A Web Mining Logical Framework for Heterogeneous Biological Ontology Data

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

Abstract: The web documents and the web adaptation for the needs of a biologist user is today's trend of web technologies. In this paper we present an logical model for the adaptive classification methods with robust, efficient and simple to use features of the documents which include biological ontologies. The logical model, named WMHBO, is a general framework for using web data mining tasks. This model perhaps the base in any ontology construction to biological commonsense knowledge base.

Key-Words: logical model, heterogeneous data, document classification, biological ontology.

1 Introduction

A uniform view on different operators in Web Mining (WM) is an important point in supporting WM processes in the adaptive hypermedia systems and recommender systems. 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. The recommender refers to the information filtering technique that attempts to present information items (movies, music, books, news, images, web pages) that are likely of interest to the biologist user. 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 the concept mining and biological ontology to the user model is solved by an logical model called Web Mining Heterogeneous Biological Ontology hierarchies (shortly WMHBO). 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.
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 be made by an algebraic model which we will later name Web Mining Heterogeneous Biological Ontology - WMHBO.

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 Logical Framework for Heterogeneous Biological Web Documents

The WMHBO model, that serves to the Web Mining process, can be thought of on more layers by the representation of a universal algebra \((D, \Omega)\) associated with the fitting layer, where \(D\) 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 \(\Omega\) is the operation set that applies to the elements in \(D\) (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 fulfill 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 WMHBO 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 \((D, \Omega)\). \(D\) is the object set from web space. \(\Omega\) is the operation set like tasks, methods, functions and routines which acts on the \(D\) 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.

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 WMHBO($i$) as a documents’ base from the web space in relation with the algebraic structure WMHBO($i+1$) from the same hierarchy, which we simply denote as

$$\mathcal{B} = \langle \mathcal{D}, \Omega_d \rangle,$$

where $\mathcal{D}$ is the support set of the WMHBO($i$), and $\Omega_d$ 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 $\Omega_d$ 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 WMHBO($i+1$) from the same hierarchy is specified based on $\mathcal{B}$ under the form of a triplet:

$$\mathcal{H} = \langle \mathcal{D}, (\mathcal{D}_d)_{d \in \mathcal{D}}, \Sigma = (\Sigma_m)_{m \subseteq \Omega_D}, F \rangle$$

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 family of documents.

$\Sigma = (\Sigma_m)_{m \subseteq \Omega_D}$ 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:

$$\Sigma_d = \{ \sigma = <d_1 \cdots d_n> | d = \sigma(d_1, d_2, \ldots d_n) \}$$

Finally, in the specifying structure WMHBO($i + 1$) from WMHBO($i$) 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 $\sigma \in \Sigma_m$ an $\omega \in \Omega_d$ heterogeneous operation specific to WMHBO($i + 1$). The definition domain, the values domain, the $n$-arity and the operation symbol, are all obviously established by the $\sigma$ scheme and the way of action is typical of WMHBO($i + 1$) 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 $\sigma$ operation scheme is $\sigma = < n, s_0 s_1 \ldots s_n, d_1d_2\ldots d_d>$, then $F(\sigma)$ is a specified operation in WMHBO($i + 1$) and in the same time behaves like the function:

$$F(\sigma): \mathcal{D}_0 \times \mathcal{D}_2 \times \ldots \times \mathcal{D}_n \rightarrow \mathcal{D}_d$$

So, an object from the $\mathcal{D}$ support set of the 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 WMHBO($i+1$) becomes a $(\mathcal{D}_d)_{d \in \mathcal{D}}$ family of elements over which we can act with a family of $(\Sigma_m)_{m \subseteq \Omega}$ 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 specific mechanism of a Web Data Mining instrument.

Therefore, a combination of many algorithms at same level can improve classification by applying the $\sigma$ operation scheme ($\sigma = < n, s_0 s_1 \ldots s_n, d_1d_2\ldots d_d>$) of classifiers. From the Web applications designing perspective, a $\sigma$ 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 WMHBO 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 the 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 access prediction [7].

A possible architecture of the Web Mining platform with adaptive futures is showed in Fig. 3. In WMHBO model at level \( i \) - WMHBO(i) exist \( \mathcal{D} \) a document corpus and many operation \( \Omega(i) \) on the documents. By passing to the \( i+1 \) level - WMHBO(i+1), the family of the scheme operation \( \Sigma= (\Sigma_\omega)_{\omega \in \Omega} \) 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.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-Nearest Neighbor (KNN)</td>
<td>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.</td>
</tr>
<tr>
<td>KNN using average cosine</td>
<td>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</td>
</tr>
</tbody>
</table>

![Fig. 3: General architecture of web system with web access prediction](image-url)
class \(i\), \(Avg\_Cosine(i)\).
3. Assign (i.e., classify) the testing document a class label which has largest average cosine.

