USING AGENTS TO REORGANIZE THE DYNAMIC INFORMATION RESOURCES IN A SELF ORGANIZED WEB SYSTEM

AMJAD RATTROUT
University Claude BernardLyon France
Al-Quds University, Jerusalem, Palestine
amjad.rattrout@liris.cnrs.fr

RASHA ASSAF
Al-Quds University Jerusalem-Palestine
rassaf@science.alquds.edu

SALIMA HASSAS
University Claude Bernard Lyon France
hassas@bat710.univ-lyon1.fr

1. ABSTRACT

The Web is a constantly growing dynamic environment where the components are changed in non-linear ways. These components represent the targets to researches in order to better understand the behavior of the Web, where the owners and the users in this environment exist as out factors. Web page Usage information is the term which describes ways and methods of using the Web. Various factors affect the use of the diversity of the resources in the Web, The non-linear way of its growth, and the evolution in the methods for how we build the Web pages which eventually leads to reflecting the users' interests. Search engines are created to meet the users need for information on the Web. Generally, the researchers seek user's satisfaction through utilizing these search engines to serve the user. One of the most efficient methods in this domain is the use of semantic measure algorithms to recognize the outputs of the information resources according to the users' needs. The Web is represented as three aspects: Content, Structure, and Usage. Three components can lead to having a semantic Web in order to reinforce the semantic value. This paper will present a model that uses new Web Usage information to see the effects on the semantic values, and how it will help us achieve a robust well organized Web. It will consider the Usage space as the field of our research as we will simulate this environment in the MAS "Multi Agents System" and CAS "Complex Adaptive System "paradigm.

Keywords: Complex Systems, Complex Web, Web Content, Web Usage, Web Structure, and Multi-Agents System.

2. INTRODUCTION

Growth of information is the key aspect of World Wide Web. World Wide Web has become the main source of information. Users think that they can find the optimal solution because retrieval time is less and search gives more related topics which they are looking for. Satisfaction of the users requires a filtered, organized, and maintained data, but because the web environment is a Complex System, it is presented as a direct graph where its nodes are websites and its vertices are links connected nodes with each other. That is why it is hard to predict the growth of information behavior on it. It is hard for a search engine to completely understand the user's desire from the given keywords. The Semantic Web is the solution to this problem which is based on a vision of Tim Berners Lee, the inventor of the WWW; he defined the semantic web as an extension of the current web in which information is given as a well-defined meaning, better enabling computers and people to work in cooperation". Agent software can utilize computation of semantic value to provide personalized user services. The main goal is to offer an efficient utilization of information. Achieving such a goal requires adopting the perspective of the CAS "Complex Adaptive System" model. This leads to finding the WebPages of relevant topics during the search process. Information on the Web is characterized in terms of three main aspects characterized as Content, Structure, and Usage. Inferring from Holland's CAS [1] properties and

mechanisms, and using similarity rules adopted by Menczer model [17], to illustrate a combination of these components.

In section 2, reviewing of CAS characteristic behaviors, in sections 4, we illustrate the WACO model for organizing the content on the web dynamically, and in sections 5, 6 and 7 we propose a model that combines Content, Usage and Structure. We use the combination between the three respects (Content, Structure, and Usage), specially the Usage information to reinforce the semantic value to achieve a self organization Web. Finally, we illustrate the results obtained.

3. THE WEB FROM A CAS PERSPECTIVE

It's believed by most researchers in this field [1, 2] that CAS consists of many interacting parts which give rise to emergent patterns of behavior. The behavior is believed to emerge due to the fact that at the macroscopic level, the system demonstrates new complex properties which are not found at the local level of the different components. CAS is in no need for central control or rules governing it's behaviors since it is adapted and adjusted to changes in the environment. CAS systems are non –linear systems, i.e. "the whole is more than the sum of its components" [1] Global pattern arises from the local interaction of individual agents. Such pattern cannot be predicted at the local level. Such systems allow the emergence of order through a process of self-organization [2]. The study of CAS has been applied to various fields and areas of knowledge such as economy [3], organization [4], ecology (5), biology, the immune system as well as the brain.

3.1. CAS INTERACTIONS

CAS has no centralized control which governs the system overall behavior. At the local level agents govern their own rules of environment and the emergence of order takes palace at the macroscopic level. There is No global control or authority control web page creation since web authors all over the world are allowed to do all the changes they need about web pages and websites as well as being able to create hyperlinks to any page or node in the web graph. However, the web organizes itself into web communities according to hyperlinks Structure analysis. Flake defines a web community in [6] as "a collection of web pages that each member page has more hyperlinks in either direction" within the community than the outside of the interests arise with no central control what so ever.

