

Social Data Mining to Improve Bioinspired Intelligent Systems

Alberto Ochoa^{1, ٥}, Arturo Hernández², Saúl González¹, Arnulfo Castro¹,
Alexander Gelbukh³, Alberto Hernández⁴ and Halina Iztebegovič⁵

¹Juarez City University,

^٥CIATEC,

²CIMAT,

³CIC-IPN,

⁴CIICAp-UAEM,

⁵Montenegro University

^{1, 2, 3, 4} ٥ México

⁵Montenegro

1. Introduction

The proposal of this chapter is to explain the implementation of social data mining to improve results in bioinspired intelligent systems using generation of clusters, associative rules; decision trees, associated models, dioramas and multivariable analysis for obtain knowledge about any issue related with a topic. This kind of intelligent systems using bioinspired computing – specially, group intelligence techniques such as: Ant Colony, Particle Swarm Optimization and Cultural Algorithms- that try to simulate biological processes that occur in the nature. Intelligent agents use this information to make decisions to improve a needed heuristic optimization in different fields such as: negotiation, argumentation or artificial societies simulation.

First in section 2 of this chapter, we approach different concepts related with social data mining and how to use different ways to analyze and model the necessary information to support the correct decision of agents; in next three sections we explain the way to generate a specific behaviour by using group intelligence techniques –ant colony (section 3), particle swarm optimization (section 4) and cultural algorithms (section 5), In section 6, we apply this knowledge in diverse fields and application domains that use a heuristic optimization. In section 7 we compare different cases of studies: Eurovision Voting problem, and the Distribution of Elements. Finally in section 8 we provide our conclusions and outline our future research.

2. Social data mining – basic notions

Social data mining systems enable people to share opinions and obtain a benefit from each other's experience. These systems do this by mining and redistributing information from computational records of social activity such as Usenet messages, system usage history,

Source: Data Mining in Medical and Biological Research, Book edited by: Eugenia G. Giannopoulou,
ISBN 978-953-7619-30-5, pp. 320, December 2008, I-Tech, Vienna, Austria

citations, and hyperlinks among others. Two general questions for evaluating such systems are: (1) is the extracted information valuable? , and (2) do interfaces based on extracted information improve user tasks performance?

We report here on social data mining applications, systems that mine information from the structure and content of web pages and provide an exploratory information workspace interface. We carried out experiments that yielded positive answers to both evaluation questions. First, a number of automatically computable features about web sites do a good job of predicting expert quality judgments about sites. Second, compared to popular web search interfaces, the Topic Shop interface to this information lets users select significantly higher quality sites, in less time and with less effort, and to organize the sites they select into personally meaningful collections quickly and easily. We conclude by discussing how our results may be applied and considering how they touch on general issues concerning quality, expertise, and consensus

The motivation for the social data mining approach goes back at least to Vannevar Bush's *As We May Think* essay. Bush envisioned scholars blazing trails through repositories of information and realized that these trails subsequently could be followed by others. Everyone could walk in the footsteps of the masters. In our work, we have formulated a similar intuition using the metaphor of a path through the woods. However, this metaphor highlights the role of collective effort, rather than the individual. A path results from the decisions of many individuals, united only by where they choose to walk, yet still reflects a rough notion of what the walkers find to be a good path. The path both reflects history of use and serves as a resource for future users.

Social data mining approaches seek analogous situations in the computational world. Researchers look for situations where groups of people are producing computational records (such as documents, Usenet messages, or web sites and links) as part of their normal activity. Potentially useful information implicit in these records is identified, computational techniques to harvest and aggregate the information are invented, and visualization techniques to present the results are designed. Thus, computation discovers and makes explicit the "paths through the woods" created by particular user communities. And, unlike ratings-based *collaborative filtering* systems (Resnick et al., 1994), social data mining systems do not require users to engage in any new activity; rather, they seek to exploit user preference information implicit in records of existing activity. The "history-enriched digital objects" line of work (Hill et al., 1992) was a seminal effort in this approach. It began from the observation that objects in the real world accumulate *wear* over the history of their use, and that this wear – such as the path through the woods or the dog-eared pages in a paperback book or the smudges on certain recipes in a cookbook – informs future usage. *Edit Wear* and *Read Wear* were terms used to describe computational analogues of these phenomena. Statistics such as time spent reading various parts of a document, counts of spreadsheet cell recalculations, and menu selections were captured. These statistics were then used to modify the appearance of documents and other interface objects in accordance with prior use. For example, scrollbars were annotated with horizontal lines of differing length and color to represent amount of editing (or reading) by various users.

Other work has focused on extracting information from online conversations such as Usenet. PHOAKS (Hill & Terveen, 1996) mines messages in Usenet newsgroups looking for mentions of web pages. It categorizes and aggregates mentions to create lists of popular web pages for each group. In (Viegas et al, 1990) have harvested information from Usenet newsgroups and chats and have used them to create visualizations of the conversation.

These visualizations can be used to find conversations with desirable properties, such as equality of participation or many regular participants. (Smith & Fiore, 2001) also extracted information from newsgroups and designed visualizations of the conversational thread structure, contributions by individual posters, and the relationships between posters.

Still other work has focused on extracting information from web usage logs. Footprints (Wexelblat et al., 1999) records user browsing history, analyzes it to find commonly traversed links between web pages, and constructs several different visualizations of this data to aid user navigation through a web site. Pursuing the metaphor of navigation, some researchers have used the term *social navigation* to characterize work of this nature (Munro et al., 1999). Finally, a distinct technical approach was taken by (Chalmers et al., 1998). They used the activity *path* – e.g., a sequence of URLs visited during a browsing session – as the basic unit. They have developed techniques to compute similarities between paths and to make recommendations on this basis – for example, to recommend pages to you that others browsed in close proximity to pages you browsed.

