A Logical Framework for Web Data Mining Based on Heterogeneous Algebraic Structure Hierarchies

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

EMIL MARIN POPA
Department of Computer Science
Lucian Blaga University
5-7 Ioan Ratiu Str, Sibiu
ROMANIA

Abstract: The World Wide Web has become ubiquitous. Heterogeneity of the Web space consists in web pages conceptuality, content of documents and variety of the files. This heterogeneity is becoming a challenge in the information retrieval, information integration, information exploring and mining of the web. The main idea of this paper is to create a novel approach for the modelling of the Web space by an algebraic mechanism which models the heterogeneous Web warehouse. The need for computational efficiency is well-recognized by the data mining community, which sprung from the database community concern for efficient manipulation of large datasets. We present our initial development of an algebraic model named WDHAS for gathering, analyzing, and redeploying web data. Not dissimilar to conventional data mining, the general idea is that good use of web data first requires the careful selection of data (both usage and content data), the deployment of appropriate learning methods, and the evaluation of the results of applying the results of learning in a web application. Our algebraic model includes a framework which can deal with tools for building, using, and visualizing web abstractions. This paper is organized as follows. In the first section we get the arguments for algebraic modelling. Section 2 presents details of the WDHAS model. Section 3 emphases the properties of the WDHAS hierarchy base. Section 4 presents an application of the model to Hypertext Abstract Machine, showing different pattern-making like abstract objects and their operations.

Key-Words: Algebraic Modelling, Heterogeneity, Web Space, Web Mining.

1 Introduction

The enormous amount of the accumulated data into Web space represents a new challenge for the humankind which works with data sets. The Web is a huge collection of semi-structured data gathered in Web pages. The Web space is more than plain HTML and other text dominant formats and we would like to search well other data types. [2]. Most Web pages are HTML documents, with tags containing meta information about pieces of the document. Some Web pages are XML documents with standard semantic metadata and schema; XML data being associated to semantic information [3]. Others are various documents, among them we have dynamic pages, multimedia objects. There are three distinct data types in the Web: semi structured multimedia content (or Web pages), hyperlink (or just link) structure and usage data in the form of Web logs [9], [13].

If the Semantic Web becomes a reality in spite of all the social issues that need to be solved, we may have an XML-based Web. In that world, information retrieval (IR) becomes easier, and even multimedia search is simplified. Spam should disappear in this setting and it is easier to recognize good content. On the other hand, new retrieval problems appear, such as XML processing and retrieval information, and Web mining on semi structured data. [1].

The World Wide Web (or just Web) can be thought of as textual and multimedia distributed databases that reach hundreds of terabytes. This amount of data represents a new challenge in many areas of computer science, particularly to applications for distributed databases, information retrieval, component based design applications and language notation for intelligent business too. [8]

This problem is solved by an algebraic model for mining the Web which is presented below. The model which we are developing is called *Web Data Heterogeneous Algebraically Structures hierarchies* (shortly WDHAS), and we are showing an algebraic model for creating web abstractions, using them, and representing them to aid in their understanding. Also, 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. These operations also include data cleaning, session identification, merging logs from several applications and removal of robots amongst other

things.

The model is based on heterogeneous algebraically structures 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 warehouse. In the Web space, Web data set is modelled like more levels of Web warehouse, where each level represents a universal algebra of the Web objects and operations (tasks, 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.

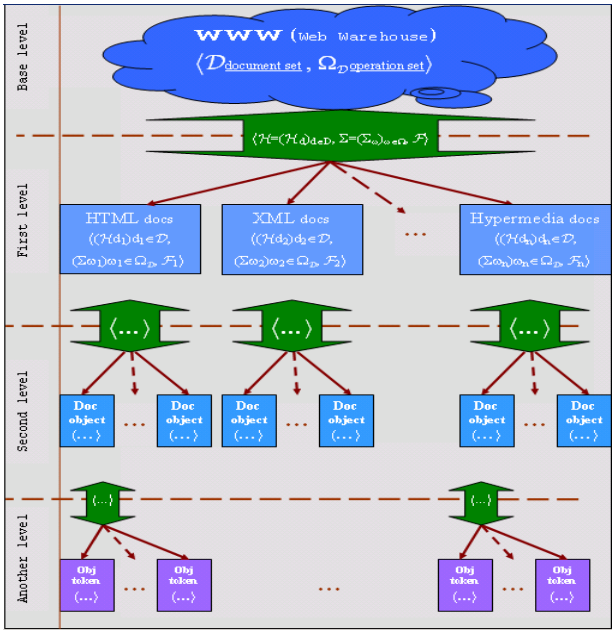

Figure 1. Simple schemata of the heterogeneous algebraically structures hierarchy for the Web space.

