

# New Implementations of Data Mining in a Plethora of Human Activities

Alberto Ochoa<sup>1,2</sup>, Julio Ponce<sup>3,4</sup>, Francisco Ornelas<sup>4</sup>,  
Rubén Jaramillo<sup>7</sup>, Ramón Zatarain<sup>5</sup>, María Barrón<sup>5</sup>,  
Claudia Gómez<sup>6</sup>, José Martínez<sup>6</sup> and Arturo Elias<sup>3</sup>

<sup>1</sup>*Juarez City University*

<sup>2</sup>*UNICAMP Instituto de Computação*

<sup>3</sup>*Aguascalientes University*

<sup>4</sup>*Cuauhtémoc University*

<sup>5</sup>*ITC*

<sup>6</sup>*ITCM*

<sup>7</sup>*CIMAT*

<sup>1,3,4,5,6,7</sup>*México*

<sup>2</sup>*Brazil*

## 1. Introduction

The fast growth of the societies along with the development and use of the technology, due to this at the moment have much information which can be analyzed in the search of relevant information to make predictions or decision making. Knowledge Discovery and Data Mining are powerful data analysis tools. The term Data mining is used to describe the non-trivial extraction of implicit, Data Mining is a discovery process in large and complex data set, refers to extracting knowledge from data bases. Data mining is a multidisciplinary field with many techniques. With this techniques you can create a mining model that describe the data that you will use (Ponce et al., 2009a).

Typical Data Mining techniques include clustering, association rule mining, classification, and regression.

We show an overview of some algorithms that used the data mining to solve problems that arisen from the human activities like: Electrical Power Design, Trash Collectors Routes, Frauds in Saving Houses, Vehicle Routing Problem.

One of the reasons why the Data Mining techniques are widely used is that there is a need to transform a large amount of data on information and knowledge useful.

Having a large amount of data and not have tools that can process a phenomenon has been described as rich in data but poverty in information (Han & Kamber, 2006). This steady growth of data, which is stored in large databases, has exceeded the ability of human beings to understand. Moreover, various problems they might present a constant stream of data, which may be more difficult to analyze the power of information.

### 1.1 Tree decisions to improve electrical power design

A decision tree (DT) is a directed acyclic graph, consisting of a node called root, which has no input arcs, and a set of nodes that have an entrance arch. Those nodes with output arcs are called internal nodes or nodes of evidence and those with no output arcs are known as leaf nodes or terminal nodes of decision (Rokach & Maimon, 2005).

The main objectives pursued by creating a DT (Safavian & Landgrebe, 1991) are:

- Correctly classify the largest number of objects in the training set (TS).
- Generalize, during construction of the tree, the TS to ensure that new objects are classified with the highest percentage of correct answers possible.
- If the dataset is dynamic, the structure of DT should be upgraded easily.

An algorithm for decision tree generation consists of two stages: the first is the induction stage of the tree and the second stage of classification. In the first stage is constructed decision tree from training set, commonly each internal node of the tree is composed of an attribute of the portion of the test and training set present in the node is divided according to the values that can take that attribute. The construction of the tree starts generating its root node, choosing a test attribute and partitioning the training set into two or more subsets, for each partition generates a new node and so on. When nodes are more objects of a class generate an internal node, when it contains objects of a class, they form a sheet which is assigned the class label. In the second stage of the algorithm, each new object is classified by the tree constructed, the tree is traversed from the root to a leaf node, from which membership is determined to some kind of object. The way forward in the tree is determined by decisions made at each internal node, according to attribute this to the test.

Pattern Recognition one of the most studied problems is the supervised classification, where it is known that a universe of objects is grouped into a given number of classes which have of each, a sample of known objects belong to it and the problem is given a new order to establish their relationships with each of those classes (Ruiz et al., 1999).

Supervised classification algorithms are designed to determine the membership of an object (described by a set of attributes) to one or more classes, based on the information contained in a previously classified set of objects (training set - TS).

Among the algorithms used for solving supervised classification are decision trees. A decision tree is a structure that consists of nodes (internal and leaves) and arches. Its internal nodes are characterized by one or more attributes of these nodes test and emerge one or more arcs. These arcs have an associated attribute value test and these values determine which path to follow in the path of the tree.

Leaf nodes contain information that determines the object belongs to a class. The main characteristics of a decision tree are: simple construction, no need to predetermine parameters for their construction, can treat multi-class problems the same way he works with two-class problems, ability to be represented by a set of rules and the easy interpretation of its structure.

#### 1.1.1 Classifications of decision trees

There are various classifications of decision trees, for example according to the number of test attributes in their internal nodes there are two types of trees:

- Single-valued: only contain a test attribute on each node. Examples of these algorithms include ID3 (Mitchell, 1997), C4.5 (Quinlan, 1993), CART (Breiman et al., 1984), FACT (Vanichsetakul & Loh, 1988), QUEST (Shis & Loh, 1988), Model Trees (Shou et al., 2005),

CTC (Perez et al., 2007), ID5R (Utgoff, 1989), ITI (Utgoff et al., 1997), UFFT (Gama & Medes, 2005), StreamTree (Jin & Agrawal, 2003), FDT (Janikowo, 1998), G-DT (Pedrycz, 2005) and Spider (Wang, et al., 2007).

- Multivalued: they have to a subset of attributes in each of its nodes. For example, PT2 (Utgoff & Brodley, 1990), LMDT (Utgoff & Brodley, 1995), GALE (Llora & Wilson, 2004) and C-DT.

According to the type of decision made by the tree, there are two types of trees:

- Fuzzy: give a degree of membership of each class of the data set, for example, C-DT, FDT, G-DT and Spider.
- Drives: assign the object belongs to only one class, so the object is or does not belong to a class, are examples of such algorithms: ID3, C4.5, CART, FACT, QUEST, Model Trees, CTC, LMDT, GALE, ID5R, ITI, UFFT, and StreamTree PT2.

The algorithms for generation of decision trees can be classified according to their ability to process dynamic data sets, i.e. sets in which lets you add new objects.

According to this there are two types of algorithms for generation of decision trees:

- Incremental: can handle dynamic data sets which are getting a partial solution as they are looking at the objects. Examples of such algorithms are: ID5R, ITI, UFFT, and StreamTree PT2.
- No Incremental: can only work on static data sets as needed for the solution to the dataset in its entirety. Examples include: ID3, C4.5, CART, FACT, QUEST, Model Trees, CTC, FDT, G-DT, Spider LMDT, GALE and C-DT.

### 1.1.2 Decision tree application

To diagnose the electric power apparatus, the decision tree method can be a highly recommended classification tool because it provides the if-then-rule in visible, and thus we may have a possibility to connect the physical phenomena to the observed signals. The most important point in constructing the diagnosing system is to make clear the relations between the faults and the corresponding signals. Such a database system can be built up in the laboratory using a model electric power apparatus, and we have made it. The next important thing is the feature extraction (Llora & Wilson, 2004).

## 2. Trash collectors routes organized by profiles

Waste. It is something that we produce as part of everyday living, but we do not normally think too much about our waste. Actually many cities generates a waste stream of great complexity, toxicity, and volume (see fig. 1). It includes municipal solid waste, industrial solid waste, hazardous waste, and other specialty wastes, such as medical, nuclear, mining, agricultural waste, construction and demolition (C&D) waste, household waste, etc. (OECD, 2008).

In the management of solid waste have the problem relates to the household waste is the individual decision-making over waste generation and disposal. When the people decide how much to consume and what to consume, they do not take into account how much waste they produce.

Therefore garbage collection is a very complex (even though in most cases do not perceive it) as not only identify routes used by vehicles for this purpose (which by itself is highly complex, to be taken into consideration many factors including the capability of vehicles, the

amount of waste that can each container, the type of waste, which is held in each container, the distance between containers, street address, etc.), but to determine what the best way to make such collection (Marquez, 2009).

Currently a major concern in the world is the way which must be stored, recycled or destroy the waste that we produce (as they have done studies that indicate that the daily waste production per person is about an extra kilogram to the produced in the manufacture of the products we use daily) which starts with the garbage collection process.

There are many algorithms and techniques being used to improve the collection process, creating different routes on the basis of the different profiles from those who generate the garbage and of the type of waste, some of these algorithms and techniques are: Ant Colony Algorithms, Hybrid Genetic Algorithms, Data Mining, among others.

