Available online at www.sciencedirect.com
ScienceDirect
Journal of Computer and System Sciences 74 (2008) 133–145
JOURNAL OF COMPUTER AND SYSTEM SCIENCES
www.elsevier.com/locate/jcss

The web from a complex adaptive systems perspective ☆

Maya Rupert [a,b,*], Amjad Rattrout [b], Salima Hassas [b]

[a] *Thompson Rivers University, Kamloops, BC, V2C 5N3, Canada*
[b] *LIRIS-University of Lyon 1, 69622 Villeurbanne Cedex, France*

Received 15 June 2006; received in revised form 31 October 2006

Available online 20 April 2007

Abstract

The web continues to grow at a phenomenal rate, and the amount of information on the web is overwhelming. Finding the relevant information remains a big challenge. Due to its wide distribution, its openness and high dynamics, the web is a complex system, for which we have to imagine mechanisms of content maintaining, filtering and organizing that are able to deal with its evolving dynamics and distribution. Integrating mechanisms of self-organization of the web content is an attractive perspective, to match with these requirements. Self-organized complex systems can be programmed using situated multi-agent systems with a coupling between the agents' social organization and spatial organization. This paper explores the web from a complex adaptive system (CAS) perspective. It reviews some characteristic behaviors of CASs and shows how the web exhibits similar behaviors. We propose a model and a prototype of a system that addresses the dynamic web content organization, adopting the CAS vision and using the multi-agent paradigm.

Keywords: Web; Complex adaptive systems; Multi-agent systems; Web content organization

1. Introduction

The web is growing at a tremendous rate. It contains a huge amount of unstructured, distributed, multi-media data. This content provides a great potential source for knowledge acquisition that needs to be filtered, organized, and maintained in order to permit efficient use. The web content's organization and maintenance are tasks hard to achieve because of the wide distribution of the web, its openness and high dynamics. The web is a complex open dynamic network, exhibiting a self-organizing adaptive behavior similar to a complex adaptive system (CAS). In this paper, we analyze the web from a CAS perspective and present a methodology for programming complex systems using the multi-agent paradigm. Adopting the CAS vision, we propose WACO, an approach inspired by social insects for organizing dynamically the content on the web.

The paper is organized as follows: in the first section, a review of CAS characteristic behaviors is presented, outlining how the web exhibits similar behaviors. A new methodology for programming complex systems is presented in the second section and in section three, we describe a model for organizing the web content in which we adopt the

☆ Expanded version of a paper presented at the Workshop Web And Mobile Information Systems (WAMIS 2006).

* Corresponding author.
E-mail addresses: mrupert@tru.ca (M. Rupert), amjad.rattrout@liris.cnrs.fr (A. Rattrout), hassas@liris.cnrs.fr (S. Hassas).

0022-0000/$ – see front matter
doi:10.1016/j.jcss.2007.04.001

CAS perspective and based on the methodology outlined in section three. In section four, we present experiments to demonstrate the effectiveness of this approach, and discuss some observed results. Related work is outlined in section five, and we conclude in section six.

1.1. Complex adaptive systems

There is no single definition of Complex Adaptive System (CAS). Most researchers in this field [1,2] agree that CASs are composed of many interacting parts, giving rise to emergent patterns of behavior. The behavior is said to be emergent because at the macroscopic level, the system exhibits new complex properties that are not found at the local level of the different components. CASs self-organize and adapt to changes in the environment without central control or rules governing their behaviors. CASs are non-linear systems where the "whole is more than the sum of its components" [1]. Each CAS is formed of many interacting agents. From the local interaction of individual agents arises a global behavior or a pattern that cannot be predicted at the local level. In such systems, order can emerge through a process of self-organization [2]. The study of CASs has been applied to different fields such as economies [3], organizations [4], ecologies [5], biology, the immune system and the brain. CASs share several common characteristic features. In the next sections a review of these characteristics will be presented, outlining how these features apply to the web.

1.2. Emergence and self-organization

The main characteristic of a complex system is that it has many interacting components or agents. Examples include atoms, molecules, neurons in the brain, and ants in a colony. These semi-autonomous agents are interconnected, and they interact with each other to form the behavior of the whole system. In the web, there is a large number of agents including users, web authors, search engines, web pages, hyperlinks, web services and software agents interacting with each other in a non-linear way.

From the local interaction between the different agents emerges an organized global behavior of the system. The whole is more than the sum of its parts, and reductionism theory cannot be applied to understand this emergence. The web has been represented by a directed graph whose nodes or vertices are the web pages and whose arcs or edges are the hyperlinks between the web pages [6]. In the web graph, the emergence of scaling is noted [7], i.e. the probability that a web page has k pages linking to it (*indegree*) or a web page is linked to k pages (*outdegree*), follows the power law distribution $P(k) = k^{-\lambda}$. The web graph is not a random network, but it exhibits the property of a scale free network, which develops a degree of self-organization. This topology shows how order is emerging on the web.

In a CAS, there is no centralized control that dictates the system's overall behavior. Agents govern their own rules of behavior at the local level, adapt to their environment, and at the macroscopic level, order emerges. In the web, there is no global control or authority governing web page creation. Web authors are free to add and delete pages and websites and create hyperlinks to any page or node in the web graph. Despite this decentralized process, the web self-organizes into web communities. A web community can be defined as "a collection of web pages such that each member page has more hyperlinks in either direction within the community than outside of the community" [8].