The Hypertext, as a virtual object system managed by the World Wide Web, gives the possibility of linking documents that are related by words and phrases. None of the models showed above is based on algebraic structures. In the algebraic modelling of the a hypermedia document system we can start from the existence of the Algebraic State Machine pattern, the Agent Algebras pattern or the multi-sorted algebraic pattern or Heterogeneous Algebraic Structures hierarchies. The growing amount of data found in the hypermedia system is heterogeneous and semi-structured. [5]. This heterogeneousness of the data needs a model that best reflects reality, and this can by made by an algebraic pattern which we will later name Web Data Mining Heterogeneous Algebraic Structure hierarchy - WDHAS . [6], [12].

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. [11]. This pattern, that serves the Web Mining process, can be thought of on more layers by the representation of a universal algebra ($\mathcal{D}$, Ω) 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 operations 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.). [10].

The mechanism is known like the HAS hierarchy (Heterogeneous Algebraic Structure hierarchy), and this was introduced by T. Rus in 1975. [12] Such a theory must be organized as an algebraic hierarchy 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 next, 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.

2 The Model

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. [4], [7]. The architecture of the Web Mining is showed in Figure 2.

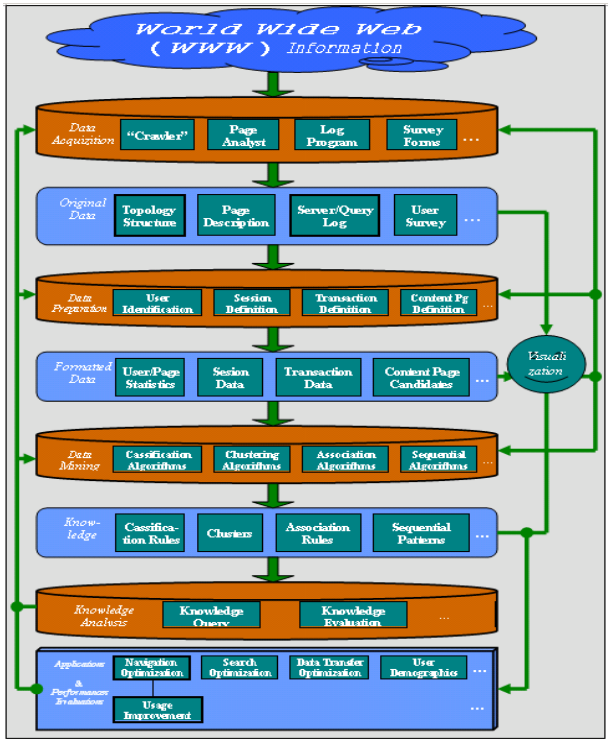

Figure 2. General framework for WM. [13].

To fulfil our main goals, that is, modularity, flexibility, extensibility, and scalability we decided to use an algebraic approach for the model, coupled with the support of efficient internal algorithms. The WDHAS 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, we are represented the web space as 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.

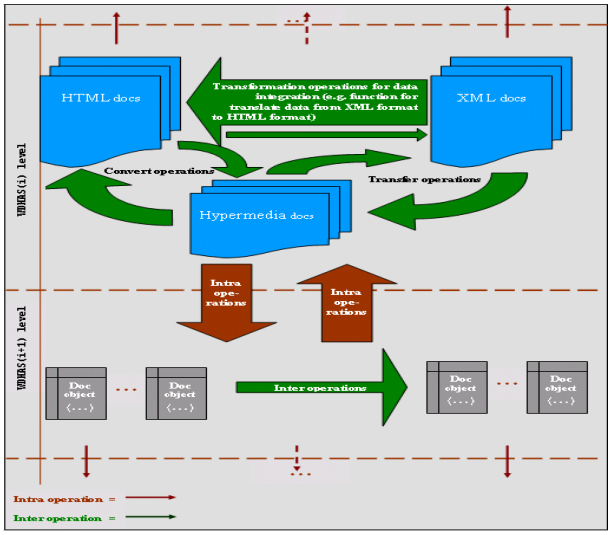

Figure 3. The mechanism of operation deed in WDHAS model.

Here hypermedia means all the documents from the WWW system, documents that are created with different web techniques, which respect the accepted standards by WWW service.