3.2. USING HOLLAND'S ROPERTIES AND MECHANIZMS

A bottom up approach is required for modeling such a system as CAS. They are composed of agents interacting with each other, adapting and co-evolving in their environment. Such an approach should be able to identify the various agents and their rules of behavior and interactions. The inside of the system gives rise to emergent properties. In this Paper, we used properties and mechanism proposed by John Holland which identifies a CAS as [1]:-

Aggregation: It is the property through which agent group gives rise to categories or Meta-agents which recombine to a higher level (Meta –agents), thus creating the Complex System. Meta–agents emerge because of agent interactions at the lower level. We group Content and Structure by users need into a web page which is grouped into websites which are grouped into web communities (Meta –agents). These Meta-agents arise and self-organize without any centralized control. Self-organization is a consequence of a retroactive interaction between Usage, Content and Structure. Web page designers change the Content and Structure of their web pages due to the fact that the user needs are developing rapidly. Besides, web communities are emerging continuously. Further more, the emergence of hubs and authorities in the web make aggregate behavior observable [7]

Tagging: A Tag might be considered as the major topic of a web community or the word vector "bags of words" of a certain web page that is used in text analysis as well as web page similarity Analysis.

Non-linearity: is the property where the emergent behaviour of the system is the result of a non-proportionate response to its stimulus. That means the behaviour resulting from the interactions between aggregate agents is more complicated than a simple summation or average of the simple agents. Thus the system can not be predicted by simply understanding how each component works and behaves. The growth of the web is a nonlinear process.

Flows: The physical resources or the information circulating through the nodes of a complex network.

Diversity: The diversity of skills, experiments, strategies, rules of different agents ensure the dynamic adaptive behaviour of a complex adaptive system. The web has a large number of interacting constituents and this diversity in the web is contributing to its robustness. We observe diversity in its Usage, Structure and Content. In [8] users were classified into random users, rational users and recurrent users. Web page authors come from different backgrounds, creating a vast variety of topics. Web pages are also diverse in their Structure, like hubs and authorities pages [7], and web pages were divided into five categories: Strongly Connected Components SCC, IN, OUT, tendrils and tubes, and disconnected.

Internal models or schemas: They are the functions or rules that the agents use to interact with each other and with their environment. These schemas direct agent's behaviours.

Building blocks: 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 [9], and web services are building blocks for distributed web based applications [10].

4. Related work

Many methods were presented in this domain of research, but the most efficient was using the semantic measures algorithms. These algorithms were used to reorganize the outputs of the information resources according to the users needs. The researchers have many problems and challenges to achieve for a durable solutions [13,12,17,20]. Some researchers [20] used the method of CAS to understand better the behaviour of the Web as a complex system, where others use the MAS to simulate in a virtual way [13]. In the recent years, researchers start to do combinations between these two systems, and represent this environment as a heterogeneous paradigm that analyses the web from the point of view of the CAS once and MAS in the other. Using these two systems was to come over the non linear growth in the Web information resources and in its complexity. Users are a very important factor in these systems; force the researchers to study their behaviours in order to understand better the Web and the changes that emerge according to this usage. In [19] they present the Web components as an accurate access to the information in the Web information. However, some researchers [18] used some usage information to reorganize the Web resources.

In [11,19, 20] Hassas, Rattrout and Rupert, exploiting the web as a CAS, they have proposed a framework for developing CAS for complex networks such as the internet and the Web using stigmergy mechanism.

They used the situated multi-agents paradigm and behavioral intelligence for identifying the agents, and their roles (tags). Further, they used a mechanism of communication between agents, based on a spatial representation and mediated by the environment, such as the stigmergy mechanism. This favours the aggregation of control information and its spreading through the distributed environment. Finally, they Maintained equilibrium between exploration and exploitation in the behavior of different agents, to allow aggregation (reinforcement) of the agents and diversity (randomness).

In [11, 12, 18, 19, 20] Hassas, Rupert and Rattrout illustrated a model where the CAS principles are applied in the context of web Content organization called WACO. WACO (Web Ants Content Organization) is an approach, to organize dynamically the web Content. The internal organization system that is followed by the WACO model is very closed to various functions followed by the ants in there Complex System of work.

This system is made up of four agents with various jobs. The first agent (*Explorers Web Ants*) would be responsible for the process of discovery of the places of web document in random way to sort it. The second agent (*Collectors Web Ants*), whose function is to keep and organized semantically collected documents. The third agent (*Searchers Web Ants*) whose job is to enforce cluster of collected documents by searching the web for similar documents to add to the cluster. The fourth agent (*Requests Satisfying Web Ants*) whose job is search for the appropriate clustered based on user query .The various groups of Web Ants can achieve their tasks and work together following the feedback system without having to get direction from the internal center. Each semantic topic is identified by a kind of pheromone. But WebAnts can work in a group through a stigmergic, using a multi-Structured electronic pheromone. Synthetic pheromone is coded by a Structure with these different fields:

- *Label* (W_{ij})*:* decide the sort of information classified by the pheromone, which is in our Content the semantic value of a document (weighted keyword).

$$W_{ij} = Lc.Hc.T_f.IDF$$

T_F is the number of times of the term in the present document, the H_C is a Header constant (H_C >1 if the word shown in a titlE, =1 otherwise), which increases the weight of the term if it is shown in the title of the document, and IDF_K is the inverse of document frequency. The linkage constant L_C (L_C >1 if the word is shown in a link, =1 otherwise)

- *Intensity* (t_{ij})*:* which is the term that shows how continuous is the flow of information about a certain topic and how high is the value of the pieces of information and also their attractive power. Every time (t+1) anew document is found, it adds to the site i and so to the topic j, as:

$$t_{ij}(t+1) = r_j t_{ij}(t) + \sum_{k=1,[D_{ij}]} Dt_{il}^{K}(t)$$

r_j represents the persistence rate ((1-r_j) the evaporation rate), Dt_{ij}^{k} (t) the intensity of pheromone comes out by a document k, on the site i for a topic j at time t, and D_{ij} is the set of documents addressing topic j on the site i.

- *Evaporation rate:* it shows the stability of the rate of information on the specific field. So, if the value of information is low, its influence will be longer, and that value is always calculated in relation to the ratio of documents related to that topic and compared to all documents in that site i.

$$r_j = |D_{ij}| / |D_i|$$

D_{ij} is the group of documents about the topic j on the site i, and D_i is the amount of all documents on the site i. Our job is to make the clustering of documents of a certain topic more relative than separated ones. When the site has different semantic Content, it is none as insufficient pertinent and then the joined pheromone is going to evaporate faster than that emitted by the different Content.

- *Diffusion rate*: when a piece of information has higher value than other pieces of information then its scope of information is grater and it spreads in the environment faster. When we browse the Web looking for a topic we express this distance d_{ij} using the linkage topology information. And so we can explore the topic of interest by associating to each site i. The distance of the topic is characterized as the longest path from the site to the last site addressing the topic j, following a depth first search

$$d_{ij} = Max_k (d_{ij}^{k})$$

k is the number of links addressing topic j, from a site i. The idea here is to make sites that are a good entrance point for a search, have a wider diffusion scope than the other ones. Doing so, we could guide the search process to handle queries like "Give me all documents k links away this one". The Web Ants in WACO are created in a dynamic way and they adapt to their environment and co-evolve. Two mechanisms direct their life cycle: *duplication* (birth) and *disappearance* (death).

The WACO Model has achieved the results it has promised. The number of sites in neighborhood can give us an idea about the function of that site locally and its relation to other sites with similar Content. By study, we noticed that disorder decreases all the time in that system while new document's appearances increase. They measure disorder by the total number of document minus the number of clustered documents and this can also show us the effectiveness of the clustering behavior. They tell that there is an evaluation and an increase is taking place every time in sizes of clusters and tell us that clustering behavior reinforce of the creation of clusters. These agents are increased when clusters are formed in large numbers and they decrease when clusters are not formed. If there is a sudden increase in there energy, new cluster appear specially when new documents are discovered or new sites are created. Agents increase in population and a lot of evaluation happen during time which leads to regulation of their activities. They know that all agents are active, and by the time the active agents increase and the inactive ones disappear. And by this way they reduced the number of initial agents. But during the formation of new clusters and the creation of new sites, all agents become active agents again.

5. WEBCOMP MODEL

In this model, we present the problem from three concepts as in figure1.

Figure1: Interrelated relation between the three major items that represent the problematic and the proposed solution

5.1. A MODEL TO COMBINE WEB COMPONENTS

In [18] Rattrout et al. suggest a multi-scale space model called Webcomp. Pages connect to each other whereby hyperlink. Web Pages have Content, Usage, Structure, and values of semantic. Finding the effective approach is the main idea which makes us able to do aggregation between various spaces with out losing the semantic [17] value of pages. We can enable the interaction within a single space of different spaces by using the agent's communication. Figure.2 represents the interactions between these components.

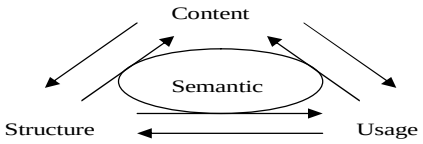

Figure 2: the interaction between the components of the Web

There are two levels of space of this model: the real main space, and the virtual sub space. The virtual spaces enrich the real ones by adding information that are related to their dynamic Structure. Different agents can form the interaction between two pages specially the agents that have the same Head Tags and different values of semantic. We can also combine previous agents that are based on identical positive tags by using the resulted similarities. We can also combine a Multi Agent System by using further agents. We can know about interactions between different space levels by following the virtual dimension and according to the degree of similarity.