**KNN using combination**

This combines objective functions of both classical KNN and KNN using average cosine.

**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., \(t_{xx}\)).

**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.

<table>
<thead>
<tr>
<th>Combination algorithm</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CB_KNN</td>
<td>1. Compute a concept vector for each category using true label information of training documents and then construct concept vector matrix (C(w\text{-by}-c)), where (c) is the number of categories. 2. Do projection of VSM model (A(w\text{-by}-d)) using concept vector matrix (C(w\text{-by}-c)) (i.e., (C^t*A)) 3. Apply KNN with the projected VSM model (i.e., (c\text{-by}-d) matrix)</td>
</tr>
</tbody>
</table>

Table 1: Important document classification algorithms and possible combinings.

## 4 WMHBO 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 WMHBO model trained on a large background corpus, while the dynamic element represents smaller, instance-specific evidence from the WMHBO model that are regularly retained as new data is accrued. 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 WMHBO 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].

Acknowledgment: This work bennefits of founding from the Romanian Ministry of Education and Research, under research grant CNCSIS 33/2007.

References:
Multi-Agent System model for optimization the monitoring process within a Natura 2000 site

DANA SIMIAN, FLORIN STOICA
University Lucian Blaga of Sibiu
Faculty of Sciences
Department of Computer Science
Str. Dr. Ion Ratiu 5-7, 550012, Sibiu
ROMANIA
dana.simian@ulbsibiu.ro, florin.stoica@ulbsibiu.ro

ANGELA CURTEAN-BANADUC
University Lucian Blaga of Sibiu
Faculty of Sciences
Department of Ecology
Str. Dr. Ion Ratiu 5-7, 550012, Sibiu
ROMANIA
banaduc@yahoo.com

Abstract: The aim of this paper is to present the main problems which must be solve for assuring the conservation of species and of habitat within a Natura 200 site and to construct a model for a multi-agent system used for optimization the monitoring process.

Key–Words: Insect-behavior, Wasp agent, Multi-agent system, Conservation

1 Introduction and motivation

The relationship between man and nature has been accepted as something essential and irreplaceable. After the second world war human pressure on natural ecosystems grew at an even faster rate. The main man pressures on biodiversity in Europe comes from changes in land use, pollution, changes in management of agricultural land, waters and forests, the introduction and spread of alien species, the over exploitation of natural resources and tourism. The obvious danger of the disappearance of a significant percentage of plant and animal species has become increasingly clear in the last decade through the massive loss of natural and semi-natural habitats, tending to become too small to maintain certain species. This specific threat is becoming more and more acute under the continuous and accelerated development of infrastructure in Europe today.

At the same time, coordinated efforts were made to create a programme of nature conservation, as a result of the appreciation of its importance by the international community. These efforts were made concrete through a considerable growth in the total area covered by protected areas, but the implementation of protection measures in these areas was not constant and did not succeed in preventing the general decline.

The shrinking and fragmentation of natural and semi-natural habitats is a great problem. Linking protected areas into networks by creating green corridors and migration areas for wild fauna has become an important pan-European project. At the moment, there are very few signs that we can stop the decline of the number of species and habitats. Hopes are linked, in the first stage, to the opportunities in Central and Eastern Europe. Romania, as a direct consequence of its geographical position, and through various favorable elements of climate, relief and socio-economic development, is still the store of remarkably high diversity at the levels of ecosystem, species and genetics.

The European project Natura 2000, has appeared for preventing species and habitat decline and for establish measures for a good conservation of species and habitat. Natura 2000 is a network of protected areas. The conservation of biodiversity is no longer limited only to the protection of the endangered species and of the genetic diversity, but also, now includes the protection of habitats. The European network of sites called Natura 2000 includes the Birds Directive, which demands the establishment of Special Protection Areas (SPAs) for bird species and the Habitats Directive, which similarly requires Special Areas of Conservation (SACs) to be designated for other species, and for habitats. At once the sites were established, a monitoring of basis conservation indicators must be done. One of the aims is to reach, within the site, the favorable conservation state establishing a monitoring strategy and a suitable management plan.

There are many indices which can be monitored in order to decide the level of conservation state. Indices which express the conservation state of species are: the size of the population, the structure of the population, limits of some indicators, resident population, the degree of isolation. There are also indices which express the conservation state of habitat: conservation degree of the habitat characteristics which are important for that species, the degree of vegetation cover,