1.3. Adaptation, co-evolution and dynamics

At the edge of chaos where order starts to disappear, agents need to adapt to a changing environment. They change their internal models and behaviors according to their temporal-spatial organization. They also co-evolve to ensure survival in the new environment. In fact, the adaptive behavior of the system cannot be the result of completely random dynamics. Holland [9] describes the evolution of these systems as the result of a strategy combining (random) exploration to maintain a certain diversity, and exploitation to reinforce promising tracks, allowing adaptation. Since its creation, the structure, content and usage of the web have been coevolving and adapting to each other. Websites' personalization and adaptation have emerged as the number of users is increasing constantly, and there is a higher need for websites' providers to adapt their websites to different usage in order to deliver better content.

A CAS changes constantly because of the continuous interactions and interdependence between the different agents and their environment. The dynamics are non-linear, and one cannot predict the behavior of a CAS. The web is a dynamic graph, and it evolves by constantly adding new pages and removing some old ones (Growth model). There is a higher probability of linking new pages to a more connected page (preferential attachment model) [7] and a good example is the Google's web page. The dynamics of the web i.e. the lifetime, age and newness of web pages were studied [10,11]. Both studies conclude that a Poisson process can mathematically model web pages' changes.

1.4. Holland's properties

As CASs are formed of agents interacting with each other, adapting and co-evolving, modeling such systems requires a bottom–up approach consisting of identifying the different agents and their rules of behaviors and interactions. Emergent properties will arise from within the system. John Holland identified seven basic elements of a CAS [1]:

Aggregation is the property by which agents group to form categories or meta-agents that in turn can recombine to a higher level (meta-meta-agents) leading to the complex system. Web pages can be grouped into websites, and websites into web communities that emerge and self organize without centralized control. This self-organization is a result of a retroactive interaction between usage, content and structure. Aggregate behavior is also observed in the appearance of hubs and authorities in the web [12].

Tagging is the mechanism that facilitates the formation of aggregates by assigning attributes or tags used for agent identification. A tag could be the main topic of a web community or the word vector "bags of words" of a specific web page used in text analysis.

Non-linearity is the property in which the emergent behavior of the system is the result of a non-proportionate response to its stimulus. That means the behavior resulting from the interactions between aggregate agents is more complicated than a simple summation or average of the simple agents. Thus the system cannot be predicted by simply understanding how each component works and behaves, and this phenomenon can be noticed in the evolution and growth of the web.

Flows are the physical resources or the information circulating through the nodes of a complex network, and in the case of the web, it is the flow of information from page to page through hyperlinks.

Diversity: The diversity of skills, experiments, strategies and rules of different agents ensures the dynamic adaptive behavior of a CAS. The web has a large number of interacting constituents and this diversity in the web contribute to its robustness. We observe diversity in its usage, structure and content. Web users are classified into random users, rational users and recurrent users [13], and web page authors come from different backgrounds, creating a vast variety of topics. Web pages are also diverse in their structure, much like hubs and authorities pages [12], and they can be divided into five categories: Strongly Connected Components (SCC), IN, OUT, tendrils and tubes, and disconnected [14].

Internal models or schemas are the functions or rules agents use to interact with each other and with their environment. These schemas direct agents' behaviors.

Building blocks are the component parts that can be combined and reused for each instance of a model. Identifying these blocks is the first step in modeling a CAS. Sub-graphs' motifs form the building blocks for the WWW network [15], and web services are the building blocks for distributed web based applications [16].

1.5. Stigmergy and the web

Stigmergy is a concept introduced by the French biologist Grassé [17]. He studied the behavior of termites during the construction of their nest, and noticed that the behavior of workers during the construction process is influenced by the structure of the constructions themselves. This mechanism is a powerful principle of cooperation in insect societies. It has been observed in many insects such as wasps, bees and ants. It is based on the use of the environment as a medium of inscription of past behaviors' effects to influence the future ones. More generally, this mechanism shows how simple systems can produce a wide range of more complex coordinated behaviors, simply by exploiting the influence of the environment. It allows the self-structuring of the environment through the agents' activity: the state of the environment and the current repartition of agents in the environment determine their respective future evolutions. There are several attempts to associate the web with the notion of stigmergy. The possibility of using stigmergic strategies to create self-organized websites was explored [18]. The site structure is flexible and website visitors interact with each other, changing its structure and content with the help of software agents. A stigmergic mechanism can be used for locating sources of information similar to the process used by ants colonies when locating food sources.

The concept of stigmergy was linked to the web [19] since weblogs, web communities, and the research engine Google exhibit a stigmergic behavior. The different weblogs users (or agents) communicate indirectly with each other, leaving traces in their environment and eventually changing it. When one user modifies the environment, the other replies to the new environment and modifies it and so on. The emergence of self-organized web communities is due to an indirect interaction between the creators of web pages and their environment (the web), and pages sharing the same topics point to each other creating clusters. And finally, the PageRank algorithm [20] used in Google "interprets a link from page A to page B as a vote, by page A for page B, and votes cast by pages that are themselves 'important' weigh more heavily and help to make other pages 'important' " [21].

2. Programming CASs with agents

Hassas has proposed a framework for developing CASs for complex networks such as the Internet and the web [22]. The framework combines Holland's basic properties to the stigmergy mechanism:

- Using the situated multi-agents paradigm and embodied intelligence for identifying the building blocks (agents), their internal models and their roles. Situated agents have a constructivist approach, they act based on the situation presented to them. Their perceptions/actions are situated. The agents are embodied in their environment and have a strong relation to it.
- Using the web as the physical environment to facilitate the emergence of aggregates and the flows of information. This spatial representation allows the crystallization of the crucial information on which the system's dynamic is based. It is represented by a dynamic complex graph.
- Using a mechanism of communication between the agents based on a spatial representation and mediated by the environment, such as the stigmergy mechanism. This favors the aggregation of control information and its spreading through the distributed environment.
- Maintaining equilibrium between exploration and exploitation in the behavior of different agents, to allow aggregation (reinforcement) of the building blocks and diversity (randomness). Random exploration allows for the diversification of a search in the environment to discover new tracks, and exploitation allows for the reinforcement of promising tracks to be discovered in a self-catalytic way (i.e. the more a track is promising, the more it will be exploited).