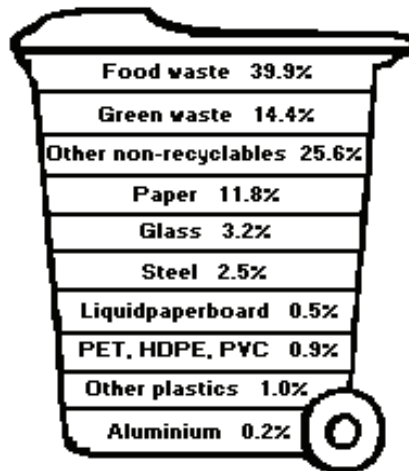


Fig. 1. Example of composition by weight of household garbage

### 3. Fraud analysis in saving houses

Fraud is an illegal activity, which has many variants and is almost as old as mankind. Fraud tries to take advantage in some way, usually economic, by the fraudster with respect to the shame. Specifically in the case of plastic card fraud there are several variants (Sánchez et al., 2009). The total cost of plastic card fraud is bigger respect to other forms of payment. The first line of defence against fraud is based on preventive measures such as the Chip and PIN cards. Next step is formed by methods employed to identify potential fraud trying to minimize potential losses. These methods are called fraud detection systems (FDS), and a variety of ways are used to detect the most behavior potential fraudulent.

#### 3.1 Techniques for detection of frauds

There are two major frameworks to detect fraud through statistical methods. If fraud is conducted in a known way, the pattern recognition techniques are typically used, especially supervised classification schemes (Whitrow et al., 2009). On the other hand if the way in which fraud is not know, for example, when there are new fraudulent behaviors, outlier analysis

methods are recommended (Kou et al., 2004). Previous research has established that the use of outlier analysis is one of the best techniques for the detection of fraud in general. Some studies show simple techniques for anomaly detection analysis to discover plastic card fraud. (Juszczak et al., 2008). However, to establish patterns to identify anomalies, these patterns are learned by the fraudsters and then they change the way to make de fraud. Other problem with this approach is not always abnormal behaviors are fraudulent, so a successful system must locate the true positive events, that is, transactions that are detected as fraud, but they really are fraud and not only appear to be fraudulent. Time is a factor against it, because to reduce losses, fraud detection should be done as quickly as possible. In practical applications it is possible to use supervised and unsupervised methods together.

### 3.1.1 Clustering

The clustering is primarily a technique of unsupervised approach, even if the semi-supervised clustering has also been studied frequently (Basu et al., 2004). Although often clustering and anomaly detection appear to be fundamentally different from one another, have developed many techniques to detect anomalies based on clustering, which can be grouped into three categories which depend on three different assumptions regarding (Chandola et al., 2009):

- a. Normal data instances belong to a pooled data set, while the anomalies do not belong to any group clustered.
- b. Normal instances of data are close to the cluster centroids, while anomalies are further away from these centroids.
- c. The normal data belongs to large, dense clusters, whereas the anomalies belong to small and sparse clusters.

Each of the above assumptions has their own forms of detect outliers which have advantages and disadvantages between them.

### 3.1.2 Hybrid systems

However, as in many aspects of artificial intelligence, the hybridization is a very current trend to detect abnormalities. The reason is because many developed algorithms do not follow entirely the concepts of a simple classical metaheuristic (Lozano et al., 2010), to solve this problem is looking for the best from a combination of metaheuristics (and any other kind of optimization methods) that perform together to complement each other and produce a profitable synergy, to which is called hybridization (Raidl, 2006).

Some possible reasons for the hybridization are (Grosan et al., 2007):

1. Improve the performance of evolutionary algorithms.
2. Improve the quality of solutions obtained by evolutionary algorithms.
3. Incorporate evolutionary algorithms as part of a larger system.

In this way, Evolutionary Algorithms (EAs) have been the most frequently technique of hybridization used for clustering. However previous research in this respect has been limited to the single objective case: criteria based on cluster compactness have been the objectives most commonly employed, as the measures provide smooth incremental guidance in all parts of search space.

Since many years ago there has been a growing interest in developing and applying of EAs in multi-objective optimization (Deb, 2001).

The recent studies on evolutionary algorithms have shown that the population-based algorithms are potential candidate to solve multi-objective optimization problems and can be efficiently used to eliminate most of the difficulties of classical single objective methods such as the sensitivity to the shape of the Pareto-optimal front and the necessity of multiple runs to find multiple Pareto-optimal solutions.

In general, the goal of a multi-objective optimization algorithm is not only to guide the search towards the Pareto-optimal front but also to maintain population diversity in the set of the Pareto optimal solutions. In this way the following three main goals need to be achieved:

- Maximize the number of elements of the Pareto optimal set found.
- Minimize the distance of the Pareto front produced by the algorithm with respect to the true (global) Pareto front (assuming we know its location).
- Maximize the spreads of solutions found, so that we can have a distribution of vectors as smooth and uniform as possible (Dehuri et al., 2009).

So it looks like a good proposal to develop a FDS with a foundation of multi-objective clustering, which places the problem of detecting fraud in an appropriate context to reality. In the same way, the system is strengthened through hybridization using PSO for the creation of clusters, and then finds the anomalies using the clustering outlier concept.

The FDS is running on the plastic card issuing institution. When a transaction arrived is sent to the FDS to be verified, the FDS receives the card details and purchase value to verify if the transaction is genuine, by calculating the anomalies, based on the expenditure profile of each cardholder, purchasing and billing locations, time of purchase, etc. When FDS confirms that the transaction is malicious, it activates an alarm and the financial institution decline the transaction. The cardholder concerned is contacted and alerted about the possibility that your card is at risk.

To find information dynamically observation for individual transactions of the cardholder, stored transactions are subject to a clustering algorithm. In general, transactions are stored in a database of the financial institution, which contain too many attributes. Although there are several factors to consider, many proposals working only with the transaction amount, with the idea of reducing the dimensionality of the problem. However, to improve the accuracy of the system is recommended to use other factors such as location and time of the transaction. So, if the purchase amount exceeds a certain value, the time between the uses of the card is low or the locations where different transactions are distant are facts to consider activating the alarm. Therefore, the alarm must be activated with a high level of accuracy.

Overall accuracy is simply the percentage of correct predictions of a classifier on a test set of "ground truth". TP means the rate of predicting "true positives" (the ratio of correctly predicted frauds over all of the true frauds), FP means the rate of predicting "false positives" (the ratio of incorrectly predicted frauds over those test examples that were not frauds, otherwise known as the "false alarm rate") (Stolfo et al., 1997).

Other two types of rates are considered for the results delivered by FDS, FN means the rate of predicting "false negatives" (the ratio of no predicted frauds over all the true frauds) and TN means the rate of predicting "true negatives" (the ratio of normal transactions detected). Table I shows the classification rate of results obtained by the FDS after analyzing a transaction.

Once clusters are established, new transaction is entered and evaluated in the FDS, to see if it belongs to a cluster set or is outside of it, seeing the transaction as an anomaly and becoming a candidate to be fraudulent. All this required the calculation of anomalies through the clustering of transaction information through a multi-objective Pareto front with the support of Particle Swarm Optimization (PSO).

Outcome	Classification
<i>Miss</i>	False Negative (FN)
<i>False Alarm</i>	False Positive (FP)
<i>Hit</i>	True Positive (TP)
<i>Normal</i>	True Negative (TN)

Table 1. Classification rate of results.

The accuracy of the FDS is represented as the fraction of total transactions (both genuine and fraudulent) that are detected as correct, which can be expressed as follows (Stolfo et al., 2000). The equation 1 shows the way to computing the precision.

$$Precision = \frac{\# \text{ of TN} + \# \text{ of TP}}{\text{Total of carry out transaction}} \quad (1)$$

Fig. 2 shows the idea of the full flow of the process proposed for the FDS. As shown in the figure, the FDS is divided into two parts, one that involves the creation of clusters and the second in the detection of anomalies.

Transactions outside of clusters are candidates to be considered fraudulent, however as mentioned above the accuracy of the system is a factor to be considered, which is expected to maximize in order to increase the functionality of the FDS.

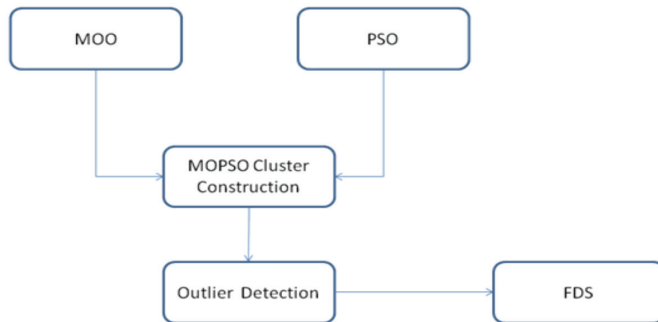


Fig. 2. Research model

#### 4. Data mining in vehicle routing problem

With the rapid development of the World-Wide Web (WWW), the increased popularity and ease of use of its tools, the World-Wide Web is becoming the most important media for collecting, sharing and distributing information. Progress in digital data acquisition and storage technology has resulted in the growth of huge distributed databases. Due that interest has grown in the possibility of tapping these data, of extracting from them information that might be of value to the owner of the database.