Mining the Web

Most relevant to the concerns of this paper is the work that mines the structure of the World Wide Web itself. The Web, with its rich content, link structure, and usage logs, has been a major domain for social data mining research. A basic intuition is that a link from one web site to another may indicate both similarity of content between the sites and an endorsement of the linked-to site. An intellectual antecedent for this work is the field of bibliometrics, which studies patterns of co-citation in texts (Egghe et al., 1990). Various clustering and rating algorithms have been designed to extract information from link structure.

(Ochoa et al, 2007) was developed a categorization algorithm that used hyperlink structure (as well as text similarity and user's data access) to categorize web pages into various functional roles and using Cultural Algorithms. Before (Pitkow & Pirolli, 1997) experimented with clustering algorithms based on co-citation analysis, in which pairs of documents were clustered based on the number of times they were both cited by a third document Kleinberg formalized the notion of document quality within a hyper-linked collection using the concept of *authority* (Kleinberg, 1998). At first pass, an authoritative document is one that many other documents link to. However, this notion can be strengthened by observing that links from all documents aren't equally valuable – some documents are better *hubs* for a given topic. Hubs and authorities stand in a mutually reinforcing relationship: a good authority is a document that is linked to by many good hubs, and a good hub is a document that links to many authorities. Kleinberg developed an iterative algorithm for computing authorities and hubs. He presented examples that suggested the algorithm could help to filter out irrelevant or poor quality documents (i.e., they would have low authority scores) and identify high-quality documents (they would have high authority scores). He also showed that his algorithm could be used to cluster pages within a collection, in effect disambiguating the query that generated the collection. For example, a query on "Jaguar" returned items concerning the animal, the car, and the NFL team, but Kleinberg's algorithm splits the pages into three sets, corresponding to the three meanings.

Several researchers have extended this basic algorithm. Weight links based on the similarity of the text that surrounded the hyperlink in the source document to the query that defined the topic made several important extensions. First, they weighted documents based on their

similarity to the query topic. Second, they count only links between documents from different *hosts*, and average the contribution of links from any given host to a specific document. That is, if there are k links from documents on one host to a document D on another host, then each of the links is assigned a weight of $1/k$ when the authority score of D is computed. In experiments, they showed that their extensions led to significant improvements over the basic authority algorithm.

PageRank is another link-based algorithm for ranking documents. Like Kleinberg's algorithm, this is an iterative algorithm that computes a document's score based on the scores of documents that link to it. PageRank puts more emphasis on the quality of the links to a particular document. Documents linked to by other documents with high PageRank scores will themselves receive a higher PageRank score than documents linked to by low scoring documents.

In summary, much recent research has experimented with algorithms for extracting information from web structure. A major motivation for these algorithms is that they can be used to compute measures of document quality. Yet this work has proceeded without much experimental evaluation, leaving two basic questions unanswered: first, what benefits do the more complicated link-based algorithms provide beyond simple link counts? And second, how well do the various link-based metrics (in-links, authority scores, PageRank scores) actually correlate with human quality judgments? We will report on an experiment that investigates these issues.

Information Workspaces

Once information has been extracted, it must be presented in a user interface. Users must be able to evaluate collections of items, select items they find useful, and organize them into personally meaningful collections. (Card et al, 1991) introduced the concept of *information workspaces* to refer to environments in which information items can be stored and manipulated. A departure point for most such systems is the file manager popularized by the Apple Macintosh and then in Microsoft Windows. Such systems typically include a list view, which shows various properties of items, and an icon view, which lets users organize icons representing the items in a 2D space. (Mander et al., 1992) enhanced the basic metaphor with the addition of "piles". Users could create and manipulate piles of items. Interesting interaction techniques for displaying, browsing, and searching piles were designed and tested. Bookmarks are the most popular way to create personal information workspaces of web resources. Bookmarks consist of lists of URLs; typically the title of the web page is used as the label for the URL. Users may organize their bookmarks into a hierarchical category structure. (Abrams et al., 1998) carried out an extensive study of how several hundred web users used bookmarks. They observed a number of strategies for organizing bookmarks, including a flat ordered list, a single level of folders, and hierarchical folders. They also made four design recommendations to help users manage their bookmarks more effectively. First, bookmarks must be easy to organize, e.g., via automatic sorting techniques. Second, visualization techniques are necessary to provide comprehensive overviews of large sets of bookmarks. Third, rich representations of sites are required; many users noted that site titles are not accurate descriptors of site content.

Finally, tools for managing bookmarks must be well integrated with web browsers. Many researchers have created experimental information workspace interfaces, often designed expressly for web documents. (Card et al., 1996) describe the Web Book, which uses a book

metaphor to group a collection of related web pages for viewing and interaction, and the Web Forager, an interface that lets users view and manage multiple Web Books. In addition to these novel interfaces, they also presented a set of automatic methods for generating collections (Web Books) of related pages, such as recursively following all relative links from a specified web page, following all (absolute) links from a page one level, extracting "book-like" structures by following "next" and "previous" links, and grouping pages returned from a search query. (Mackinlay et al., 1995) developed a novel user interface for accessing articles from a citation database.

The central UI object is a "Butterfly", which represents an article, its references, and its citers. The interface makes it easy for users to browse among related articles, group articles, and generate queries to retrieve articles that stand in a particular relationship to the current article. The Data Mountain of (Robertson et al., 1998) represents documents as thumbnail images in a 3D virtual space. Users can move and group the images freely, with various interesting visual and audio cues used to help users arrange the documents.