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

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

where $\mathcal{D}$ is the support set of the WDHAS(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 WDHAS (*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 an 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:

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

Finally, in the specifying structure WDHAS (*i* +1) from WDHAS (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 WDHAS (*i* +1). The definition domain, the values domain, n-arity and the operation symbol, all are obviously established by the σ scheme and the way of action is typical of WDHAS (*i* +1) and therefore it 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 [12]. 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 WDHAS (*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 WDHAS (i+1) it becomes a $(\mathcal{D}_d)_{d \in \mathcal{D}}$ family of elements over which we can act with a family of $(\Sigma_\omega)_{\omega \in \Omega \mathcal{D}}$ 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 (see the DOM - Document Object Model – standard system recommended by the W3C) 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.

3 The characteristics of a document Web base

The goal of the WDHAS hierarchy concept is to offer a formal mechanism of accurate modelling for the hypertext space concept, for Web Data Mining processes. To do this we must investigate, besides the building of a heterogeneous algebraic structure, some characteristic formal properties which offer a mathematical foundation for the modelling we are discussing. Discussion of this type of properties must be based on certain base characteristics that specify the heterogeneous algebraic structure we are talking about. This kind of properties will be at last inherited from the WDHAS hierarchy.

In the web space a *document* is considerate like an object dealt with as a file by the operating systems of the Internet system, and the client/server applications (browsers, crawlers etc) deals with a document as a semi-structured entity which contains heterogeneous data (text, graphics elements, video, audio, links etc.). This object is built from other heterogeneous objects organized with the help of different web technologies in the recognized structures by the W3C standards.

Therefore, let **(1)** be a base of documents (universal algebra). Further on we will use the habitual signs of the operations and relations from set theory, which are $\cup, \cap, \subseteq, \in$ in relation to universal algebras, with the meaning that they act on the algebras support. In other words if $\mathcal{B}_1 = \langle \mathcal{D}_1, \Omega \rangle$ and $\mathcal{B}_2 = \langle \mathcal{D}_2, \Omega \rangle$, then $\mathcal{B}_1 \cup \mathcal{B}_2 = \langle \mathcal{D}_1 \cup \mathcal{D}_2, \Omega \rangle$; $\mathcal{B}_1 \cap \mathcal{B}_2 = \langle \mathcal{D}_1 \cap \mathcal{D}_2, \Omega \rangle$; $\in$ is the relation of affiliation relative to algebraic support to that refers; and $\subseteq$ is the inclusion relation relative to the algebraic support to that refers.

Further on we will denote with $\mathcal{C}(\Omega)$ the class of all algebraic structures which have the set Ω domain of the operations, that is the class of similarity of the Ω domain. The similarity concept must be looked into more thoroughly, in this context, thus: for universal algebras the set Ω need considerate as a fixed set of those elements are the operation schemes.

This means that each $\sigma \in \Omega$ element has the following significance: $\sigma \in \Omega$ denotes a algebraic operation in $\mathcal{D}$; each $\sigma \in \Omega$ has associated a *n*-arity, meaning that once with σ one consider given the *n*-arity of the established operation from the $\boldsymbol{o}$ operation. In other words, the $\boldsymbol{\Omega}$ set needs to be looked as a function $n: \Omega \rightarrow \omega$ that determine the partitioning of $\Omega = (\Omega_n)_{n \in \omega}$, $\boldsymbol{o} \in \Omega_n$, if and only if

$n(\boldsymbol{o}) = n$. So, the $\langle \Omega, n \rangle$ pair, where n: $\Omega \to \omega$ is equivalent with the algebraic structure type used in universal algebras. That is why at any time we will use the similarity class concept we will mean the class of all algebras of a certain type given in above sense.

The law of effective action of the operations from the $\boldsymbol{\Omega}$ about the elements from $\boldsymbol{\mathcal{B}}$ is a particular question, specific of the (1) algebraic structure and the $\mathcal{C}(\Omega)$ can be looked at as a class of the similar documents that is represented by an algebraic operation. This means that the (1) pair as an algebraic structure is specified by the following three elements: $\mathcal{D}$ set; $<\Omega, n>$ pair, n: $\Omega \to \omega$, thus that $\Omega = (\Omega_n)_{n\in\omega}$; for each $\boldsymbol{o}\in\Omega_n$, the $\boldsymbol{o}$ denote a law of composition well specified that associates to each n-tuple $(d_1, d_2, \ldots, d_n)\in\mathcal{D}_n$ a $d\in\mathcal{D}$ element.