The discipline concerned with this task has become known as data mining, is the analysis of observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. The relationships and summaries derived through a data mining exercise are often referred to as models or

patterns. Examples include linear equations, rules, clusters, graphs, tree structures and recurrent patterns in time series.

These patterns provide knowledge on the application domain that is represented by the document collection. Such a pattern can also be seen as a query or implying a query that, when addressed to the collection, retrieves a set of documents. Thus the data mining tools also identify interesting queries which can be used to browse the collection. The system searches for interesting concept sets and relations between concept sets, using explicit bias for capturing interestingness. A set of concepts (terms, phrases or keywords) directly corresponds to a query that can be placed to the document collection for retrieving those documents that contain all the concepts of the set.

In this work, a new ant-colony algorithm, Adaptive Neighboring-Ant Search (AdaNAS), for the semantic query routing problem (SQRP) in a P2P network is presented. The proposed algorithm incorporates an adaptive control parameter tuning technique for runtime estimation of the time-to-live (TTL) of the ants. AdaNAS uses three strategies that take advantage of the local environment: learning, characterization, and exploration. Two classical learning rules are used to gain experience on past performance using three new learning functions based on the distance travelled and the resources found by the ants. These strategies are aimed to produce a greater amount of results in a lesser amount of time. The time-to-live (TTL) parameter is tuned at runtime, though a deterministic rule based on the information acquired by these three local strategies.

#### 4.1 Semantic Query Routing Problem (SQRP)

SQRP is the problem of locating information in a network based on a query formed by keywords. The goal in SQRP is to determine shorter routes from a node that issues a query to those nodes of the network that can appropriately answer the query by providing the requested information. Each query traverses the network, moving from the initiating node to a neighboring node and then to a neighbor of a neighbor and so forth, until it locates the requested resource or gives up in its absence. Due to the complexity of the problem (Amaral, 2004) (Lui et al., 2005) (Tempich et al., 2004), (Wu et al., 2006) solutions proposed to SQRP typically limit to special cases.

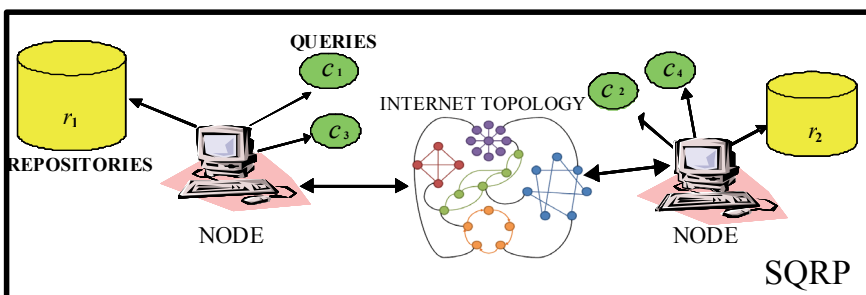


Fig. 3. SQRP Componets

The general strategies of SQRP algorithms are the following. Each node maintains a local database of documents  $r_i$  called the repository. The search mechanism is based on nodes sending messages to the neighboring nodes to query the contents of their repositories. The queries  $q_i$  are messages that contain keywords that describe for possible matches. If this

examination produces results to the query, the node responds by creating another message informing the node that launched the query of the resources available in the responding node. If there are no results or there are too few results, the node that received the query forwards it to one or more of its neighbors. This process is repeated until some predefined stopping criteria is reached. An important observation is that in a P2P network the connection pattern varies among the net (heterogeneous topology), moreover the connections may change in time, and this may alter the routes available for messages to take. As showed in the Figure 1 each node has associated a database of documents  $r_i$  (repository). Those are available to all nodes connected in the network. A node seeks information at the repository sending messages to its nodes neighbors.

#### 4.2 Neighboring-Ant Search (NAS)

NAS (Cruz et al., 2008) is also an ant-colony system, but incorporates a local structural measure to guide the ants towards nodes that have better connectivity. The algorithm has three main phases: an evaluation phase that examines the local repository and incorporates the classical lookahead technique (Mihail et al., 2004), a transition phase in which the query propagates in the network until its TTL is reached, and a retrieval phase in which the pheromone tables are updated.

Most relevant aspects of former works have been incorporated into the proposed NAS algorithm. The framework of AntNet algorithm is modified to correspond to the problem conditions: in AntNet the final addresses are known, while NAS algorithm does not have a priori knowledge of where the resources are located. On the other hand, differently to AntSearch, the SemAnt algorithm and NAS are focused on the same problem conditions, and both use algorithms based on AntNet algorithm. However, the difference between the SemAnt and NAS is that SemAnt only learns from past experience, whereas NAS takes advantage of the local environment. This means that the search in NAS takes place in terms of the classic local exploration method of Lookahead (Mihail et al., 2004), the local structural metric DDC (Ortega, 2009) its measures the differences between the degree of a node and the degree of its neighbors, and three local functions of the past algorithm performance.

#### 4.3 Adaptative Neighboring-Ant Search (AdaNAS)

AdaNAS is a metaheuristic algorithm, where a set of independent agents called ants cooperate indirectly and sporadically to achieve a common goal.

The algorithm has two objectives: it seeks to maximize the number of resources found by the ants and to minimize the number of steps taken by the ants. AdaNAS guides the queries toward nodes that have better connectivity using the local structural metric degree; in addition, it uses the well known lookahead technique, which, by means of data structures, allows to know the repository of the neighboring nodes of a specific node.

The algorithm performs in parallel all the queries using query ants. Each node has only a query ant, which generates a Forward Ant for attending only one user query, assigning the searched keyword  $t$  to the Forward Ant. Moreover the query ants realize periodically the local pheromone evaporation of the node where it is. In the Algorithm is shown the process realized by the Forward Ant. As can be observed all Forward Ants act in parallel. In an initial phase (lines 4-8), the ant checks the local repository, and if it finds matching documents then creates a backward ant. Afterwards, it realizes the search process (lines 9-25) while it has live and has not found  $R$  documents. The search process has three sections: Evaluation of results, evaluation and application of the extension of TTL and selection of next node (lines 24-28).

The first section, the evaluation of results (lines 10-15) implements the classical Lookahead technique. That is, the ant  $x$  located in a node  $r$ , checks the lookahead structure, that indicates how many matching documents are in each neighbor node of  $r$ . This function needs three parameters: the current node ( $r$ ), the keyword ( $t$ ) and the set of known nodes ( $known$ ) by the ant. The set  $known$  indicates what nodes the lookahead function should ignore, because their matching documents have already taken into account. If some resource is found, the Forward Ant creates a backward ant and updates the quantity of found matching documents.

**Algorithm: Forward ant algorithm**

```

1  in parallel for each Forward Ant  $x(r,t,R)$ 
2  initialization:  $TTL = TTL_{max}$ ,  $hops = 0$ 
3  initialization:  $path = r$ ,  $\Lambda = r$ ,  $known = r$ 
4   $Results = get\_local\_documents(r)$ 
5  if  $results > 0$  then
6    create backward ant  $y(path, results, t)$ 
7    activate  $y$ 
8  End
9  while  $TTL < 0$  and  $results < R$  do
10    $La\_results = look\_ahead(r,t,known)$ 
11   if  $la\_results > 0$  then
12     create backward ant  $y(path, la\_results, t)$ 
13     activate  $y$ 
14      $results = results + la\_results$ 
15   End
16   if  $TTL > 0$  then
17      $TTL = TTL - 1$ 
18   Else
19     if  $(results < R)$  and  $(\Delta TTL(x, results, hops) > 0)$  then
20        $TTL = TTL + \Delta TTL(x, results, hops)$ 
21       change parameters:  $q = 1$ ,  $W_{deg} = 0$ ,  $\beta_2 = 0$ 
22     End
23   End
24    $Hops = hops + 1$ 
25    $Known = known \cup [ (r \cup \Gamma(r))$ 
25    $\Lambda = \Lambda \cup r$ 
27    $r = \ell(x,r,t)$ 
28   add to  $path(r)$ 
29   End
30   create update ant  $z(x, path, t)$ 
31   activate  $z$ 
32   kill  $x$ 
33   end of in parallel

```

Fig. 4. AdaNAS algorithm

The second section (lines 16-23) is evaluation and application of the extension of TTL. In this section the ant verifies if TTL reaches zero, if it is true, the ant intends to extend its life, if it

can do it, it changes the normal transition rule modifying some parameters (line 21) in order to create the modified transition rule. The third section of the search process phase is the selection of the next node. Here, the transition rule (normal or modified) is applied for selecting the next node and some structures are updated. The final phase occurs when the search process finishes; then, the Forward Ant creates an update ant for doing the pheromone update.

