realistic classification model for spam filtering should take into account the fact that spam evolves over time. It should also take into account the fact that each individual spam filtering instance will have its own characteristics, due to the variation in email usage, but at the same time much evidence about the nature of spam versus genuine email will be common across all (or at least most) instances.
In this light we extend our model to incorporate both a static and dynamic element. The static element represents evidence contributed by the WMHAS model trained on a large background corpus, while the dynamic element represents smaller, instance-specific evidence from the WMHAS model that are regularly retained as new data is accured. However if a greater number of fields were used, a more complex algorithm will need to be investigated [8].

5. Conclusions

Heterogeneous and semistructured data from Web repository require efficient and scalable data mining techniques. The explorative and iterative data mining process needs a strong user interaction. Ideas towards a solution of these issues were presented in this paper. First, a uniform framework was proposed, based on interoperability concepts and interestingness users. The model consists of data objects and Web mining operators. Several operators on Web objects supporting an efficient and integrated view to the Web mining process.
We have also presented an efficient, adaptive classification model for semi-structured documents that extends similar work in the semistructured and hybrid generative/discriminative classification fields.
We demonstrate that our model is efficient at combining evidence from distinct training distributions (an important attribute for adaptive classification) which suggests that the model is well suited to spam filtering, maintaining high levels of genuine recall without loss of overall accuracy.
The second contribution of the WMHAS model was the discussion of the implementation of combining different operators for Web mining process. The tight coupling of mining algorithms and Web object is an important research issue in scaling up data mining algorithms, because different kinds of patterns are necessary for implementing the operators in adaptive module for the Web sites or the Web applications.
Other issues are the development of optimization rules for the global, uniform model as well as the mapping between the algebraic operators and document classification. With many applications of adaptive Web-based systems, with small clusters of collaborating, distributed, modular systems, the idea of adaptive systems can still become a "next big thing" [6].

References:

[1]. I. Pop, E.M. Popa, *A Logical Framework for Web Data Mining Based on Heterogeneous Algebraic Structure Hierarchies*, in procedings of The 8th WSEAS International Conference (MMACTEE '06), 2006.
[2]. I. Pop, E.M. Popa, *Algebraic Modelling for Web Mining Applications and for Component Based Web Design*, WSEAS Transactions On Information Sciences And Applications, journal, Issue 1, Vol. 4, pp.97-103, January 2007.
[3]. I.Pop, I.M. Neamtu, *Applying Classification Techniques through PMML*, WSEAS Transactions On Information Sciences And Applications, Issue 1, Vol. 4, pp.103-109, January 2007.
[4]. B. Grilheres, S. Brunessaux, P. Leray, *Combining classifiers for harmful document filtering*, in procedings of Conference RIAO 2004, available at http://www.riao.org/sites/RIAO-2004/Proceedings-2004/, (retreived 8 feb. 2007).
[5]. S. Lawless, V. Wade, *Dynamic Content Discovery, Harvesting and Delivery, from Open Corpus Sources, for Adaptive Systems*, AH 2006, LNCS 4018, Springer-Verlag Berlin Heidelberg, pp. 445 – 451, 2006
[6]. P. De Bra, L. Aroyo, V. Chepegin, *The Next Big Thing: Adaptive Web-Based Systems*, Journal of Digital Information, Volume 5 Issue 1, Article No. 247, 2004.
[7]. J. Snopek, I. Jelínek, *Web Access Predictive Models*, in procedings of *CompSysTech' 2005*, available at http://ecet.ecs.ru.acad.bg/cst05/, (retreived 1 mar. 2007).
[8]. B. Medlock, *An Adaptive, SemiStructured Language Model Approach to Spam Filtering on a New Corpus*, available at www.ceas.cc/2006/, (retreived 18 mar. 2007).
[9]. K.L. Ong, W.K. Ng, E.P. Lim, *A Web Mining Platform for Enhancig Knowledge Management on the Web*, (Workshop in conj. IEEE Int. Conf. on Data Mining), November 2001.
[10]. J. Erman, M. Arlitt, A. Mahanti, *Traffic Classification Using Clustering Algorithms*, SIGCOMM'06 Workshops, Sept. 2006, Pisa, available at www.sigcomm.org/sigcomm2006/papers/minenet-01.pdf, (retrived feb. 2007).
[11]. P. Brusilovsky, M. T. Maybury, *From adaptive hypermedia to the adaptive web*, ACM Press New York, Volume 45, Issue 5 The Adaptive Web, 2002, pp. 30 – 33.