In a study comparing the use of Data Mountain to Internet Explorer Favorites, Data Mountain users retrieved items more quickly, with less incorrect or failed retrieval. Other researchers have created interfaces to support users in constructing, evolving, and managing collections of information resources. SenseMaker (Baldonado et al, 1997) focuses on supporting users in the contextual evolution of their interest in a topic. It attempts to make it easy to evolve a collection, e.g., expanding it by query-by-example operations or limiting it by applying a filter. Scatter/Gather (Pirolli et al, 1996) supports the browsing of large collections of text, allowing users to iteratively reveal topic structure and locate desirable documents.

VIKI system (Marshall et al., 1994) lets user organize collections of items by arranging them in 2D space. Hierarchical collections are supported. Later extensions (Shipman et al., 1999) added automatic visual layouts, specifically non-linear layouts such as fisheye views. (Hightower et al., 1998) addressed the observation that users often return to previously visited pages. They used Pad++ (Bederson et al., 1996) to implement PadPrints, browser companion software that presents a zoomable interface to a user's browsing history there are a number of important issues that deserve further investigation. One direction is to seek new sources for mining information about user preferences. As we have discussed, researchers have investigated hyperlink structure, electronic conversations, navigation histories and other usage logs, and purchasing history. One area with great potential is electronic media usage, in particular, listening to digital music. By observing what music someone is listening to, a system can infer the songs, artists, and genres that person prefers, and use this information to recommend additional songs and artists, and to put the person in touch with other people with similar interests. We took a step in this direction with a system that lets users visualize individual and group listening histories and define new play lists relative to listening history (Terveen et al., 2002). (Crossen et al., 2002) reported on a system that learns user preferences from the music they listen to, then selects songs to play in a shared physical environment, based in part on the preferences of all people present.

As user preferences are extracted from more and more sources, the issue of combining different types of preferences becomes important. For example, PHOAKS extracted preferences about web pages from Usenet messages and presented them to users. As users browsed through this information, PHOAKS tracked which pages users clicked on (another type of implicit preference), and users also could rate web pages (explicit preferences).

Developing general techniques for combining different types of preferences is a challenge. (Billsus et al. 1998) was presented a method for weighting different types of contributions; however, whether this is the best combination method and how to determine appropriate weights are still open issues. It is worth pointing out that the task that TopicShop supports – selecting a subset of items from a large set and then organizing the subset arises – is quite general and occurs in other contexts. For example, of the many people I exchange email with, a small subset are “contacts” whom I wish to keep track of, and organize into groups which I can use to manage my communication. We have applied this intuition in a project with Steve Whittaker, developing a new interface for the ContactMap contact management system (Whittaker et al., 2002). Features about potential contacts including their organization and frequency and regency of communication are extracted from email archives and presented in a table; as in TopicShop, the table can be sorted by any of the columns. And, when users find important contacts, they add them to their “map” (equivalent to the TopicShop Work Area) by dragging and dropping. Contacts on the map are organized by spatial arrangement and color coding. This experience illustrates that the general interaction paradigm of TopicShop can be applied in an altogether different domain. While our experiment compared TopicShop to the state-of-the-art Web directory Yahoo, it may have occurred to the reader that our techniques are suitable for integration with such a system. This is absolutely correct. Both directory systems, which contain categories of web sites typically built by a person, and search engines, which retrieve documents based on their similarity to a query, could benefit. An effective way to apply the results of our research would be to enhance (say) Yahoo by (1) using a WebCrawler/analyzer to augment each manually constructed collection of pages with the sorts of profiles our experiments showed effective, and (2) providing a TopicShop-style information workspace interface. Such a system would combine the advantages of people – applying judgment to select the initial set of collections – and computers – applying analysis techniques to provide enhanced information and to keep the collections up to date. A similar tactic could be taken by a search engine; this would be most efficient for one such as Google that already maintains a database of links between web pages. Finally, note that this argument shows that even a very large, manually constructed set of “seed” pages can be enhanced significantly by providing additional features, grouping pages into sites, and offering a good user interface

3. Behaviour in group intelligence techniques – ant colony

Various biologically inspired approaches to problem solving using a social metaphor have been proposed. For example, both Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) have been employed to solve problems in optimization and design. Both approaches employ simple social interactions between agents to produce emergent social structures that are used to solve a given problem. In this paper we investigate the emergence and power of more complex social systems based upon principles of cultural evolution. Cultural Algorithms employ a basic set of knowledge sources, each related to knowledge observed in various social species. These knowledge sources are then combined to direct the decisions of the individual agents in solving optimization problems. Here we develop an algorithm based upon an analogy to the marginal value theorem in foraging theory to guide the integration of these different knowledge sources to direct the agent population. Various phases of problem solving emerge from the combined use of these knowledge sources and these phases result in the emergence of individual roles within the

population in terms of leaders and followers. These roles result in organized swarming in the population level and knowledge swarms in the social belief space. Application to real-valued function optimization in engineering design is used to illustrate the principles.

Ant Colony Algorithm

Ant Colony Optimization (ACO) is based on the observation, in the laboratory, of colonies of ants. Investigators found that ants are able to find the shortest path from the source of food to the nest without using visual tracks (Hölldobler et al, 1990). Also it was observed that they were able to adapt to the changes on the environment, for example, using a new path once the previously used stops being feasible due an obstacle.