Figure 5 shows the results of the different experiments applied to NAS and AdaNAS on thirty runnings for each ninety different instances generated with the characteristics showed in (Cruz et al., 2004). It can be seen from it that on all the instances the AdaNAS algorithm outperforms NAS. On average, AdaNAS had an efficiency 81% better than NAS. The topology and the repositories were created static, whereas the queries were launched randomly during the simulation. Each simulation was run for 15,000 queries during 500 time units, each unit has 100ms. The average performance was studied by computing three performance measures of each 100 queries. Average efficiency, defined as the average of resources found per traversed edge (hits/hops).

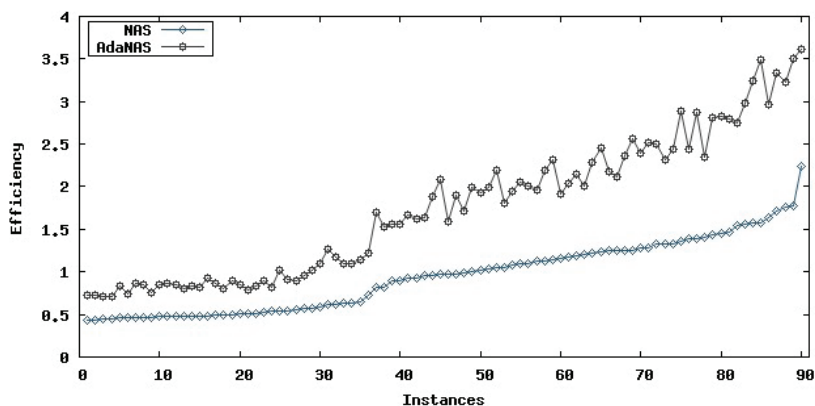


Fig. 5. Comparison between NAS and AdaNAS experimenting with 90 instances.

## 5. Text mining in the media

Today it is common to use computational tools to retrieve information, in fact it is an everyday and in many cases necessary. Information retrieval is performed on structured or unstructured data, IR systems commonly have recovered information from unstructured text (text without markup) while the database systems has been created to query relational data (sets of records that have values for predefined) , the principal differences between are in terms of retrieval model, data structures and query language. (Christopher et al., 2009).

According to the literature reviewed, nowadays do not exist techniques for Natural Language Processing to achieve 100% accurate results, either with the statistical approach, or the linguistic approach, in such a situation some researchers have blended both techniques (Chaudhuri et al. , 2006) (Gonzalez et al., 2007) (Vallez & Pedraza, 2007). For example, in (Sayyadian, 2004) they propose several methods to exploit structured information in databases and present a query expansion mechanism based on information extraction from structured data. The experimental results obtained show that using more structured information to expand the textual queries to improve performance in the recovery of entities in texts.

It is common that the amount of data with which one interacts is considerably larger and cannot be worked and in some cases it would be very difficult to work with these manually, in addition, these digital resources increase rapidly every day, reason by which the World Wide Web has become so popular, and is notorious as well as increased information systems. Because of this, it is very important to retrieve information efficiently (Hristidis & Papakonstantinou, 2002).

The search motor of Google, is the clearest example of how a computational tool can facilitate a user the information retrieval, unfortunately does not allow elaborate searches successfully, since it is designed mainly to operate with key words on documentary data bases; email servers are other type of tools very useful and popular.

Due to the diversity of existing digital media (heterogeneous data) has been investigated in diverse areas, as much in information retrieval as in natural language processing, whose final objective is to facilitate access to information and improve performance . In (Vallez & Pedraza, 2007) classified research areas as follows:

- The information extraction is the removal of a text or a set of texts entities, events and relationships between existing elements.
- The generation of summaries must like objective condense the most relevant information of a text. The techniques used vary according to compression rate, the purpose of summary, the genre of the text, the language (or languages) of the source texts, among other factors.
- The quest for answers can give a concrete answer to the question raised by the user, is important that the information needs to be well defined: dates, places, people, etc.
- The multilingual information retrieval consists of the possibility of recovering information although the question and/or the documents are in different languages, situation that reigns at the moment in the Web.

Automatic classification techniques Search text automatically assign a set of documents to predefined classification categories, mainly by using statistical techniques, processing and parameterization.

IR systems not only seek to identify only one object in a collection, but several items that can answer the query that satisfy user requirements, objects are usually text documents, but may be of multimedia content such as image, video or audio. For recovery to be efficient, the data are transformed into adequate representation, in addition, to answer satisfactorily the demands made by the user, the system can use various techniques and models, for example, the statistical processing that represents the classical model the information retrieval systems. In (Noy, 2006) use data mining to test their analytical approach, whereas in (Oren, 2002) use the genetic programming paradigm with satisfactory results.

In (Iskandar, 2007) "The retrieval strategy has been evaluated using Wikipedia, a social media collection that is an online encyclopedia. Social media describes the online tools and platforms that people use to share opinions, insights, experiences, and perspectives with each other. Social media can take many different forms, including text, images, audio, and video. Popular social mediums include blogs, message boards, podcasts, wikis, and blogs", see Figure 6.

## 5.1 Experiments

We simulated by means of the developed tool -WREID- the expectations of successfully in a circuit of Wrestling and interests of obtain popularity based on their performance associated with specific features. One of most interesting characteristics observed in the experimental

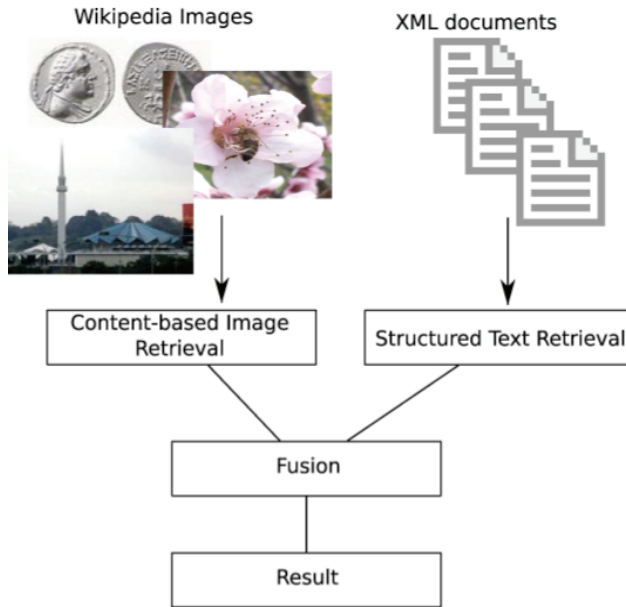


Fig. 6. Social Media Retrieval using image features and structured text

analysis were the diversity of cultural patterns established by each society because the selection of different attributes in a potential best wrestler: Agility, ability to fight, Emotional Control, Force, Stamina, Speed, Intelligence. The structured scenes associated the agents cannot be reproduced in general, so that the time and space belong to a given moment in them. They represent a unique form, needs and innovator of adaptive behavior which solves a followed computational problem of a complex change of relations. Using Social Data Mining implementing with agents was possible simulate the behavior of many followers in the selection of a best wrestler and determinate whom people support this professional career. With respect at Node attributes, we summarize the measures required to describe individual nodes of a graph. They allow identifying elements by their topological properties. The degree -or connectivity- ( $k_i$ ) of a node  $v_i$  is defined as the number of edges of this node. From the adjacency matrix, we easily obtain the degree of a given node as:

$$k_i = \sum_{j=1}^N a_{ij} \tag{2}$$

See examples of  $k$  values in figure 7. For directed graphs, we distinguish between incoming and outgoing links. Thus, we specify the degree of a node in its *indegree*,  $in_i$ , and *outdegree*,  $out_i$ . The *clustering coefficient*  $C_i$  is a local measure quantifying the likelihood that neighboring nodes of  $v_i$  are connected with each other. It is calculated by dividing the number of neighbors of  $v_i$  that are actually connected among them,  $n$ , with all possible combinations excluding autoloops, i.e.,  $ki(ki-1)$ . Formally, we have:

$$C_i = \frac{2n}{k_i(k_i - 1)} \tag{3}$$

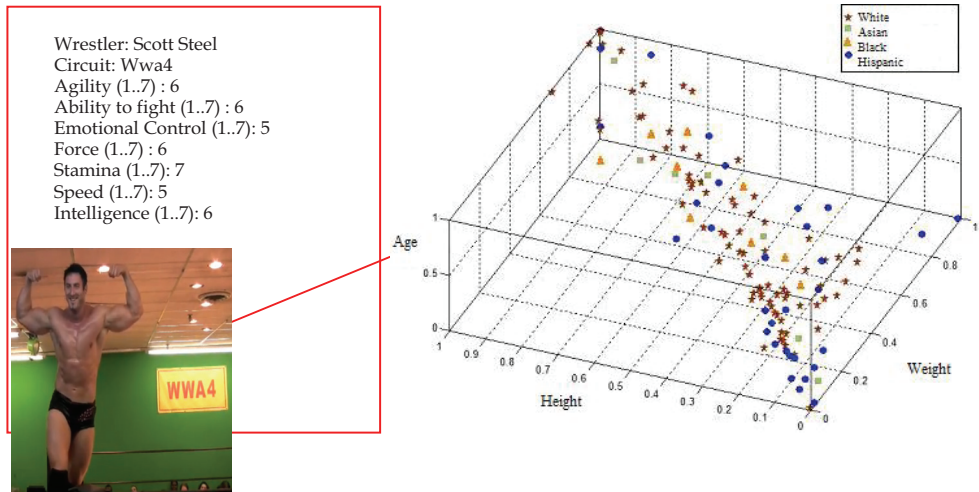


Fig. 7. Individual features of an element and classification of wrestling performance to a sample of 127 Wrestlers.