We define a spatial structure in the form of a digital pheromone. This structure is composed of the following fields:

- Label: identifies the nature of the information.
- Intensity: expresses the degree of pertinence of the information.
- Diffusion rate: expresses the distance to which information is spread in the environment.
- Evaporation rate: expresses the persistence rate of information in the environment.

The system is viewed as an artificial ecosystem, populated by agents in which the environment is at the center of the implementation. Agents interact with each other and co-evolve in a shared environment leading to self-organization. We associate to the environment two levels: a physical level represented by a spatial representation, which is coupled to the spatial organization of the multi-agents system (MAS) and a conceptual level related to the action/interactions effects of agents on the environment, which is coupled to the MAS' social organization.

2.1. Spatial organization

The physical environment is represented by a dynamic complex network and contains the set of all existing entities in the network and the different links created as a result of the usage. The topology of the network is dynamic and in perpetual change, which continuously affects the spatial organization. The physical materialization of the environment and its spatial representation as a network allow the materialization of the flux of information and resources: i.e. evaporation and propagation of digital pheromone, or document sharing. The environment self-structures through the processes of self-organization and stigmergy. The physical environment is also the place for the inscription of the users' traces related to actions and interactions. A good example is how the scale free topology of the web expresses its usage. The web represents a physical environment shared by human users. By their usage, these agents leave traces in the environment like the links between web pages. Some of these traces are aggregated by the usage, leading to the creation of strongly connected nodes (hubs). These aggregates are reinforced in a self-catalytic way; the more a node is connected, the more it has a chance to be connected in the future, (preferential attachment [23]). This corresponds to the exploitation mechanism. The different usages also define an important number of weakly connected nodes, which can eventually become strongly connected with the evolution of the usage. This corresponds to the exploration mechanism.

2.2. Social organization

The social organization in a multi-agent system consists of the following actions:

- Dividing the multi-agent society into social units or groups.
- Defining the roles that agents can play in each group (or behaviors of agents within the group) and the different relations between these roles.
- Assigning the different agents to the different groups.

Roles can be formed and assigned to agents by emergent self-organization in the system depending on the state of the environment and on the correlation with the agents' spatial organization. Examples in natural systems include Poliste wasps that may take the roles of nurses or chiefs or foragers depending on the state of the environment (food resources, brood size), and ants can spontaneously take different roles (explorers vs. pheromone followers vs. pheromone depositors). For example, the different behaviors of web communities users are self-organized socially. At the beginning, all users are surfers. Then members of a community start emerging based on common interests expressed in their web pages, forming clusters. When a user becomes aware of entering a community or being a member in that community, his role becomes to reinforce the community and recruit new members.

2.3. Coupling between the spatial organization and the social organization

The spatial organization and social organization are not independent of each other. The retroactive effect of one organization on the other should be taken into consideration. Positions of the agents in the spatial organization are influenced by the roles the agents play in different groups and vice versa; roles and groups emerge in the social organization and are identified by their relationship to the spatial organization.

3. Case study: An approach for dynamic organization of the web content

3.1. Description of the WACO system

The following model illustrates how the CAS principles are applied in the context of web content organization. WACO (Web Ants Content Organization) is an approach inspired by social insects to organize dynamically the web content [24]. In this approach, the web is considered as a complex environment, inhabited by artificial creatures called WebAnts. These creatures, implemented by mobile agents [25], are organized in a colony and mimic some behaviors of natural ants, namely the collective sorting behavior and the food foraging behavior. The content of the web is viewed by WebAnts as a potential source of food that needs to be organized and collected in an efficient way. Documents contained by the websites are considered objects to be sorted following their semantic contents, in order to construct semantic clusters, where a cluster is a set of semantically close documents with respect to a predefined similarity measure. We consider two levels in this application: a higher level corresponding to the extraction of information from documents using any algorithm of text mining and a lower level which uses a synthetic pheromone coding of the extracted information to organize and search the web content.

3.2. Spatial organization of the web

The web is considered as a potentially infinite set of URLs interconnected by links. This set is mapped into a spatial representation, where locations correspond to websites (URLs) connected to other locations via the links described in their documents.

A location may represent an existing URL or a virtual location created to cluster a set of semantically interconnected websites. The clustering is achieved through the collective sorting behavior of WebAnts. WebAnts are creatures implemented by mobile agents that are able to jump from one location to another, analyze the content of the website and assign to it a semantic value by using any standard algorithm of content mining (text mining for instance).

We consider the following entities:

- *Documents*: are the documents accessible at a website. They are modeled by a piece of food that may be collected by a WebAnt if its content is considered as interesting. Collecting a document consists of adding its URL address to the appropriate cluster.
- *Sites*: are websites accessible through an URL, and represent a location to which a WebAnt can jump.
- *Servers*: are virtual locations that are created by clustering similar documents and that serves as a deposit of collected food for WebAnts during their sorting behavior. A cluster creation consists of a dynamic creation of a web page containing all the URL addresses of the collected documents.
- *A spatial representation of the web*: is a map with different locations (sites and servers), interconnected by adaptable links. Since the web is potentially infinite, it is not our intention to represent the whole web. We consider an initial subset of the web that is permanently updated as WebAnts discover new existing sites and create virtual ones. Thus, the web represents a complex uncertain environment that WebAnts explore, and structure by interacting with one another. It is also the medium of their interactions, through the deposit and smelling of the synthetic pheromone.