It is well known that the primary manner how the ants form and maintain a path is by using pheromones. In the ant colony proposed by (Dorigo et al, 1996), an ant is frequently defined as an agent of simple calculations that iteratively constructs a solution to a problem, a problem of trajectory planning. In this model, ants deposit certain amount of pheromones, a chemical substance, whereas they walk, and each ant probabilistically prefers to follow a pheromone-rich direction. Thus, the pheromone and its density through the path are the knowledge that in ACO shares among the individual ants. The partial solutions of the problem correspond to states where each ant is moved from one state to another. ψ corresponds to a more complex partial solution. In each step σ , each ant k calculates a set of feasible expansions to its present state, and then it moves to one of these, according to a distribution of the specified probability as it follows:

$$p_{i\psi}^k = \begin{cases} \frac{\alpha \cdot \tau_{i\psi} + (1-\alpha) \cdot \eta_{i\psi}}{\sum (\alpha \cdot \tau_{iv} + (1-\alpha) \cdot \eta_{iv})}, & \text{if } i \in \text{tabu}_k, \text{ } i \notin \text{tabu}_k \\ 0, & \text{otherwise} \end{cases}$$

Here the set tabu_k represents a set of feasible movements for ant k and the parameter α defines the relative importance of the path with respect to its attraction.

After each iteration t of the algorithm, trails are updated using the following formula:

$$\tau_{i\psi}(t) = \rho \tau_{i\psi}(t-1) + \Delta \tau_{i\psi}$$

where ρ is a user-defined coefficient and $\Delta \tau_{i\psi}$ represents the sum of the contributions of all ants that use move $(i\psi)$ to construct their solution. An iterative process increases the level of those cells related to moves that were part of "good" solutions, while decreasing all others. The pseudo-code from (Maniezzo, 2000) describes how the basic Ant Colony Optimization works:

1. (Initialization)

initialize $\tau_{i\psi}, \forall i, \psi$

2. (Construction)

For each ant k do

repeat

compute $\eta_{i\psi}, \forall i, \psi$

choose in probability the state to move into

append the chosen move to the k -th ant's set tabu_k

```

until ant k has completed its solution
[apply a local optimization procedure]
enddo
3. (Trail update)
For each ant move  $(i, \psi)$  do
    Compute  $\Delta\tau_{i,\psi}$  and update the trail values
4. (Terminating condition)
If not (end_condition) go to step 2

```

ACO is applied extensively to problems of optimization of trajectory planning in many areas like symmetric and asymmetric variants of the travelling agent, as well as problems of partitioning and the associated times of telecommunication networks. What emerges from the social interaction in the colony of ants are the trajectories of high performance according to the defined terms of performance given by certain functions. Some unexpected trajectory characteristics have been demonstrated for certain kinds of problems such as to find the minimum cost path of a general graph. Recently, a number of algorithms inspired in the behaviour of social groups has been used to solve complex problems of optimization. Some of these algorithms include the particle swarm optimization (PSO) (Kennedy & Eberhart, 1995), ant colony (ACO) (Dorigo et al., 1995), and cultural algorithms (Reynolds et al., 2005). These three algorithms use a model based on a population as the base of the algorithm and solve problems by sharing information via the social interaction among agents.

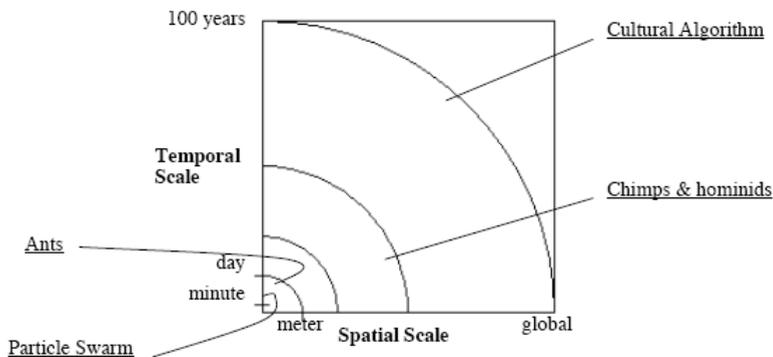


Fig. 1. Scale of Social Interaction.

Figure 1 expresses each of these approaches in terms of space and the continuous time on which the social interactions happen. We must notice that colony of ants and particle swarm are found near the left end of this continuity. For example, in particle swarm, the agents can locally change their speed and direction of movement by the interaction with other agents. In ant colony, the agents locally interchange the information in terms of the density and gradient of pheromone that marks their path. The pheromone is deposited by an ant that moves through the path. The frequency in the use of a path is indicated by the amount of pheromone that has been deposited, considering its degradation in the atmosphere up to a certain level. The cultural algorithms on the other hand allow agents to act reciprocally in diverse ways using several reflective forms of symbolic information present in complex cultural systems. The basic cultural algorithm allows that individuals communicate via a shared space of beliefs. The shared space considers five basic types of information that can

be shared mentally or symbolically. It is well known that the scale of interaction within complex systems affects the nature of the structure arising from the interaction of agents within that system (Holland, 1998). Now we examine briefly each one of the three social models for solution of problems in terms of the nature of their social interactions and their own distinctive features.

4. Behaviour in group intelligence techniques – particle swarm optimization

Particle Swarm Optimization (PSO) is a stochastic technique based in the optimization of a population, developed by Kennedy and inspired by the social behavior of the flock of birds or the shoals of fishes. In PSO, the potential solutions, called particles, move through the space of the problem following the trajectories of their optimal neighbors. Each individual particle does not lose either their last better aptitudes or the ones of its neighbors (social vision) within a fixed radius. This information determines its following direction and speed (cognitive vision). PSO is initialized with a particle group at random (solutions) and later looks for optimal degrees putting the day the generations. In each iteration, each particle is bought up to date better following both "values". The first best solution (aptitude) that has reached until now (the value of the aptitude also is stored). This value is called pbest. The other "best" S-value is followed by the optimizer of the particle accumulation, which is the best value obtained until now by any particle in the population. This second better S-value a best global and call gbest. When a particle participates in the population like its topological neighbors, the second best S-value the best premises and lbest is called. After finding better values the particle updates both its speed and position with the following equations (a) and (b):

$$v[] = v[] + c_1 * \text{rand}() * (\text{pbest}[] - \text{present}[]) + c_2 * \text{rand}() * (\text{gbest}[] - \text{present}[]) \quad (\text{a})$$

$$\text{present}[] = \text{present}[] + v[] \quad (\text{b})$$

$v[]$ is the particle velocity, $\text{present}[]$ is the current particle (solution). $\text{pbest}[]$ and $\text{gbest}[]$ are defined as started before. $\text{rand}()$ is a random number between (0,1). c_1, c_2 are learning factors. Usually $c_1, c_2 = 2$.