In another words, $\boldsymbol{\mathcal{B}}$ is specified in $\mathcal{C}(\boldsymbol{\Omega})$ specifying $\mathcal{D}$ and the law of composition what denote the elements of Ω about the elements of D. In a document base one can define the notions: the sub-algebra, the set of the sub-algebras that is a lattice in relation with $\cup$ and $\cap$ operations. Also we can introduce the notion of the function set of the type $f : \mathcal{D}_1 \to \mathcal{D}_2$ that are compatible with the action of the operations from Ω in the $\boldsymbol{\mathcal{B}}_1$ and $\boldsymbol{\mathcal{B}}_2$ algebras respective. The collection of all morphisms between the algebras from the $\mathcal{C}(\Omega)$ similarity classes, denoted HOM($\mathcal{C}(\Omega)$), can be specified under the family form:

$$\mathrm{HOM}(\mathcal{C}(\boldsymbol{\Omega})) = \mathrm{HOM}(\boldsymbol{\mathcal{B}}_1, \boldsymbol{\mathcal{B}}_2)_{(\mathcal{B}1,\, \mathcal{B}2)\ \in\ \mathcal{C}(\Omega)\times\mathcal{C}(\Omega)} \qquad (5)$$

In another words, if $h \in \mathrm{HOM}(\mathcal{C}(\boldsymbol{\Omega}))$, it exists $\boldsymbol{\mathcal{B}}_1$, $\boldsymbol{\mathcal{B}}_2$ from $\mathcal{C}(\Omega)$, so that the function

$h : \boldsymbol{\mathcal{B}}_1 \to \boldsymbol{\mathcal{B}}_2$ is a *morphism*.

Let $\boldsymbol{\mathcal{B}}_1 = \langle \mathcal{D}_1, \Omega \rangle$, $\mathfrak{B}_2 = \langle \mathcal{D}_2, \Omega \rangle$ and

$h : \mathcal{D}_1 \to \mathcal{D}_2$ be so that $h \in \mathrm{HOM}(\mathcal{C}(\Omega))$. Then we say that $h \in \mathrm{HOM}(\mathcal{C}(\boldsymbol{\Omega}))$ is: *monomorphism* if h is a injective function, *epimorphism* if h is a subjective function, *isomorphism* if h is a injective and subjective function, *endomorphism* if $\mathcal{D}_1$ coincide with $\mathcal{D}_2$, *automorphism* if h is in the same times endomorphism and isomorphism. [12].

4 Hypertext Abstract Machine is a Base in a WDHAS hierarchy

The **HAM** model is structured on the following four levels: interface with user; application; hypertext abstract machine (HAM); system of storage or host file system. The HAM model of storage is build from five object types: graphs; contents; nodes; links; attributes (semantic associates). [2]

The following operations can be done on the HAM objects: creation, deleting, destroying, modifying, filtering, querying and so on.

HAM like a base in the WDHAS hierarchy

The set of documents from the hyperspace, so caled *Web pages*, are storaged in extended memories of servers from Internet forms the space hypermedia so called *Web space*. This space can be thought of as a heterogeneous algebraic structures hierarchy model. In this hierarchical structure HAM machine manages and storages heterogeneous objects from hyperspace that data is organized semi-structured. The document set of the web space can be reprezented as a heterogeneous algebraic structure hierarchy. In this algebraic structure the HAM machine can be a base, that is it has a B set of the HAM objects and a $\mathcal{O}$ operation set on the HAM objects, which acts on theses objects. The objects from the B set are of the following type: graphs, contents, nodes, links and attributes. The operations of the $\mathcal{O}$ set are: creating, deleting, destroying, modifying, filtering, querying and so on. [8]

Therefore the HAM machine is specified of the base $\boldsymbol{\mathcal{B}}_{\mathrm{HAM}} = \langle B, \mathcal{O} \rangle$ that is lying at the i level in a WDHAS hierarchy. The passing to the higher level in this hierarchy, that is at the WDHAS(i+1) level means to specify the new level on the base $\boldsymbol{\mathcal{B}}_{\mathrm{HAM}}$ under the triplet form (2).

The specification of the WDHAS(i +1) structure by the WDHAS(i) given under form of the triplet, used the F symbol that is considered to be the symbol of a function which associates to each $\sigma\in\Sigma_{\circ}$, a $\circ\in\mathcal{O}_{\mathrm{H}}$ operation scheme. The scheme operation specifies the heterogeneous operations of the WDHAS(i +1). The action mode of F is specific to WDHAS(i +1) and the definition domain, the co domain, the n-arity and the symbol of operation are all evidently established by the σ scheme. In other words, the operation schemes are inherited from the base but the operation actions are specific to the new definite structure.