3.3. Stigmergy: Documents coding by synthetic pheromone use

Documents' contents are analyzed using any content mining algorithm (for instance text mining algorithm). A semantic value is associated with each document. This value is used as a label for a specific pheromone to which WebAnts would be sensitive when looking for documents with similar semantic values. Each semantic topic is identified by a kind of pheromone. The spatial coding of the control information and the meta-information is achieved through a synthetic pheromone with these different fields:

- *Label* (W_k): characterizes the kind of information coded by the pheromone, which is in our context the semantic value of a document (weighted keyword). For coding document content, we use a standard information retrieval technique, weighted vector representation. A keyword vector contains couples of extracted words, associated with a "$tf \times \text{IDF}$" (for term frequency times inverse document frequency) measure, which is a standard information weighting mechanism [20,26]

  $$W_k = L_c \cdot H_c \cdot T_f \cdot \text{IDF}_k,$$

  T_f is the frequency of the keyword in the current document k, H_c is a Header constant ($H_c > 1$ if the word appears in a title, $= 1$ otherwise), and IDF_k is the inverse of document frequency. The linkage constant is L_c ($L_c > 1$ if the word appears in a link, $= 1$ otherwise).
  The weighted keyword vector represents the label for a synthetic pheromone, which is spread in the environment to attract WebAnts searching for similar documents.
- *Intensity* (τ_{ij}): expresses the pertinence of the information. This value is computed at each site i, for each topic j, using the number of documents addressing the same topic, each time $(t+1)$ a new document is added:

  $$\tau_{ij}(t+1) = \rho_j \tau_{ij}(t) + \sum_{k=1,|D_{ij}|} \Delta\tau_{ij}^k(t),$$

  ρ_j represents the persistence rate ($(1-\rho_j)$ the evaporation rate), $\Delta\tau_{ij}^k(t)$ the intensity of pheromone emitted by a document k, on the site i for a topic j at time t, and Δ_{ij} is the set of documents addressing topic j on the site i.

- *Evaporation rate* $(1 - \rho_j)$: expresses the persistence rate of information in the environment. The lower its value, the longer is the influence of the spread information:

  $$\rho_j = |D_{ij}|/|D_i|,$$

  D_{ij} is the set of documents addressing topic j on the site i, and D_i is the set of all documents on the site i. If a site contains heterogeneous semantic content, this information is not considered sufficiently pertinent. So the associated pheromone will evaporate more quickly than those emitted by homogeneous content.
- *Diffusion rate*: expresses the distance to which information is spread in the environment; the higher its value the greater the scope of the information in the environment. We assign to each site i, a distance d_{ij} for each topic j addressed by i, which is computed as the longest path from the site i to the last site addressing the topic j, following a depth first search:

  $$d_{ij} = \mathrm{Max}_k\left(d_{ij}^k\right),$$

  k is the number of links addressing topic j from a site i.

3.4. Social organization of the agents

The social organization of the agents in WACO is achieved through dividing the agents' population into four groups, and assigning different roles to each group. The building blocks are the populations of WebAnts agents, which mimic the collective sorting behavior and the foraging behavior observed in natural ants. Four types of WebAnts were created, each assigned a different task (tags associated with each agent):

Explorer WebAnts are agents that explore the environment (web), following a random path, looking for documents to sort. These agents are able to perceive different kinds of pheromones, corresponding to different semantic values but are more sensitive to the pheromone value of the last document collected. This makes the agents more attracted to collecting documents for which a cluster forming process has begun, than to collecting documents not sorted yet.

Collector WebAnts are agents associated with clusters of collected documents. They have to maintain the formed cluster and organize it locally in a semantically hierarchical structure, from the more general topic to the most specific. This organization is updated permanently as new documents are gathered. These agents compute regularly the synthesis of the site pheromone and update the values of its associated parameters (label, intensity, persistence rate, diffusion rate).

Searcher WebAnts are agents associated with clusters of collected documents. They are specialized agents whose task is to search the web for documents similar to those addressed by their associated cluster, and gather them to reinforce clusters formation. These agents are launched when a cluster reaches a threshold size.

Requests Satisfying WebAnts are user requests' associated agents. They code the user request into a pheromone value, as done with documents' content, and search in the environment for the appropriate cluster, to answer the user request, by following the gradient of the associated pheromone field.

3.5. Mapping the social insects metaphor

The collective sorting behavior is achieved through the spreading by sites (existing sites and servers) of their synthetic pheromone characterizing their contents. When this pheromone is perceived by the explorer WebAnts, they collect the URL of the site and look for an appropriate cluster by smelling in the environment the appropriate pheromone. If no pheromone is found after a predefined number of jumps, they create a server to initiate the clustering operation.

The foraging behavior is accomplished through the spreading of a clustering pheromone by the searcher WebAnts. When a cluster size reaches a predefined threshold, searcher WebAnts are created to enforce the clustering operation. These agents leave the cluster to search in the environment for sites or servers with similar content. During their search behavior, they propagate the location of the cluster through the deposit of pheromones to guide both the explorer WebAnts and the request satisfying WebAnts in their search for clusters during their collecting or searching behavior. By this operation, the greater the size of a cluster, the more its location is propagated.

