Soft Computing for XML Data Mining

K. G. Srinivasa
M S Ramaiah Institute of Technology, Bangalore, India

K. R. Venugopal
University Visvesvaraya College of Engineering, Bangalore, India

L. M. Patnaik
Indian Institute of Science, Bangalore, India

INTRODUCTION

Efficient tools and algorithms for knowledge discovery in large data sets have been devised during the recent years. These methods exploit the capability of computers to search huge amounts of data in a fast and effective manner. However, the data to be analyzed is imprecise and afflicted with uncertainty. In the case of heterogeneous data sources such as text, audio and video, the data might moreover be ambiguous and partly conflicting. Besides, patterns and relationships of interest are usually vague and approximate. Thus, in order to make the information mining process more robust or say, human-like methods for searching and learning it requires tolerance towards imprecision, uncertainty and exceptions. Thus, they have approximate reasoning capabilities and are capable of handling partial truth. Properties of the aforementioned kind are typical soft computing. *Soft computing* techniques like *Genetic Algorithms* (GA), Artificial Neural Networks, Fuzzy Logic, Rough Sets and *Support Vector Machines* (SVM) when used in combination was found to be effective. Therefore, *soft computing* algorithms are used to accomplish data mining across different applications (Mitra S, Pal S K & Mitra P, 2002; Alex A Freitas, 2002).

Extensible Markup Language (XML) is emerging as a de facto standard for information exchange among various applications of World Wide Web due to XML's inherent data self-describing capacity and flexibility of organizing data. In XML representation, the semantics are associated with the contents of the document by making use of self describing tags which can be defined by the users. Hence XML can be used as a medium for interoperability over the Internet. With these advantages, the amount of data that is being published on the Web in the form of XML is growing enormously and many naïve users find the need to search over large XML document collections (Gang Gou & Rada Chirkova, 2007; Luk R et al., 2000).

BACKGROUND

The SVM is an efficient and principled method used for classification and regression purposes. The SVM is capable of classifying linearly separable and non-linearly separable data. GA is an effective technique for searching enormous, possibly unstructured solution spaces. The human search strategy which is efficient for small documents is not viable when performing search over enormous amounts of data. Hence, making search engines cognizant of the search strategy using GA can help in fast and accurate search over large document collections.

The topic categorization of XML documents poses several new challenges. The tags in XML represent the semantics of the contents of the document and thus are more significant than the contents during the process of classification. Therefore, a general framework which assigns equal priority to both, the tags and the contents of an XML document will not be able to exhibit any significant performance improvement. Thus, a *topic categorization* framework with prominence to tags will be highly efficient. The possibility of topic categorization of XML documents using SVM is explored in (Srinivasa K G et al., 2005).

A *Selective Dissemination* of Information (SDI) system helps users to cope with the large amount of information by automatically disseminating the knowledge to the users in need of it. Therefore, the selective dissemination is the task of dispatching the documents

to the users based on their interests. Such systems maintain user profiles to judge the interests of the users and their information needs. The new documents are filtered against the user profiles, and the relevant information is delivered to the corresponding users. In XML documents, the utilization of user defined tags is of great importance to improve the effectiveness of the dissemination task. The possibility of *selective dissemination* of XML documents based on a user model using Adaptive GAs is addressed in (Srinivasa K G et al., 2007:IOS Press).

The keyword search over XML documents poses many new challenges. First, the result of a search over XML documents is not the document in its entirety, but only relevant document fragments and thus, granularity of the search terms must be refined when searching over XML document corpus. Second, the result of a keyword search over XML documents must be semantically interconnected document fragments. Finally, XML documents include large amounts of textual information and part of this is rarely searched. Building a single index for the whole document will make the index bulky and difficult to manage. The possibility of retrieval and ranking of XML fragments based on keyword queries using Adaptive GA for learning tag information is explored in (Srinivasa K G et al., 2005:ICP).

MAIN FOCUS

The application of soft computing paradigms like Genetic Algorithms and Support Vector Machines are used effectively to solve the optimization problems. The XML topic categorization is efficiently carried out using SVMs. Once the XML documents are categorized, the related and relevant information has to be disseminated to the user by a genetically learned user model in combination with SVMs. Once a large number of such XML tags exist, an efficient search over such XML repository is carried out using Adaptive GAs.

Topic Categorization Using SVM

The need for categorization of XML documents into specific user interest categories is in great demand because of huge XML repositories. A machine learning approach is applied to *topic categorization* which makes use of a multi class SVM for exploiting the semantic content of XML documents. The SVM is supplemented by a feature selection technique which is used to extract the useful features. For the application of SVM to the given XML data a feature vector must be constructed. The choice of the feature set determines the overall accuracy of the categorization task. Therefore, all the distinct tags from the training set XML documents are collected. This represents the initial tag pool. The tag pool can have rarely used tags and tags spread over all the categories apart from the more frequently used tags. Such tags can deteriorate the performance of SVM. The purpose of feature selection is to select the optimal feature subset that can achieve highest accuracy. Then, the XML document is parsed and all the tags present in the document are extracted. Only binary values are assigned as dimensions of the feature vector and supplied as the input to the multi class SVM. Later, this classifier is used for categorizing a new XML document based on the topic of relevance.

Selective Dissemination of XML Fragments

As the number of documents published in the form of XML is increasing, there is a need for selective dissemination of XML documents based on user interests. A combination of Adaptive Genetic Algorithms (Srinivasa K G et al., 2007:Elsevier) and a multi class SVM is used to learn a user model. Based on the feedback from the users, the system automatically adapts to the user's preference and interests. The user model and a similarity metric are used for selective dissemination of a continuous stream of XML documents. Using GAs to learn user profiles has two advantages. First, the tag combinations which are interesting to a user can be extracted using relevance feedback mechanism. Second, the context of the search terms given by the users can be adjudged and a profile can be constructed accordingly. From the collection of user profiles, random profiles are sampled and a decision on the category to which they belong is made.

A feature extractor and an SVM are used for the user model construction. The Feature extraction is performed using the measure of expected entropy loss to rank the features that are discriminators among the categories. An extended SVM to support multiclass classification is used to build the user model. A voting vector with a dimension for each class is also used for classifica-