We first observe that Professional Wrestler Idol (support in features related with age, height and weight are considered) always plays a very significant role, which should of course not be surprising. Hidden patterns observed in the agents are related with size of circuit, match records and cultural distances (ethnicity), and the expectative of selection of a good wrestler whit specific attributes. The nodes with more value in their degree are considered more popular and obtain the best contracts. To get some insight, we run 100 regressions on 100 random samples of half the number of observations, and count the number of times each parameter affect the graph built. A Wrestler with the features similar to Scott Steel was selected as the most popular by the majority of societies because the attributes offered by it are adequate for others. In Figure 7 is shown the results of a sample of American Wrestlers.

## 6. Data mining with Ant Colony and Genetic Algorithm

### 6.1 Artificial Ant Colony

This section describes the principles of any Ant System (AS), a meta-heuristic algorithm based in the form in how the natural ants find food sources. The description starts with the ant metaphor, which is a model of this behavior. Then, it follows a discussion of how AS has evolved, and show as the ant algorithms can be applied to the Data Mining process. The Ant System was inspired by collective behavior of certain real ants (forager ants). While they are traveling in search of food, they deposit a chemical substance called pheromone on the traversed path. The communication through the pheromone is an effective way of coordinating the activities of these insects. For this reason, pheromone rapidly influences the behavior of each ant: they will choose the paths where is the biggest pheromone concentration. The behavior of real ants to search food is modeled as a probabilistic process. When there are paths without any amount of pheromone, the ants explores the neighboring area in a totally random way. In presence of an amount of pheromone, the ants follow a path with a probability based in the pheromone concentration. The ants deposit additional pheromone concentrations during his travels. Since the pheromone evaporates, the

pheromone concentration in non-used paths tends to disappear slowly. The Ant System (AS) or Ant Colony Optimization (ACO) was introduced by Marco Dorigo (Dorigo, 1991). The Ant System is inspired in the natural optimization process of real ants to create paths. This type of algorithms can be applied to the solution of many combinatorial optimization problems. The artificial ants, repeat the search process to find solutions. Each ant builds a possible solution to the optimization problem. The ants share information through the pheromone, which is a common memory (global information) that can be accessed by all. The Ant System is a multi-agent system, where the ant-agents have simple behavior but the interactions between them have like result a complex behavior of the whole ant colony. They need the collaboration of whole colony to get the final objective. The AS was originally proposed to solve the Traveling Salesman Problem (TSP), and the Quadratic Assignment Problem (QAP). Now exist a lot of applications like scheduling, machine learning, data mining, and others. There are several variants of AS designed to solve specific problems or to extend the characteristics of the basic algorithm (Ochoa et al., 2010). Some of the most important variants of AS in order of appearance are. Ant Colony Optimization (ACO) was introduced initially by Dorigo (Dorigo, 1991), the Ant-Q algorithm designed by Gambardela and Dorigo (Gambardela, 1995), Max-Min Ant System algorithm (MMAS) was developed by Stützle and Hoese (Stützle, 1996), other variant of AS, named ASrank, was developed by Bullnheimer, Hartl and Strauss (Bullnheimer et al., 1997).

Actually exist some AS to solve task of Data Mining, like classified and clustering, some of this algorithms are: ANT-LGP, ANT-BASED Clustering, AntClass, Ant-Miner, others.

The maximum clique problem is a problem classified within the NP-Hard problems; this problem has real applications eg: Codes Theory, Errors Diagnosis, Computer Vision, Clustering Analysis, Information Retrieval, Learning Automatic, Data Mining, among others. Therefore it is important to use new heuristic and/or meta-heuristics techniques to try to solve this problem (Ponce et al., 2009b). The general Ant Colony Algorithm for the maximum clique problem proposed by Fenet and Solnon (Fenet and Solnon, 2003). The proposed algorithm is based on the Ant Algorithm created to solve the clique maximum; the construction process is showed in figure 8.

To initialize the pheromone signs

To place Ants Randomly

**Repeat**

**For**  $k$  en  $1..nb$  Ants **do**:

Build the clique (Solution)  $C_k$

Update the pheromone signs  $\{C_1, \dots, C_{nbAnts}\}$

If is the first iteration to keep in lists all the solutions without repeating no one

Else only are added to the list the solutions that not exist in the list

**Until** Reaching the Number of Cycles or Finding the optimum solution

Fig. 8. Pseudo code of Ant Clustering Algorithm.

Construction of cliques: An initial vertex is selected randomly to put an ant, and iteratively it chooses vertices to add to clique of a set of candidates (all the vertices that are connected with all vertices of the partial clique), to see figure 9.

Choose the first vertex randomly  $v_f \in V$

$C \leftarrow \{v_f\}$

```

Candidates  $\leftarrow \{v_i / (v_i, v_i) \in E\}$ 
While Candidate  $\neq 0$  do
  Choose a vertex  $v_i \in$  Candidates with a probability  $p(v_i)$ , see Ec. (2)
   $C \leftarrow C \cup \{v_i\}$ 
  Candidates  $\leftarrow$  Candidates  $\cap \{v_j / (v_i, v_j) \in E\}$ 
End While
Return C

```

Fig. 9. Construction of Clique.

This Ant Colony Algorithm can be using to realize data clustering by the natural form that have a clique.

## 6.2 Genetic Algorithm with migration operator

Genetic Algorithms are algorithms that group techniques or methods based on natural evolution and genetics, taking as basis the "Theory of Evolution of Species" proposed by Charles Darwin and the discoveries made by Gregor Mendel in the field of genetics. (Holland, 1975) (Goldberg, 1989).

As in nature, the AG's evolving populations of individuals (possible solutions) usually of better quality solutions through operators for evaluation, selection, crossover and mutation. These have proved to be a good tool for solving optimization problems. Unfortunately one of its major limitations is that due to the loss of genetic diversity due to inbreeding between individuals within populations is that they tend to converge to local optima. For this reason we have proposed hybrid genetic algorithms somehow preventing the loss of diversity and achieve more efficient and fast tools.

Of these proposals are currently working largely with AG's side, where it seeks to improve the diversity of populations and their performance, this dividing both the computational load of each of the operators on different nodes for an intensification of themselves or by dividing the initial population in subpopulations that evolve individually until certain criteria laid down in that share some of the best individuals (Whitley et al., 1998) (Lu and Areibi, 2004) (Tzung-Pei et al., 2007).

Also have the AG's with immigration adapters that have a major population and a population parallel evolve independently and each number of generations are immigrants the best individuals of the population parallel to the main population (as shown in Figure 10), allowing the introduction of new genetic material in the major population allowing a greater diversity (Ornelas et al., 2009).

To evolve independently and through the parallel population has no influence from the main population evolves in a totally different which results in a process called speciation that is that genetic material that evolved independently in different conditions generates new species with very different characteristics that depend largely on the adaptive process.

The AG's with adaptive migration have been used to solve optimal route generation, water distribution networks and wastewater, design postcards, in data mining processes, among others.

These algorithms are currently used in data mining to make the process of cauterization and classification of information, and thanks to the way they work can process large volumes of information without extensive searches, which is of great importance because by the volume of information that is currently in the databases is impossible to use this type of research.