The combination of collective sorting and collective foraging behaviors permits a permanent structuring of the web into clusters representing a concentration of documents addressing the same semantic topic, and the propagation of this information on the web, making easier and more pertinent the search for related documents during the information retrieval process.

3.6. Organization and regulation of the agent population

In the WACO approach, the WebAnts are created in a dynamic way, and they adapt to their environment and co-evolve. This process requires a mechanism of managing and regulating the population of agents. We use a mechanism of energy distribution and consuming [21]. WebAnts are sensitive to some notion of order, which is obtained by semantic organization of the web content. The higher the disorder on the web, the more active are WebAnts. Activity of agents is regulated by a mechanism of energy distribution, provided by the environment and directly associated with the notion of order in the global environment. Disorder in the environment generates energy, which is captured by agents making them increase their activity and number. Initially agents are created with a quantum of energy provided by the amount of the disorder on their environment (web). Each operation achieved by an agent consumes a unit of energy. The amount of energy attributed to an agent increases by a threshold value of units when a new site (documents) is discovered.

(1) WebAnts' duplication: When the agent's energy level reaches a predefined maximal value, the agent creates a clone of itself, and divides its energy by two. Through the cloning process, agents ensure the distribution of the workload.
(2) WebAnts' disappearance: When an agent consumes all its amount of energy, it disappears. This means that there is no work to do in its neighborhood. Through this process, we ensure that as long as there are documents to sort, agents are created. When too many agents are created, the population size is regulated by the environment's action. An agent that spends all its energy in looking for some amount of work to do will disappear in an environment, whereas others achieve all tasks.

4. Simulation and results

In this section, we describe some experiments done by simulation to show the effectiveness of our approach. Experiments were made in BREVE (http://www.spiderland.org/breve/), which is a 3D simulation environment for distributed systems and artificial life, using an object-oriented programming paradigm. BREVE provides a continuous space, where mobile objects evolve, respecting some rules inspired by life. Objects are characterized by their position and speed of moving. BREVE provides a library with many facilities to develop artificial life-inspired systems.

In the simulation environment, the web is represented by a continuous space, where sites are represented by locations characterized by their spatial coordinates. Each site contains a set of documents, and is attributed a semantic value describing the semantic topic addressed by its content. The system is initialized following a uniform random distribution of sites and semantic values. An initial set of WebAnts is initialized by BREVE objects, which are distributed following a uniform distribution. To reproduce the dynamic evolution of the web, new sites and/or new topics are randomly created.

We have made four experiments with the following parameters as initial settings: 100 agents, 10 000 sites, 2000 semantic topics, 10 as a threshold of tolerated semantic distance between documents of a same cluster, 10 as a threshold size for a cluster. When this threshold is reached, clusters associated with specialized agents are created. The frequency of new sites' appearance is 0.25.

4.1. Experiment 1: Order increasing and maintained in the system

We studied the notion of order on the web. This notion was associated with the emergence of clusters with similar semantic contents. The function of local order for a given site is expressed by the number of sites with similar content in its neighborhood. Similarity is computed with respect to a specified threshold distance between associated weighted keyword vectors.

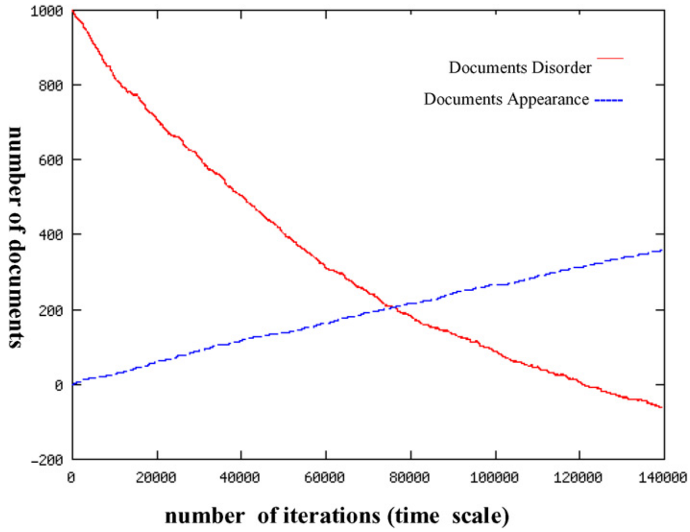

Fig. 1. Disorder decreases while new documents are created, and sorting occurs.

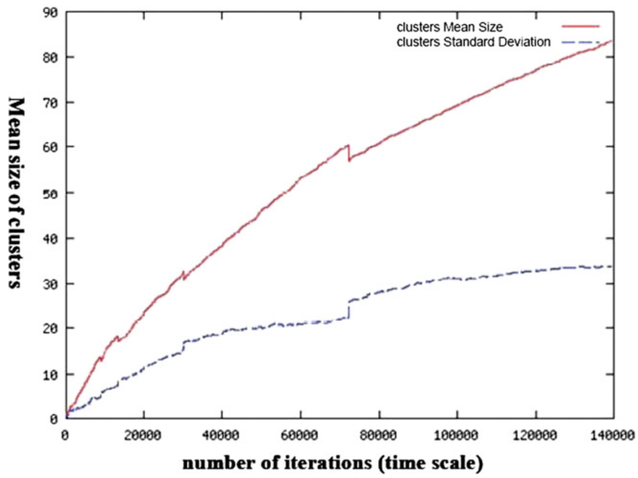

Fig. 2. Clusters mean size and its standard deviation evolution.

On Fig. 1, the x-axis represents the number of iterations (time scale) and the y-axis represents the number of documents. Disorder decreases regularly in the system whereas new documents apparition increases. Disorder is measured by the total number of documents minus the number of clustered documents.