The pseudocode of the initial version of PSO for variables of real value is determined by (Kennedy et al., 2001) of the next way:

```

For each particle
  initialize particle
End For
Do
  For each particle
    calculate fitness value
    if the fitness value is better than the best fitness value (pBest) in history
      set current value as the new pBest
  End for

```

choose the particle with the best fitness value of all the particles as the gBest

```

For each particle

```

```

  calculate particle velocity according equation (a)

```

```

  update particle position according equation (b)

```

```

End for

```

```

  While maximum iterations or minimum error criteria is not attained

```

In order to simulate the individual and interpersonal learning of the cultural transmission (social), PSO manages simplicity and effectiveness (speed of convergence). It has performed well in a variety of test problems. It operates suitably on two dimensions but they can theoretically be extended to multiple dimensions. Due to the simplicity of its social behavior with respect to its results with base in the convergence, ranks of convergence, and other unexpected characteristics have been produced. The unexpected characteristic base is the swarm of particles, or coordinated movement of individuals through the space search towards the optimal solution.

5. Behaviour in group intelligent techniques – cultural algorithms

The Cultural Algorithms (CAs) is a class of computer models derived from the observation of the process of cultural evolution in the nature (Reynolds et al., 2005). CAs has three main components: a population space, a belief space, and a protocol that it describes as the knowledge is interchanged first in both components. The population space can support any population based on a computer model, such as the Genetic Algorithms and the Evolutionary Programming. The basic framework is shown in Figure 2.

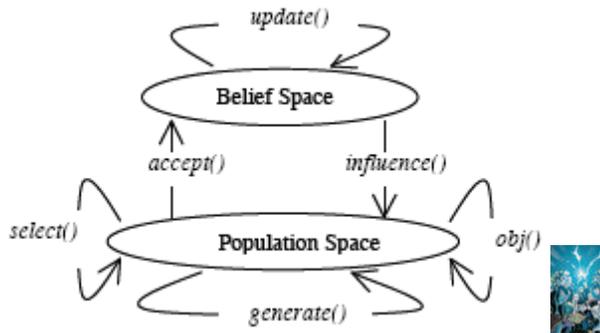


Fig. 2. Conceptual Diagram of Cultural Algorithms.

The cultural algorithms are a dual system of the inheritance that characterizes the evolution in human culture in the macro-evolutionary level, that happen within the space of beliefs, and in the micro-evolutionary level, that happens in the population space. The knowledge produced in the population at micro-evolutionary level is accepted or goes to the space of beliefs, and is used selectively to fit the knowledge structures there. This knowledge can then be used to influence the changes in the population in the following generation.

What differentiates the cultural algorithms from the PSO or the ACO approaches is the fact that cultural algorithms use five basic types of knowledge in the process of resolution of problem instead of only the transmitted value. There is evidence in cognitive sciences that each of these types of knowledge is present in several animal species and it assumes that human social systems support as minimum each of these types of knowledge. The sources of knowledge include the normative knowledge (ranges of acceptable behaviors), the circumstantial knowledge (the units or the memories of right and failed solutions among others), the domain knowledge (knowledge of the objects in the domain, the relations among them, and their interactions), the historical knowledge (temporary landlords of the behavior), and the topographic knowledge (space landlords of the behavior). This set of

categories is seen as complete for a given domain in the sense that all knowledge available can be expressed in terms of a combination of one of these classifications.

6. Evolve computing applying to diverse fields that use a heuristic optimization

The cultural algorithms have been tested with benchmarking problems (Chung et al., 1998) and also have been applied successfully in a diverse number of application areas such as modelling the evolution of agriculture, learning conceptualizing (Morrison et al., 1999), optimization of real-value functions (Jin et al., 1999), re-engineering of the knowledge bases for the manufacturing assembly processes (Rychlycky et al., 2003), modelling systems of price incentives based on agents, distribution of elements in a diorama, in predicting a ranking in Eurovisión, in the simulation of a social model in an intelligent game (Ochoa, Ponce et al., 2007), and combined with Predator/Prey Game for analyzing cultural problems (Ochoa, Quezada et al., 2008) among others. Whereas it is right, the relative complexity of the sources of knowledge and its interaction make difficult to determine because the cultural algorithms work so well. Indicated alternatively, under what conditions some systems indeed can solve a given problem and the belief space can be seen as samples For a right process to solve the problem.

We begin this section examining how cultural algorithms solve problems of resources optimization within an experimental atmosphere. In our research, first we use a simulated atmosphere of the world of cubes, and we adapted it for our experiments. Within this world, the resources are distributed in piles (cones) in the land (Sugarscape style) (Epstein et al., 1996). We can put in the landscape an arbitrary number of cones, each with a size that varies, to produce surfaces of the forage of the complexity that varies. The distribution of cones can be static, dynamic, and deceptive (in which the positioning of some cone hides better areas of agents than they climb the hill). Then the agents build reciprocally social via these several sources of knowledge to find the optimal degree, and in the dynamic atmosphere they don't lose his position, that changes in a certain rate. Then we investigated the appearance of social landlords in both the population space and the belief space when the problem is solved successfully. We use later what we have learned of this experimental atmosphere to solve complex problems in engineering design. Then we observed whether the similar social structures emerge there. We compared the operation of the systems with the optimization of the particles swarm.