Therefore, all the operations from the $\boldsymbol{\mathcal{B}}_{\mathrm{HAM}}$ base or from the WDHAS(i) can be passed by this mechanism at the WDHAS(i+1) high level. This is how the complex operations from web data mining can be translated between the levels of the heterogeneous algebraic structure hierarchy. [8]

5. Conclusions and future works

1) Now the WDHAS model aims to offer a framework to develop certain special integrated

tools (operations) for mining process on the web, with the web data mining priority, but also the general frameworks for text mining processing or semantic web. At the *i+1* level of the WDHAS hierarchy the specified operation by the $F_{(\sigma)}$ function applies one of the Web Data Mining techniques, such as: associating, classification, clustering etc. This function, which specifies such operations, solves the interoperability between the operations of the WDHAS hierarchy levels. Also, the $F(\sigma)$ function can assure a framework for the Web Data Mining techniques integration and can hint the new and complex operations that act on the objects from the managed documents. [10]

2) The WDHAS hierarchy concepts can contribute substantially, by this way of modelling, at the improving of the techniques from the Text Mining and the Semantic Web domain. They can lead to generating new techniques more complex and to the interoperability between the existing techniques from the web mining domains.

3) This pattern can lead to generating new and more complex Web Mining techniques and also to the interoperability between the existing techniques in this area. Such an example is the deelopment of the PMML (Predictive Model Markup Lenguage) standard or the BPMN (Business Process Modeling Notation). [8]

The final conclusion is the ground step for the next articles which deal with the optimization of web mining application through PMML and optimizing e-commerce application through PMML. The next researches will also deal with a study concerning the efficiency of the PMML standard towards other data mining standards, such as: XML for Analysis, JSR 73, and SQL/MM.

References:

[1] *** - An algebra for XML - www-db.cs.wisc.edu/dbseminar/fall00/talks/**stratis**.ppt

[2] *** - Hypertext Theory: http://www.gwu.edu/~gelman/train/hyperbib.htm

[3] *** - W3C Semantic Activity, web: http://www.w3.org/DesignIssues/Semantic.html

[4] A. R. Pereira Jr and N. Ziviani. Retrieving similar documents from the web. *Journal of Web Engineering (JWE)*, 2(4):247–261, 2004.

[5] B. Hammond, S.Amit, K. Kochut, *Semantic Enhancement Platform for Semantic Applications over Heterogenous Content, Real World Semantic Web Applications,* IOS Press, 2002.

[6] E.M. Popa, B. A. Brumar, E. A. Stoica, Transition Systems, *in proceedings 5th RoEduNet IEEE Conference 2006*, pp 277-280.

[7] G. Psaila, "An HTTP-Based Distributed Architecture Supporting Dynamic Cooperation Processes," dexa, *16th International Workshop on Database and Expert Systems Applications (DEXA'05),* 2005, pp. 621-625.

[8] I. Pop, (2006) An Algebraic Model for Component Based Web Design, *in proceedings 5th RoEduNet IEEE Conference 2006*, pp 208-212.

[9] R. Baeza-Yates, A. R. Pereira Jr., N. Ziviani, WIM: An Information Mining Model for the Web. *16th International Workshop on Database and Expert Systems Applications (DEXA'05)*, 2005, pp. 1155-1159,.

[10] S. Delisle, Towards a Better Integration of Data Mining and Decision Support via Computational Intelligence, *16th International Workshop on Database and Expert Systems Applications (DEXA'05),* 2005, pp. 720-724,.

[11] S. S. Bhowmick, W.-K. Ng, S. K. Madria, and M. K. Mohania. Constraint-free join processing on hyperlinked web data. *In 4th International Conference on Data Warehousing and Knowledge Discovery (DaWaK'2000)*, pages 255–264. Springer-Verlag, 2002.

[12] T. Rus *Formal tools for language specification*, Academic Publishing, Bucharest, 1983.

[13] T. Zheng, Y. Niu, R. Goebel, WebFrame: In Pursuit of Computationally and Cognitively Efficient Web Mining, *Advances in Knowledge Discovery and Data Mining, In Proceedings 6th Pacific-Asia Conference*, Taipel, Taiwan, 2002, pp 260-275.