4.2. Experiment 2: Clusters forming and size evolution

Figure 2 represents the evolution during time (x-axis) of the mean size of clusters (y-axis). We can observe the evolution of clusters forming and the increase of their sizes. This figure shows, when compared to the Fig. 1, how the clustering behavior reinforces the creation of clusters, while the disorder decreases in the system.

4.3. Experiment 3: Energy distribution and evolution in the system

Figure 3 shows the evolution during time (x-axis) of the mean value of energy (y-axis) in the system. The energy evolves in the same way as the evolution of disorder and clusters forming over time. We notice some decrease in its

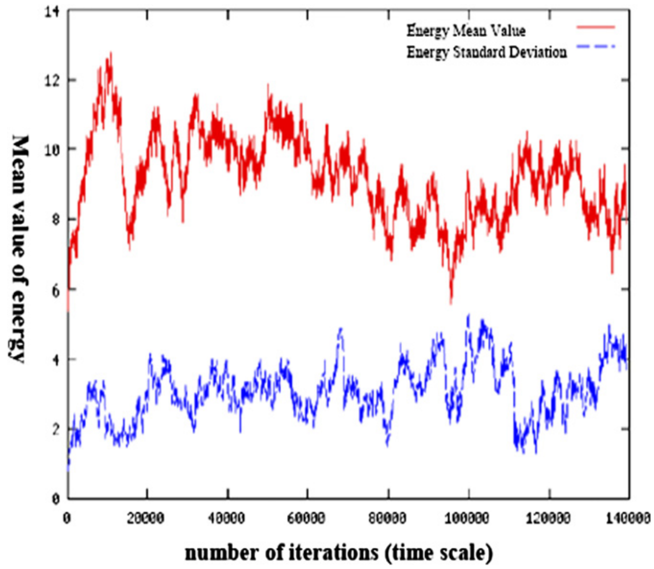

Fig. 3. Mean values and standard deviation values of agents' energy evolution.

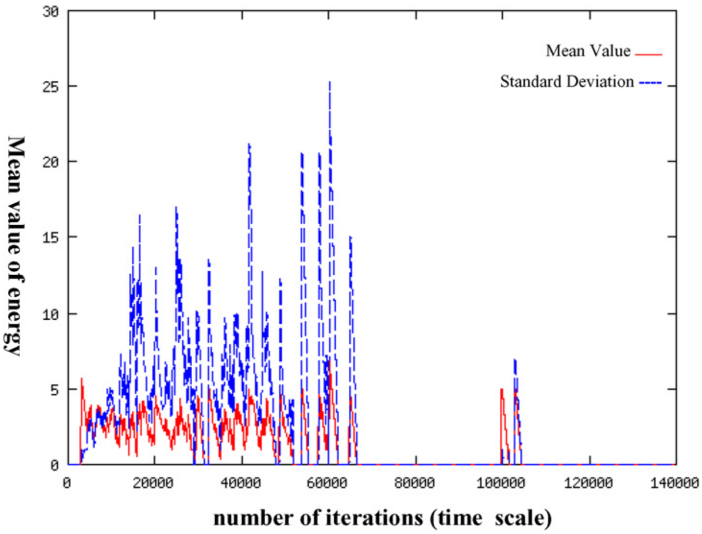

Fig. 4. Mean values and standard deviation values of Searchers WebAnts energy evolution.

value when order emerges (at time 80 000 on the figure), and a regular increasing of this value that corresponds to the activity held by agents as the apparition of new sites increases (time 100 000).

Figure 4 shows the mean value of energy of specialized agents (Searcher WebAnts). These agents increase their amount of energy during the formation of clusters. When clusters are formed (time 80 000), searcher WebAnts disappear (energy value $= 0$). We observe at time 100 000, a sudden increase of energy, which is associated with the appearance of new clusters, as new sites are created or new documents are discovered.

4.4. Experiment 4: Evolution of the population of agents and regulation of their activities

Figure 5 (y-axis represents size of population, x-axis represents time) shows the evolution of agents' populations in the system and the proportion of active agents with respect to the whole population. Until time 80 000, which is the time of emergence of a global order, all agents are active, and after this time, the number of active agents decreases.

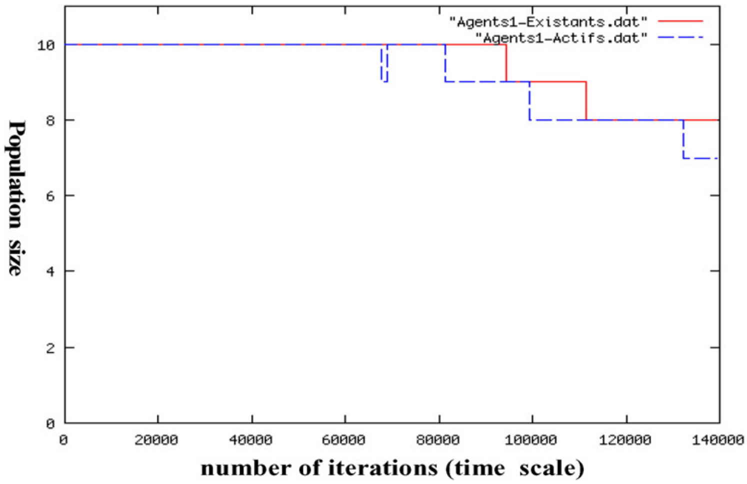

Fig. 5. Population evolution: proportion of active agents/total population of agents.