Since our problem of the cones world can be described as a search space problem, we used a framework inspired by theoretical results of the biology of populations. Specifically, the agents select diverse sources of knowledge based on which we characterized like "marginal value of información". The inspiration for this comes from the classic work (Charnov, 1976) referring to "marginal theorem of value". In certain situations, the agents who used the marginal value theorem could optimize their product of resource within an atmosphere. Simply stated, the marginal value theorem says that an agent remains within a location in the land until the present resource is minor that the predicted half value. Then one moves to another cell that satisfies each marginal constraint with the value. Here we used an approach to the integration of the knowledge that uses a principle corresponding to "marginal value of knowledge". With this approach, is more probable that an individual uses a strategy that is on the average than it uses other sources of knowledge. Then it was observed that the social organizations emerge in the population space and that the space of

beliefs assures the total success for the system. We use this approach of integration based on marginal value with an evolutionary model of programming for the population and with the five sources of knowledge as the frame of the base to the problems of solution within the simulated atmosphere and of the real world.

We wished to show that the use of the marginal value helps to the integration of the knowledge producing the following emergent structures and behaviors:

1. The appearance of certain phases to solve problems in terms of the relative operation of diverse sources of knowledge in a certain time. We labelled these phases like: heavy, granular, granules fine and backward movement. Each phase is characterized by the domination of a set or subgroup of knowledge sources that are the best to generate new solutions in that phase. Indeed, the dominant subgroup of knowledge sources is often applied in a specific sequence within each phase. It appears a type of knowledge that produces new solutions that therefore are operated on the other knowledge source. The transitions between the phases happen when the solutions produced by a phase can be exploded better by the sources of knowledge associated with the following phase.
2. The appearance of groups of individuals that move within the space of the problem as a result of the interaction of the cultural knowledge. From these phases they emerged continuously in static, dynamic, and deceptive atmospheres when the marginal approach and the integration of the s-value used. We call them "cultural groups of populations".
3. Then we observed the group knowledge in a put-level, we called them "clusters of knowledge". Due to this, the sources of the group knowledge in the goal level were produced by the interaction of the sources of knowledge via the marginal value theorem and this one serves to induce a group in the level of population.

7. Cases of studies of data mining to improve evolve computing

Diverse applications inspired in social data mining combined with evolve computing have a great value. In this section, we present two of them in order to compare their contributions; the first is related with a hybrid system that permits to determine the ranking of a debutant country in the Eurovision Song Contest. In the other, it is showed an intelligent tool which accommodate societies in a diorama which performs the accommodations by evaluating cultural and social differences.

7.1 Hybrid system using data mining and particle swarm optimization to determine the ranking of a new participant in Eurovision.

Many problems involve not structured environments which can be solved from the perspective of Particle Swarm Optimization (PSO). In this research we analyze the voting behavior in a popular song contest held every year in Europe. The dataset makes possible to analyze the determinants of success, and gives a rare opportunity to run a direct test of vote trading from logrolling. We show that they are rather driven by linguistic and cultural proximities between singers and voting countries. With this information it is possible to predict the score of a new country, and distribute the votes for a lot of the participants; this paper tries to explain this social behavior.

The Eurovision Song Contest (ESC) held for the first time in Lugano, Switzerland, in 1956, where seven countries competed. Non-European countries can also take part: Israel,

Morocco, Turkey, Armenia and Georgia are now regular participants. In 2008 Azerbaijan and San Marino will participate for the first time. Since 2002, there are 24 slots for finalists, four of which are reserved for the Big Four (France, Germany, Spain and the United Kingdom). Each ESC is broadcasted by television; in 2001 the contest was broadcasted live all around the world. Nowadays, it is watched by several hundred millions of people. The ratings are normalized so that the favorite song gets 12 points, the next one 10, and then 8, 7, 6, 5, 4, 3, 2 and 1. This allows each voting country to give positive ratings to ten other countries. Participating countries cannot vote for their nationals. The order in which candidates perform is randomly drawn before the competition starts. When performance ends, countries are asked to cast their votes. Results are announced country by country, in the same order in which participants performed. Participants are ranked according to their aggregate score. Eurovision have been studied with different perspectives: the compatibility between countries and the political and cultural structures of Europe (Rauhlen, 1997), the persistent structure of hegemony in the Eurovision Song Contest (Suaremi et al., 2006), cultural voting (Yair, 1995) and the analysis about Grand Prix which evaluate many countries participating in different years and with different many of countries competing (Yair et al., 1996), among others. This research is novel because analyze the behavior of all countries when arrived a new country in a new ESC. The objective is to predict the final ranking of Azerbaijan and San Marino, the new contenders in Eurovision Song Contest 2008. The organization of this section is the following. The analysis of the 52 ESC editions to incorporate *a priori* knowledge about the voting patterns and relationships between neighbor countries is explained. Next, the problem statement is defined. The COPSO algorithm is thoroughly explained. Our approach is then tested in the ESC 2007. The experiments and the analysis applied to estimate the final ranking of Azerbaijan and San Marino in ESC 2008 are explained and finally the conclusions are provided.

Eurovision Ranking using Data Mining.