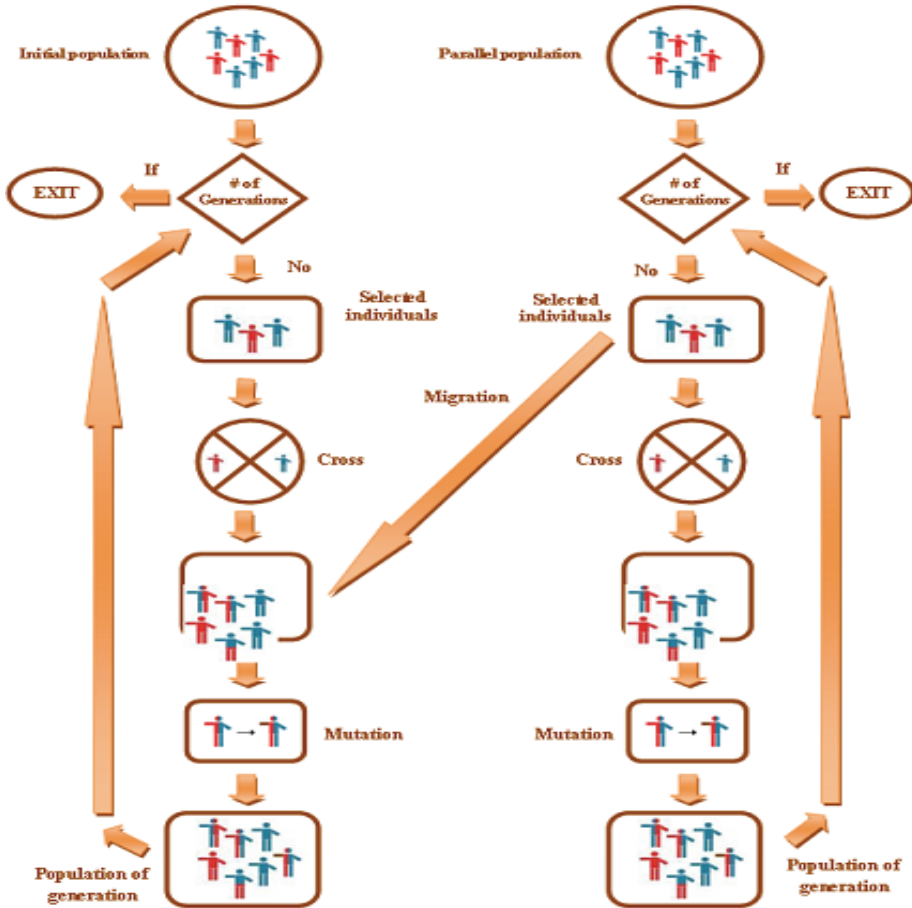


Fig. 10. Diagram of the AG's model with adaptive migration.

### 7. Intelligent Tutor Systems

Intelligent Tutoring Systems (ITS) are those computer systems that provide students with direct customization instructions or feedback without human intervention. ITSs were conceived around 1970, but not popularized until the 90's. They have four modules: the Interface Module, the Expert Module, the Student Module, and the Tutor Module. The Interface Module controls the communication between the student and the Intelligent Tutor System; the Expert Module contains a domain model that describes the knowledge or behavior that represents a high expert in the domain; the Student Module describes the student knowledge, behavior, etc.; and the Tutor Module is responsible for simulating the task of a teacher.

In this section, we present EDUCA, a Web 2.0 software tool to allow a community of authors and learners to create, share, and view learning materials and web resources for authoring Intelligent Tutoring Systems which combine collaborative, mobile and e-learning methods.

EDUCA applies different artificial intelligence techniques like a neural network and a genetic algorithm for selecting the best learning style or a recommendation-web mining system for adding and searching new learning resources.

Figure 11 illustrates the overall architecture of EDUCA. As we can observe, there are two authors: the main tutor (a teacher or instructor) and the community of learners. The student or learner is an important author of the course and participate actively adding learning resources to the courses. The learner has a user profile with information like the GPA, the particular learning style, or the recommended resources to the course. When the authors add learning material, they first create four different instances corresponding to four different learning styles according to Felder-Silverman Learner Style Model (Felder and Silverman, 1988). When a mobile course is exported to a mobile device, a XML interpreter is added to the course. A SCORM file for the course can also be exported. Once a course is created, a Course Publication Module saves it into a Course Repository. Whenever a learner accesses a course, a recommender system implemented in EDUCA presents links or Web sites with learning material related to the current topic. Such material is stored in a resource repository of EDUCA, which was searched previously by using Web mining techniques implemented also in EDUCA.

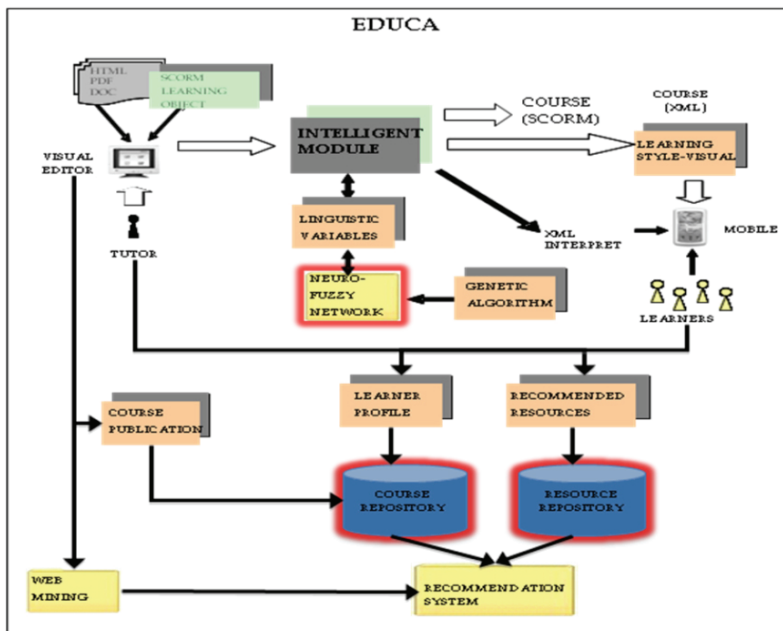


Fig. 11. EDUCA Architecture

We implemented a fuzzy-neural network using the fuzzy input values previously defined. The output of the network is the learning style for each student using a course. We also implemented a genetic algorithm (Bucket Sort) for the optimization of the weights used in the network. The network was trained for 800 generations using a population of 150 chromosomes. In order to train the network, we created three set of courses for high school students. Each course was presented in four different teaching styles according to the

Felder-Silverman model. When a mobile course is exported to a mobile device, a XML interpreter is added to the course. A SCORM file for the course can also be exported. Once a course is created, a Course Publication Module saves it into a Course Repository. Whenever a learner accesses a course, a recommender system implemented in EDUCA presents links or Web sites with learning material related to the current topic. Such material is stored in a resource repository of EDUCA, which was searched previously by using Web mining techniques implemented also in EDUCA.

We tested the tool with 15 professors/teachers and their respective students of different teaching levels. They developed different kinds of courses like a GNU/Linux course, a Basic Math Operation course, and learning material for preparation to the Mexico's Admission-Test for College EXANI-II. The students participated by reading, evaluating and adding material (Web resources) to the courses. Next, we present an example of how an author creates/updates learning material for a Basic Math course (figure 12). We first create the structure of the course (left-top). Then, we add learning material for each learning style (right-top and left-bottom). In this stage, we also assign fuzzy set values to each linguistic variable, and use recommended and actual resources for inclusion in the course. Last, we export and display the course (right-bottom).

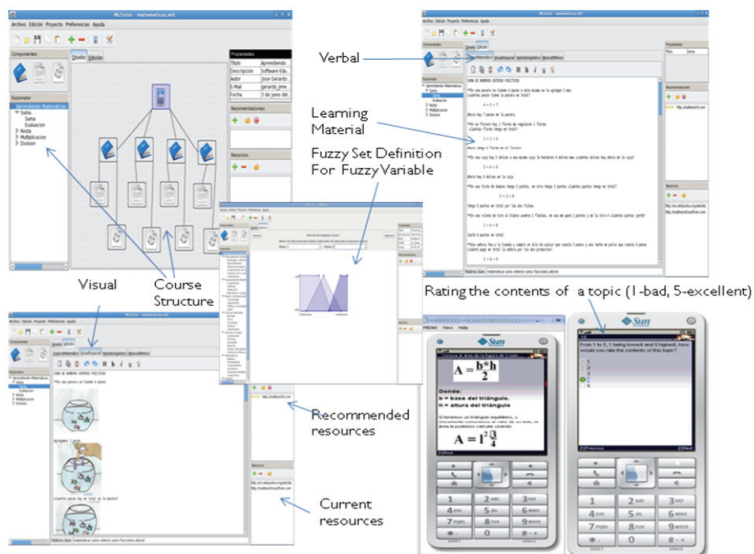


Fig. 12. Authoring Learning Material

## 8. Conclusion and the future research

Nowadays exist a lot of applications in real life problems, where is possible used data mining to analyse data base to obtain important information in different areas, in this chapter was present some algorithms and applications that us data mining such as like Electrical Power Design, Trash Collectors Routes, Fraud Analysis, Vehicle Routing Problem, Text Mining in the Media, Intelligent Tutor Systems, Ant Colony Optimization, Genetic Algorithms, Particle Swarm Optimization and Web Mining Techniques.

As shown there are multiple areas in which data mining can be used to retrieve information that is not easy to detect with the naked eye using different tools and algorithms.

We describe how decision trees work where structures are used if-then and allow the creation of recommender systems to facilitate decision making, such as diagnostic system for identifying electrical signals the device occurrence, related to physical phenomena and provide a quick and better solution to the problem presented.