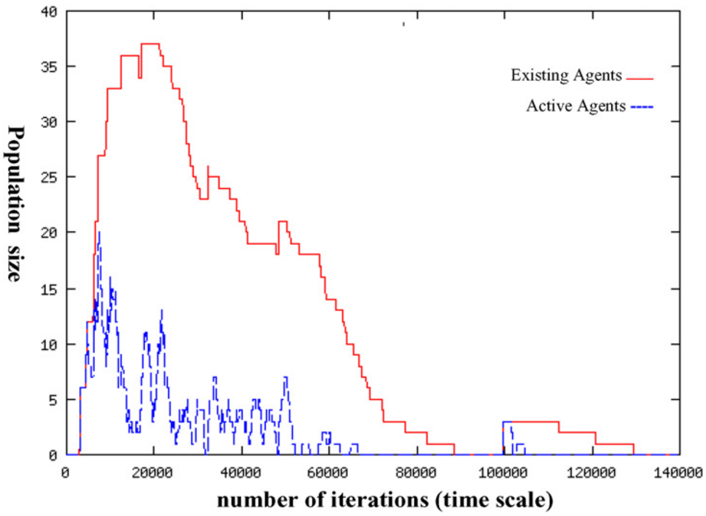

Fig. 6. Evolution of Searchers WebAnts vs. all existing agents.

Inactive agents disappear, reducing in this way the number of initial agents. All agents are active again during the formation of new clusters as new sites are created (time 100 000). After this time, we observe the same phenomenon as at time 80 000.

Figure 6 shows the evolution of the population of Searcher WebAnts. Their number increases at the beginning of the clustering operation and decreases progressively as clusters are formed until their total disappearance at time 80 000, and at time 100 000. When new cluster forming begins, their number increases again.

5. Related work

Search engines have been proposed (AltaVista, Excite, Google,) to provide users with efficiency, performance and a diversity of services. However, techniques used by current search engines suffer a lack of pertinence in content-based retrieval processes, and inaccuracy with regards to the fast dynamic evolution of the web content. This insufficiency is mainly the result of search engines creating indexes of web documents, which could be viewed as "snapshots" of the web. These indexes need to be updated as frequently as changes occur on the web. Some systems use crawlers that exhaustively visit and revisit every web page to maintain its content, but as this maintenance is achieved periodically, at any given time an index will be somewhat inaccurate or incomplete. In our approach, the system follows the permanent evolution of the web content and updates its changes as frequently as they occur.

Several agent-based methods were used to retrieve information from the web. Agents study the users' profiles, search the web retrieving the information requested and suggest appropriate recommendations. They do so via communicating protocols, coordination and negotiation in order to satisfy the users' needs. A multi-agent learning approach was proposed for web information retrieval using a neural network [27]. Agents learn from the user's interests using artificial neural networks and locate the appropriate information resources and search tools for efficient and relevant retrieval of the desired information. A multi-agent system for multi-target search [28] allows users to define requests with multiple interdependent objectives. Each searching agent will execute a sub-query corresponding to each objective and the different results will be combined and offered to the user. Machine learning techniques have also been used to build self-adaptive agents for information retrieval and information extraction [29]. These learning agents are built based on user preference instructions using theory refinement in machine learning. In the medical field, physicians need to access accurate and most recent medical information. ARCH [30] is an adaptive agent that refines the formulation of the user's query for enhanced search results.

Another point to consider is to which degree current information retrieval systems and search engines fit users' request needs. Several systems [31–34] offer websites' adaptation and personalization. By learning from web usage mining, the user's interaction with the website, a customized or personalized experience is offered to the user at the individual level. Web usage behavior in focused web sites was analyzed [35], and the generated patterns were used to dynamically adapt the websites' content to future users. A survey of different web mining techniques for web personalization can be found in [36]. Adaptive web site agents that assist the users navigating the web by making recommendations about the related documents were developed [31]. Seta2000 is a recommender system for personalization based on the multi-agent architecture, especially the agents' communication and the agents' design [37]. As web content and users' interests are both highly evolving, personalized search engines must handle long-standing queries by adapting to users' profile evolution and contributing to its diversification, by discovering new fields of interests, following the web content evolution. In our approach we deal with these issues by considering different kinds of interacting populations of agents, each with a specific role: users' requests management, users' profiles enrichment, dynamic semantic clustering of web content and content refinement and enrichment.

6. Conclusion

In this paper, we have shown that the web exhibits a complex adaptive system-like behavior. Exploiting the web efficiently remains a difficult task. An approach that takes into consideration the CAS aspect of the web seems to apprehend this difficulty. The WACO model follows the CAS perspective for organizing dynamically the web content. WACO is an agent-based system that self-organizes based on the web's dynamic, and combines the foraging and the collective sorting behaviors observed in social insects. The objective of this model is to study the multi-pheromone structure and the specialization of agents' behaviors and the regulation of their population depending on the dynamic of their environment. The efficiency of the proposed system depends on the fine-tuning of the numerous parameters defined in the digital pheromone structure. We are currently studying the relation between the social organization and the spatial organization of the different agents and the retroactive effect of one on the other. An extended work of WACO consists of developing an agent-based approach based on CAS principles that focus on the association of a semantic to information contained on the web that is represented by a multi-scale space, through the combination of the web content, usage and structure.