Data mining is the search of global patterns and the existent relationships among the data of immense databases, but that are hidden inside the vast quantity of information (Ochoa A., Meneguzzi, P. Et al, 2006). These relationships represent knowledge of value about the objects that are in the database. This information is not necessarily a faithful copy of the information stored in the databases. Rather, is the information that one can deduce from the database. One of the main problems in data mining is that the number of possible extracted relationships is exponential. Therefore, there are a great variety of machine's learning heuristics that have been proposed for the discovery of knowledge in databases. One of the most popular approaches to represent the results of data mining is to use decision trees. A decision tree provides a procedure to recognize a given case for a concept. It is a "divide and conquer" strategy for the acquisition of the concept (instance). Decision trees have been useful in a great variety of practical cases in science and engineering; in our case we use data mining to characterize the historical voting behavior for each country. Thus, we selected societies that have participated and characterized its behavior based on their previously emitted votes, which allowed to describe in great detail both the society and the individual. The purpose is to explain v_{ij} , the vote (that is, the number of points) casted by the people of country $i \in L$ in evaluating the performer of country $j \in L$ ($i \neq j$, since country i can not vote for its own candidate), where L is the total number of participating countries. If countries i and j ($i \neq j$) exchanged their votes, without taking into account any other feature, the voting equation could simply be written

$$v_{ji} = \alpha_{ij}v_{ij} + u_{ij} \quad (1)$$

Where α_{ij} is a commitment parameter, and v_{ij} a random disturbance. If exchanges of votes were "perfect", and both countries kept their commitment, α_{ij} would be equal to 1. More generally, such an equation should contain variables $k=\{1,\dots,K$ representing the characteristics (language of songs -English, French, Italian-, lyrics, music, genre and others) of a performer (singer or band) from country i , and variables representing the performances of the country i along its T_i participations in the ESC.

$$v_{ji} = \alpha_{ij}v_{ij} + \beta \sum_{k=1}^K x_{ik} + \gamma \sum_{t=1}^{T_i} z_{it} + u_{ij} \quad (2)$$

where β and γ are parameters to be estimated. The part associated with beta parameter is related with the attributes of performance of a song (music, lyrics, language among others) and her/his/their interpreter(s). The part associated with gamma parameter is the related with the performance of a country during the ESC's participations (example: Armenia has participated in 2006 and 2007). A problem arises with the fact that it will appear on the other side of the equation for the observation concerning the vote of country i for the singer representing country j . This can be dealt with in several ways. First, and this is the easiest way, instead of using v_{ij} in the right-hand side, one can use the vote cast in previous competition, say v_{ij}^{-1} , though one could think that countries would not necessarily keep their commitment over time. An alternative is to use only half of the observations along all ESC editions; thus, every v_{ij} that appears in the right-hand side of the equation is not used in the left-hand side.

The voting equation is estimated by linear methods. The influence of the order in which musicians appear in competition has often been outlined. The exogenous order in which candidates perform is thus included as determinant. Other variables include (a) a dummy for host country, determined by the citizenship of the previous year's winner-the variable takes the value 1 for the performer whose citizenship is the same as that of the host country-, (b) the language (Gelbukh et al, 2007), in which the artist sings (English, French, Spanish, Italian, in other), (c) gender of the artist, and (d) whether the artist sings alone, in a duo or in a group. The last group of variables will include linguistic and cultural distances between voters and performers, and may dispense us from using variables that characterize voters. National culture differences are represented by the four dimensions studied in (Ginsburgh et al., 2005). These studies identified and scored the four following dimensions that make for "cultural distances":

Power Distance: It measures the extent to which the less powerful members of a society accept that power is distributed unequally; it focuses in the degree of equality between individuals;

Individualism: It measures the degree to which individuals in a society are integrated into a group; it focuses on the degree a society reinforces individual or collective achievements and interpersonal relationships;

Masculinity: It refers to the distribution of roles between genders in a society; it focuses on the degree a society reinforces the traditional masculine work role of male achievement, control, and power;

Uncertainty avoidance: It deals with the society's tolerance for uncertainty or ambiguity, and refers to man's search for truth.

Table 1: Correlations between Cultural Distances and Linguistic

	Language	Power	Indiv.	Masc.	U. A.
Language	1				
Power	0.205	1			
Indiv.	0.254	0.111	1		
Masc.	-0.092	0.031	-0.128	1	
U. A.	0.319	0.567	0.404	0.083	1

Table 2: Cultural Distances vs Contender Characteristics

	(a)	(b)	(c)	(d)
Quality	0.911 (0.03)	0.914 (0.03)	0.901 (0.03)	0.905 (0.03)
Logrolling	0.028 (0.01)	0.022 (0.01)	0.018 (0.01)	0.016 (0.01)
Order of perf.	0.003 (0.01)	0.002 (0.01)	0.004 (0.01)	0.003 (0.01)
Host country	0.177 (0.24)	0.191 (0.24)	0.155 (0.24)	0.171 (0.24)
Sung in english	0.14 (0.14)	0.193 (0.14)	0.101 (0.14)	0.135 (0.14)
Sung in french	0.353 (0.17)	0.354 (0.17)	0.343 (0.18)	0.347 (0.18)
Male singer	0.139 (0.13)	0.148 (0.13)	0.147 (0.13)	0.154 (0.13)
Duet	0.223 (0.20)	0.147 (0.20)	0.203 (0.20)	0.174 (0.20)
Group	0.1 (0.13)	0.08 (0.13)	0.087 (0.13)	0.079 (0.13)
Language	-	-1.142 (0.22)	-	-0.634 (0.24)

Table 1 illustrates the correlations between the cultural distance and native languages for the countries (Gelbukh & Sidorov, 2006) that are present in our sample. Uncertainty avoidance is correlated with three other variables, but otherwise, distances seem to pick up very different dimensions of people's behavior. One of the most interesting characteristics observed in this experiment were the diversity of the cultural patterns established by each community. The structured scenes associated with the agents cannot be reproduced in general, so time and space belong to a give moment. They represent a unique form, needs and innovator of adaptive behavior which solves a followed computational problem of a complex change of relations. The generated configurations can metaphorically be related to the knowledge of the behavior of the community with respect to an optimization problem (to make alliances to obtain a better ranking). Columns (a) to (d) of Table 2 contain the results of an OLS estimation of equation 2. We first observe that quality always plays a very significant role, which should of course not be surprising. Logrolling is significant only in (a), in which no account is taken of linguistic and cultural distances. It ceases to be in all the other equations once linguistic and/or cultural distances are also accounted for. Note that even when the coefficient is significantly different from zero, its value is very small. Order of appearance plays no role, while among the other variables, the only one which has some influence is "sung in French". Though not all distance coefficients are significantly different from 0 at the level of 5 percent of probability, they all pick negative signs (the larger the