For the problem of garbage collection to do a catheterization to determine how best to plan it based on the type of waste, areas collection, type and number of vehicles used for this purpose, among others, using algorithms such as Ant Colony Optimization, Genetic Algorithms and Particle Swarm Optimization.

Once clustered can use these same tools to generate optimal routes that shorten the distance travelled, fuel consumption, deterioration of vehicles, among others.

The methodologies for the detection of fraud have their own strengths and weaknesses characteristics. The overall strength of FDS using anomaly detection is the adaptability to new patterns fraudsters, in the particular case of this study is strengthened with the application of hybridization clustering processes giving a greater dynamism to the system and making it look like a promising component within the fraud detection systems with potential advantages in regard to: upgrade and management of the heterogeneity of customers and their transactions, achieving a better accuracy in the results, and greater dynamism in the system.

Additionally, the multi-objective approach place it in a better position compared to other systems, due to the characteristics of fraud detection problem where there are several factors to consider for best results.

For the solution of SQRP, we proposed a novel algorithm called AdaNAS that is based on existing ant-colony algorithms. This algorithm incorporates parameters adaptive control techniques to estimate a proper TTL value for dynamic text query routing. In addition, it incorporates local ruler that take advantage of the environment on local level, three functions were used to learn from past performance. This combination resulted in a lower hop count and an improved hit count, outperforming the NAS algorithm. Our experiments confirmed that the proposed techniques are more effective at improving search efficiency. Specifically the AdaNAS algorithm in the efficiency showed an improvement of the 81% in the performance efficiency over the NAS algorithm.

Using Social Data Mining in Media Richness we improve the understanding of change for the best paradigm substantially, because we classify the communities of agents appropriately based on their related attributes approach, this allows determine a "American Wrestler Idol" which exists with base on the determination of acceptance function by part of the remaining communities to demonstrate best performance. Each year 7000 new wrestlers arrive to different American Wrestling Circuits. Social Data Mining offers a powerful alternative for optimization problems, for that reason it provides a comprehensible panoramic of the cultural phenomenon (Ochoa et al., 2006). This technique lead us about the possibility of the experimental knowledge generation, created by the community of agents for a given application domain. How much the degree of this knowledge is cognitive for the community of agents is a topic for future work. The answer can be similar to the involved in the hard work of communication between two different societies and their respective perspectives. A new Artificial Intelligence that can be in charge of these systems, continues being distant into the horizon, in the same way that we still lack of methods to understand the original and peculiar things of each society.

As future work is to continue working with various tools and algorithms that allow us to improve data mining and this allowed us to knowledge based on information extracted from databases (information that can not be extracted directly and that features not visible to the naked eye) to improve many existing systems and create developments that take into account factors that so far can not be displayed using other tools.

Applied the models proposed in several areas for example establishing the need for FDS to be increasingly proactive in order to adapt to the greatest extent possible so changing the behaviour presented by fraudsters or in singers of Mexican Society and determine the possible "New Musical Idols or Bands" where only 27% record their second album, this for different genders according their profiles, the principal problem is the confidentiality of this information and its use for this propose.

## 9. References

- Amaral, L. and Ottino, J. (2004) Complex systems and networks: Challenges and opportunities for chemical and biological engineers. *Chemical Engineering Scientist*, 59:1653-1666.
- Basu, S.; Bilenko, M. and Mooney R. (2004). A probabilistic framework for semi-supervised clustering. *Proceedings of the Tenth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining*. Seattle, WA : ACM, press, pp. 59-68.
- Breiman, L.; Friedman, J. and Olshen, R. (1984). Classification and Regression Trees, *Wadsworth International Group*. Belmont, CA.
- Bullnheimer, B.; Hartl, R. and Strauss C. (1997). A New Rank Based Version of the Ant System: A Computational Study, *Technical report*, Institute of Management Science, University of Vienna, Austria, 1997.
- Chandola, V.; Banerjee A. and Kumar, V. (2009). Anomaly detection: A survey. *Journal of Computing Surveys*. ACM. pp. 1-58.
- Chaudhuri, S.; Das, G.; Hristidis, V. and Weikum, G. (2006). Probabilistic information retrieval approach for ranking of database query results, *ACM Trans. Database Syst.*, 31(3), pp. 1134-1168
- Cruz, L.; Gómez, C.; Aguirre, M.; Schaeffer, S.; Turrubiates, T.; Ortega, R. and Fraire, H.(2008). NAS algorithm for semantic query routing systems in complex networks. *In DCAI*, volume 50 of Advances in Soft Computing, pages 284-292. Springer.
- Deb, K. (2001). Multi-objective optimization using evolutionary algorithms, Book Chichester, Uk : John Wiley and Sons.
- Dehuri, S. and Cho, S.B. (2009). Multi-criterion Pareto based particle swarm optimized polynomial neural network for classification: A review and state-of-the-art, *Journal of Computer Science Review*. pp. 19-40.
- Dorigo, M. (1991). Positive Feedback as a Search Strategy. *Technical Report*. No. 91-016. Politecnico Di Milano, Italy.
- Felder, R. and Silverman, L. (1988). Learning and Teaching Styles In Engineering Education, *Journal of Engineering Education*. North Carolina State University and Institute for the Study of Advanced Development.. 78(7), pp. 674\_681.
- Fenet, S. and Solnon, C. (2003) Searching for Maximum Cliques with Ant Colony Optimization, *EvoWorkshops 2003*, LNCS 2611, 236-245.
- Gama, J. and Medes, P. (2005) Learning decision trees from dynamic data streams. *Journal of Universal Computer Science*.

- Gambardella, L.M. and Dorigo M.(1995). Ant-Q: A Reinforcement Learning Approach to the Traveling Salesman Problem. *Proceedings of ML-95, Twelfth International Conference on Machine Learning, Tahoe City, CA*, A. Prieditis and S. Russell (Eds.), Morgan Kaufmann, pp. 252-260.
- Goldberg, D.(1989). Genetic Algorithms in Search, Optimization, and Machine Learning. *Addison Wesley*. ISBN: 0-201-15767-5..
- González, J.J.; Pazos, R.; Gelbukh, A.; Sidorov, G.; Fraire, H. and Cruz, I. (2007). Prepositions and Conjunctions in a Natural Language Interfaces to Databases, *Lecture Notes in Computer Science*, Vol. 4743, pp. 173-182.
- Grosan, C. and Abraham, A.(2007) Hybrid evolutionary algorithms: methodologies, architectures, and reviews. Hybrid evolutionary algorithms. Book auth. Grosan C., Abraham A. and Ishibuchi H.. - Berlin : Springer Verlag-Heidelberg.
- Han, J. and Kamber, M. (2006). Data Mining, Concepts and Techniques. *Morgan Kaufmann Publishers is an imprint of Elsevier*. ISBN 13: 978-1-55860-901-3, ISBN 10: 1-55860-901-6.
- Holland J. (1975). Adaptation in Natural and Artificial Systems An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. *The University of Michigan Press*.
- Hristidis, V. and Papakonstantinou, Y. (2002). Discover: Keyword Search in Relational Databases, *VLDB '02: Proc. of the 28th International Conference on Very Large Data Bases*, Hong Kong, China, pp. 670-681.
- Iskandar, D.; Pehcevski, J.; Thom, J. and Tahaghoghi, S. (2007). Social Media Retrieval using Image Features and Structured Text, *In N. Fuhr, M. Lalmas, and A. Trotman (eds)*.
- Janikowo, C. (2008) Fuzzy decision trees: Issues and methods. *IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics*. 28-1. 1.14.
- Juszczak, P.; Adams, N.; Hand, D.; Whitrow, C. and Weston, D. (2008). Off-the-peg and bespoke classifiers for fraud detection. *Computational Statistics & Data Analysis* -Vol. 52. pp. 4521-4532.
- Jin, R. and Agrawal, G. (2003) Efficient decision tree construction on streaming data. *In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 571 576.
- Kou, Y.; Sirwongwattana, C. and Huang, S. (2004). Survey of fraud detection techniques. *IEEE International Conference on Networking, Sensing and Control*. Taipei : IEEE press, pp. 749-754. ISSN: 1810-7869. Print ISBN: 0-7803-8193-9.
- Liu, L.; XiaoLong, J. and Kwock, C. (2005). Autonomy oriented computing – from problem solving to complex system modeling. *In Springer Science + Business Media Inc*, pages 27-54.
- Llora, X. and Wilson, S. (2004). Mixed Decision Trees: Minimizing Knowledge representation bias in LCS. *Genetic and Evolutionary Computation. GECCO. Lecture Notes in Computer Science* -Vol. 3103/2204. pp. 797 809.
- Lozano, M. and García, C. (2010). Hybrid metaheuristics with evolutionary algorithms specializing in intensification and diversification: Overview and progress report. *In Journal of Computers and Operations Research*. pp. 481-497.
- Lu, G. and Areibi, S.(2004). An Island-Based GA Implementation for VLSI Standard-Cell Placement. *In GECCO 2004*. K. Deb et al. (Eds.), LNCS 3103, pp. 1138-1150, Springer-Verlag, 2004.
- Manning, C.; Raghavan, P. and Schütze, H. (2009). An Introduction to Information Retrieval, *Cambridge University Press*, Cambridge, England, pag.195.