References

[1] J. Holland, Hidden Order: How Adaptation Builds Complexity, Perseus Books, Massachusetts, 1995.
[2] S. Kauffman, The Origins of Order: Self-Organization and Selection in Evolution, Oxford University Press, 1993.
[3] W.B. Arthur, S. Durlauf, D. Lane, The Economy as an Evolving Complex System II, Addison–Wesley, Longman, 1997.
[4] E. Mitleton-Kelly, Organisations as co-evolving complex adaptive systems, in: British Academy of Management Conference, 1997.
[5] S. Levin, Ecosystems and the biosphere as complex adaptive systems, Ecosystems 1 (1998) 431–436.
[6] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tompkins, E. Upfal, The web as a graph, in: Proceedings of the Nineteenth ACM SIGMOD–SIGACT–SIGART Symposium on Principles of Database Systems, ACM Press, 2000, pp. 1–10.
[7] A.L. Barabasi, E. Bonabeau, Scale-free networks, Scientific American 288 (2003) 60–69.

[8] G.W. Flake, S. Lawrence, C.L. Giles, F.M. Coetzee, Self-organization of the web and identification of communities, IEEE Comput. 35 (2002) 66–71.
[9] J. Holland, Adaptation in Natural and Artificial Systems, MIT Press, Cambridge, MA, 1992.
[10] B. Brewington, G. Cybenko, How dynamic is the web?, in: 9th International World Wide Web Conference (WWW9), Geneva, 2000.
[11] J. Cho, H. Garcia-Molina, The evolution of the web and implications for an incremental crawler, in: The Twenty-Sixth International Conference on Very Large Databases, 2000.
[12] J. Kleinberg, Authoritative sources in a hyperlinked environment, in: ACM–SIAM Symposium on Discrete Algorithms, 1998, pp. 668–677.
[13] J. Liu, S. Zhang, Y. Ye, Understanding emergent web regularities with information foraging agents, in: Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems, ACM Press, 2002, pp. 457–458.
[14] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph structure in the web, Comput. Netw. Int. J. Comput. Telecommun. Netw. 33 (2000) 309–320.
[15] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Network motifs: Simple building blocks of complex networks, Science 298 (2002) 824–827.
[16] M. Kirtland, The programmable web: Web services provides building blocks for the Microsoft .NET framework, MSDN Magazine, 2000.
[17] P.P. Grassé, La reconstruction du nid et les coordinations inter-individuelles chez bellicoitermes natalenis et cubitermes, la théorie de la stigmergie, Insectes Sociaux 6 (1959) 41–81.
[18] P. Small, Stigmergic systems, online at http://www.stigmergicsystems.com/index.html?, 2004.
[19] J. Gregorio, Stigmergy and the World Wide Web, online at http://bitworking.org/news/Stigmergy, 2002.
[20] S. Brin, L. Page, The anatomy of a large-scale hypertextual {web} search engine, Comput. Netw. ISDN Systems 30 (1998) 07–117.
[21] Google Technology, http://www.google.com/technology/index.html, 2004.
[22] S. Hassas, Systèmes complexes à base de multi-agents situés, University Claude Bernard Lyon, 2003.
[23] A.L. Barabasi, The physics of the web, Phys. World 14 (2001) 33–38.
[24] S. Hassas, Using swarm intelligence for dynamic web content organization, in: IEEE Swarm Intelligence Symposium, IEEE Press, Indianapolis, IN, 2003, pp. 19–25.
[25] S. Fenet, S. Hassas, A.N.T.: A distributed network control framework based on mobile agents, in: Proceedings of International Congress on Intelligent Systems and Applications, Australia, 2000.
[26] G. Salton, C. Buckley, Text weighting approaches in automatic text retrieval, Cornell University Technical Report, 1987, pp. 870–881.
[27] Y.S. Choi, S.I. Yoo, Multi-agent learning approach to WWW information retrieval using neural network, in: IUI '99: Proceedings of the 4th International Conference on Intelligent User Interfaces, ACM Press, 1999, pp. 23–30.
[28] E.D. Nitto, C. Ghezzi, P. Selvini, Using agents for multi-target search on the web, in: SAC '03: Proceedings of the 2003 ACM Symposium on Applied Computing, ACM Press, 2003, pp. 828–833.
[29] T. Eliassi-Rad, J. Shavlik (Eds.), Intelligent Web Agents that Learn to Retrieve and Extract Information, Springer-Verlag, 2003.
[30] S. Parent, B. Mobasher, S. Lytinen, An adaptive agent for web exploration based on concept hierarchies, in: 9th International Conference on Human Computer Interaction, New Orleans, 2001.
[31] M.J. Pazzani, D. Billsus, Adaptive web site agents, Auton. Agents Multi-Agent Systems 5 (2002) 205–218.
[32] M. Eirinaki, C. Lampos, S. Paulakis, M. Vazirgiannis, Web personalization integrating content semantics and navigational patterns, in: Workshop on Web Information and Data Management, ACM Press, 2004, pp. 72–79.
[33] F. Menczer (Ed.), Complementing Search Engines with Online Web Mining Agents, Elsevier Science Publishers, B.V., 2003.
[34] D. Pierrakos, G. Paliouras, C. Papatheodorou, C.D. Spyropoulos, Web usage mining as a tool for personalization: A survey, User Model. User-Adapted Interact. 13 (2003) 311–372.
[35] M. El-Ramly, E. Stroulia, Analysis of web-usage behavior for focused web sites: A case study, J. Softw. Maint. Evol.-Res. Practice 16 (2004) 129–150.
[36] M. Eirinaki, M. Vazirgiannis (Eds.), Web Mining for Web Personalization, ACM Press, 2003.
[37] L. Ardissono, A. Goy, G. Petrone, M. Segnan, A multi-agent infrastructure for developing personalized web-based systems, ACM Trans. Internet Technol. 5 (2005) 47–69.