Table 3: Performance Rates

Country	2008	2007
Armenia	0.87	0.64
Ukraine	0.81	0.77
Georgia	0.79	0.61
Serbia	0.78	0.55
Azerbaijan	0.77	-
Ireland	0.68	0.69
Belarus	0.66	0.61
Sweden	0.65	0.64
Turkey	0.63	0.6
Finland	0.62	0.51
Malta	0.61	0.58
Russia	0.60	0.59
Albania	0.59	0.58
Greece	0.58	0.55
Israel	0.57	0.53
Slovenia	0.56	0.54
Bosnia & Herzegovina	0.55	0.51
Hungary	0.54	0.51
Poland	0.53	0.52
Croatia	0.52	0.51
Latvia	0.51	0.49
Belgium	0.49	0.47
France	0.48	0.46
Romania	0.46	0.43
Germany	0.45	0.42
Spain	0.44	0.37
FYR Macedonia	0.43	0.42
United Kingdom	0.42	0.43
Bulgaria	0.40	0.41
Norway	0.39	0.38
The Netherlands	0.37	0.39
Iceland	0.36	0.35
Estonia	0.35	0.34
Portugal	0.34	0.37
Lithuania	0.33	0.34
Moldova	0.32	0.36
Denmark	0.31	0.33
Cyprus	0.30	0.28
Montenegro	0.29	0.21
Switzerland	0.25	0.26
Czech Republic	0.22	0.21
San Marino	0.14	-
Andorra	0.11	0.08

distance, the lower the rating). The Table 3 presents the expected performance rates for 2008. The performance rate tries to predict the country rank through environment variables observed along 52 previous ESC editions. The Table 3 shows the performance rate of the last ESC where Ukraine had the highest rate. In the ESC 2007 the winner was Serbia which had a performance rate of 0.55, below the top-10. The performance rates were estimated based on the characteristics listed in Table 4 and the country performance along previously participations in every ESC edition. For example, in ESC 2007 participated 42 countries hence it was more complex to obtain a second place than in 1981 for example, when only 20 countries participated. Obviously, for the new contenders, Azerbaiyan and San Marino,

there is not historical information available. The information obtained through data mining, denotes a similar behavior of countries into the same neighborhood and with similar characteristics (language, territorial extension, religion, in others). Thus, the historical performance for Azerbaijan was calculated from Armenia, Georgia, Bosnia & Herzegovina and Turkey; and for San Marino was calculated from Italy, Switzerland, Andorra, Monaco, Malta and Luxembourg. The parameters used by the model to calculate the performance rate are: $\beta=0.4$ and $\gamma=0.6$. The model used to calculate the values of Table 3 is the following:

$$r_i = 0.4 \sum_{k=1}^7 x_{ik} + 0.6 \sum_{t=1}^{T_i} z_{it} \tag{3}$$

Where T_i is the number of ESC editions that country i has participated. Equation 3 is a synthesis of the voting model presented in Equation 2. The missed term $\alpha_{ij}v_{ij}$ represents the voting behavior expected between countries i and j . A robust model was developed adding probability terms that reflect the voting history between a judge country I and a contender country $j(v_{ij})$. The complete model and its implicit problem are explained in the next section.

Problem Statement.

The objective of this study is to estimate the position rank of the new contenders, Azerbaijan and San Marino. This implies to estimate the final voting matrix, where every cell j, i represents the score given to contender i by country j ; that is v_{ji} . For attaining a well prediction, the model should controls the voting behavior between judges and contenders taking into account the historical performance that reflects the cultural empathy, the commonality of regions, the returning voting patterns, in others. The estimated performance rate could guide the model towards an optimal voting configuration according to the current expectations of the experts.

The next objective function posses these two important features of the ESC, the voting behavior and the performance rate explained in the previous Section. Notice that Equation 3 is part of Equation 4.

$$\text{Maximize } f = \sum_{i=1}^C \sum_{j=1}^N c_{ij} + 4 \sum_{i=1}^C \sum_{k=1}^S p_{ik} + \frac{2}{max_S} \sum_{i=1}^C s_i * r_i \tag{4}$$

Subject to:

- Country j can not vote for itself.
- Country j just can vote one time for contender i .
- Country j just can give a score k to only one contender i .

Where N is the number of voting countries, C is the number of contenders, S is the number of available scores $S=\{12,10,8,7,6,5,4,3,2,1\}$ and $max_s=12$ is the maximum score. The first two terms represent the performance of the final ranking. In the first term of equation 4, c_{ij} is the probability that a score k was given by country j for a contender country i . Table 5 shows an example of the probabilities c_{ij} of Finland for score $s=12$. Along 52 ESC editions, Finland has received 19 times a score of 12 points from 11 different countries. Sweden and Iceland are the countries which have voted more times for Finland, both with 3 editions. Therefore, they are the countries with highest probabilities C_{ij} . In the second term of Equation 4, P_{ik} is the probability that a country i receives a score k from country j . Table 6 shows an example of

Thank You for previewing this eBook

You can read the full version of this eBook in different formats:

- HTML (Free /Available to everyone)
- PDF / TXT (Available to V.I.P. members. Free Standard members can access up to 5 PDF/TXT eBooks per month each month)
- Epub & Mobipocket (Exclusive to V.I.P. members)

To download this full book, simply select the format you desire below