- Márquez, M. Y. (2009). Determinación de perfiles de generación de RSD por tipología familiar a través de minería de datos: Estudios de casos en tres comunidades de Mexicali, B. C. *Tesis Doctoral*. UABC.
- Michlmayr, E. (2007). Ant Algorithms for Self-Organization in Social Networks. *PhD thesis*, Vienna University of Technology.
- Mihail, M.; Saberi, A. and Tetali, P.(2004). Random walks with lookahead in power law random graphs. *Internet Mathematics*, 3, 2004.
- Mitchell, T. (1997). *Machine Learning*. McGraw Hill.
- Noy, A.; Raban, D. and Ravid, G. (2006). Testing Social Theories in CMC through Gaming and Simulation. *Journal of Simulation and Gaming*, 37(2), pp. 174-194.
- OECD (2008) *Household Behaviour and the Environment*.
- Ochoa, A.; Hernández, A.; Cruz, L.; Ponce, J.; Montes, F.; Li, L. and Janacek, L. (2010) New Achievements in Evolutionary Computation, *Book edited by: Peter Korosec*, ISBN 978-953-307-053-7, pp. 318, INTECH, Croatia, downloaded from SCIYO.COM
- Ochoa, A.; Sehr, M.; Sarchimelia, M.; Meriam, G. et al. (2006). Italianità: Discovering a Pygmalion effect on Italian Communities Using Data Mining. *In Proceedings of CORE'2006*.
- Oren, N. (2002). Improving the effectiveness of Information Retrieval with Genetic Programming, *MSc research report*, University of the Witwatersrand, South Africa.
- Ortega, R.(2009) Estudio de las Propiedades Topológicas en Redes Complejas con Diferente Distribución del Grado y su Aplicación en la Búsqueda de Recursos Distribuidos. *PhD thesis*, Instituto Politécnico Nacional, México.
- Ornelas, F.; Padilla, A.; Padilla F.; Ponce de León E. and Ochoa, A. (2009) Genetic Algorithm using Migration and Modified GSX as Support. *Artificial Intelligence & Applications*, *Book Edited by: A. Gelbukh*, ISBN 978-607-95367-0-1, pp. 21-28, SMIA, Mexico.
- Pedrycz, W. and Sosnowski (2005). Genetically optimized fuzzy decision trees. *IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics*. pp. 633- 641.
- Perez, J.; Muguera, J.; Arbelaitz, O.; Gurrutxaga, I. and Martin, J. (2007). Combining multiple class distribution modified subsampled in a single tree. *Pattern Recognition Letters*. pp. 414-422.
- Ponce, J.; Hernández, A.; Ochoa, A.; Padilla, F.; Padilla A.; Álvarez, F. and Ponce de León, E. (2009a). *Data Mining in Web Applications*. *Data Mining and Knowledge Discovery in Real Life Applications book*. Edited by Julio Ponce and Adem Karahoca. ISBN 978-3-902613-53-0, 436 pages
- Ponce, J.; Padilla, F.; Ochoa, A.; Padilla, A.; Ponce de León, E. and Quezada, F. (2009b). Ant Colony Algorithm for Clustering through of Cliques, *Artificial Intelligence & Applications*, A. Gelbukh (Ed.), ISBN: 978-607-95367-0-1, pp. 29-34, November 2009, Mexico.
- Quinlan, J. (1993). C4.5: Programs for Machine Learning. *Morgan Kaufmann*, San Mateo, CA.
- Raidl, G. (2006). A unified view on hybrid metaheuristics. *In Proceedings of Hybrid Metaheuristics, Third International Workshop*. Berlin : Springer Verlag, pp. 1-12.
- Rokach, L. and Maimon, O.(2005) Top-down induction of decision trees Classifiers - a survey. *IEEE Transactions on Systems, Man and Cybernetics*. Reviews - Vol. 35-4. pp. 476-487. ISSN: 1094-6977.
- Ruiz, R.; Guzman, A. and Martinez J. (1999) Enfoque Lógico Combinatorio al Reconocimiento de Patrones. Instituto Politecnico Nacional, 1999.
- Safavian, S. and Landgrebe, D. (1991) A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man and Cybernetics*. pp. 660 674.)

- Sánchez, D.; Vila, M.; Cerda, L. and Serrano, J. (2009). Association rules applied to credit card fraud detection. *Expert Systems with Applications: An International Journal* – Vol. 36, pp. 3630-3640. ISSN:0957-4174.
- Sayyadian, M.; Shakeri, A.; Doan, A. and Zhai, C. (2004). Toward Entity Retrieval over Structured and Text Data, *In Proc. of ACM SIGIR 2004 Workshop on Information Retrieval and Databases*.
- Shih, Y. and Loh, W. (1997) Split Selection Methods for Classification trees. *Statistica Sinica*, pp. 815-840.
- Shou Chih, C., Hsing Kuo, P. and Yuh Jye, L.(2005) Model Trees for Classification of hybrid data types. *In Intelligent Data Engineering and Automated Learning - IDEAL: 6th International Conference*. pp. 32-39.
- Stolfo, S.; Fan, D.; Lee, W.; Prodrmidis, A. and Chan, P. (1997). Credit Card Fraud Detection Using Metalearning: *Issues and Initial Results*. *AAAI Workshop AI Methods in Fraud and Risk Management*. Columbia : AAAI Press, pp. 83-90.
- Stolfo, S.; Fan, D.; Lee, W.; Prodrmidis, A. and Chan P. (2000). Cost-Based Modeling for Fraud and Intrusion Detection: Results from the JAM Project. *DARPA Information Survivability Conference & Exposition* – Vol. 2. Hilton Head : IEEE Press, pp. 130-144. ISBN: 0-7695-0490-6.
- Stützle, T. and Hoos, H. H.(1996). Improving the Ant System: A detailed report on the MAXMIN Ant System. *Technical report AIDA-96-12*, FG Intellektik, FB Informatik, TU Darmstadt.
- Tempich, C.; Staab, S. and Wranik, A. (2004). REMINDIN': Semantic Query Routing in Peer-to-Peer Networks based on Social Metaphers, *In 13th World Wide Web Conference (WWW)*.
- Tzung-Pei, H.; Wen-Yang, L.; Shu-Min, L. and Jiann-Horng L. (2007). Dynamically Adjusting Migration Rates for Multi-Population Genetic Algorithms. *Journal of Advanced Computational and Intelligent Informatics*. Vol. 11, No. 4, pp. 410-417.
- Utgoff, P. (1989) Incremental induction of decision trees. *Machine Learning*. pp. 161-186.
- Utgoff, P.; Berkman, N. and Clouse, J. (1997). Decision tree induction based on efficient tree Restructuring. *Machine Learning*. 5-44.
- Utgoff, P. and Brodley, C. (1990). An Incremental Method for Finding multivariate splits for decision trees. *In Proc.7th International Conference on Machine Learning*. pp. 58-65.
- Utgoff, P. and Brodley, C. (1995). Multivariate decision trees. *Machine Learning*. pp. 45-77.
- Vallez, M. and Pedraza-Jimenez, R. (2007). Natural Language Processing in Textual Information Retrieval and Related Topics, [on line]. "Hipertext.net", num. 5, 2007. <<http://www.hipertext.net>> [Consulted: 07/15/10]. ISSN 1695-5498
- Vanichsetakul, N. and Wei-Yin, L. (1988). Tree-Structured Classification via Generalized Discriminant analysis. *Journal of the American Statistical Association*, pp. 715-728.
- Wang, X.; Nauck, D. and Spott, M.(2007). Intelligent data analysis with fuzzy decision trees. *Soft Computing - A Fusion of Foundations, Methodologies and Applications*. pp. 439-457.
- Whitley, D.; Rana, S. and Heckendorn (1998). The Island Model Genetic Algorithm: On Separability, Population Size and Convergence. *Journal of Computing and Information Technology*, Vol. 7, pp. 33-47, Colorado State University.
- Whitrow, C.; Hand, D.; Juszczak, P.; Weston, D. and Adams, N. (2009). Transaction aggregation as a strategy for credit card fraud detection. *Journal of Data Mining and Knowledge Discovery*. pp. 30-55.
- Wu, C.-J.; Yang, K.-H. and Ho.(2006). AntSearch: An ant search algorithm in unstructured peer-to-peer networks. *In ISCC*, pages 429-434.