5.2. THE MULTI-SCALE SPACE DEFINITION

The Total Documentary Space (TDS) is group of pages tagging from the Web by searching engines according to one (or a set of) keyword(s) (k_j). We can use algorithms of Page rank, Best First Search, InfoSpider [13-16] to make a variety of Web documents (W). These pages (p_i) are downloaded locally with the goal of applying all the needed process.

The page (information) is defined as follows:

$$TDS = \{\forall p_i \in W / k_j \in p\}, \forall p_i \in W \ \& \ p_i \in TDS$$

p_i=<information Content(I_C), information Structure (I_S), information Usage (I_U)>.

- $I_C = \{k / k \in p \& w_{p,K} = f(p,k).i(k)\}$, Where f (p, k) is the frequency of Keywords K in the page. i(k) is the opposite of the logarithmic frequency with regard to the page.
- $I_i = \{1 \in L / i \in Linput\_p \quad OR \quad i \in Loutput\_p\}$, where I_I=<IS_I, IS_O> is the keyword frequency in the page. i(k) is the inverse frequency. We should consider an extraction algorithm for the set of L_{output_p}, L_{input_p}.
- $I_u = \{idp_i / idp_i \in Hp_i\}$, *where* idp_i is the page identifier according to its historic Hp_i.

L Stands for the set of links that exist in TDS. Two verities of agents are known for the Usage of a page pi, a user agent (AgU_j) and an abstract agent. A user agent communicates with the agent of the page p_i (AgP_i) to register the last page visited by the user from pi. So, a history is shared between different agents for the same page. The last information is the major type head tag connected directly to the following level of document spaces. I_C denotes the head tag for the agent representing the Content space. I_S denotes the head tag for the agent representing the semantic space. I_l denotes the head tag for the agent representing the space of links. Finally the I_U denotes the head tag for the agent representing the Usage space. An abstract agent is created for each space according to the type of its main tag.

5.3. MAIN LEVEL SPACES

1. Link Space (E_L): is a graph space resulting from link similarity where, $E_L = \{\forall \ p_{i,} p_j \in TDS / p_i s_L \quad p_j\}$.
2. Content Space (E_C): is a graph space resulting from similarity computation based on Content similarity where, $E_C = \{\forall \ p_{i,} p_j \in TDS / p_i s_C \quad p_j\}$.
3. Usage Space (E_U): is a graph space resulting from Usage similarity computation based on the Usage information similarity where

$$EU = \{p_k \in (Ep_i \cup Ep_j) \,\&\, p \notin TDS\}$$

(Ep_i) is the space of visited pages starting from p_i whoever the user is (u_i). The similarity is calculated between the two spaces of the two pages (p_i, p_j).

4. Semantic space $E_S = \{\forall p_i, p_j \in TDS / p_i s_S \quad p_j\}$. is applied on the three different spaces in order to have a semantic significance for spaces.

5.4. CONTENT SIMILARITY σc

Every relation between two pages in the Web is based on their relation for n terms. The similarity measures σ can be defined from distance measures δ using the relationship [15-17]:

$$\sigma = 1/\delta + 1 \quad Where \quad sc(p,q) = \left(\vec{p},\vec{q}\right) \Big/ \left(\left\|\vec{p}\right\|.\left\|\vec{q}\right\|\right)$$

$\vec{p},\vec{q}$ are representations of the pages in word vector space after removing stop words and stemming. This is actually the "cosine similarity" function, traditionally used in information retrieval IR.

5.5. SEMANTIC SIMILARITY

A semantic similarity between two documents is defined in [17] using the entropy of the documents' respective topics:

$$s_s(d_1,d_2) = \frac{2\log \Pr[t_0(d_1,d_2)]}{\log \Pr[t(d_1)] + \log \Pr[t(d_2)]}$$

Where t(d) is the topic node containing d in the ontology, t0 is the lowest common ancestor topic for d1 and d2 in the tree, and Pr[t] represents the prior probability that any document is classified under topic t.

5.6. LINK SIMILARITY

Link similarity is defined with

$$s_l(p,q) = \left|\mathbf{U}_p \cap \mathbf{U}_q\right| \Big/ \left|\mathbf{U}_p \cup \mathbf{U}_q\right|$$

Where, U_p is the set containing the URLs of page p's out-links, in links, and of p itself. Out-links are obtained from the pages themselves, while a set of in-links to each page in the sample is obtained from the list of the table of the out links that point to the pages exists . This Jacquard coefficient measures in [17] the degree of clustering between the two pages, with a high value indicating that the two pages belong to a clique.

5.7. SIMILARITYCORRELATION AND COMBINATIONS

In [17], there was study by Menczer to know if the various similarity measures are correlated or not. We should analyses the relation between Content, Structure, and semantic similarity functions so as to make a map of the correlations and functional relationships between the three measures a cross tow pages. Menczer tried to approximate the accuracy of semantic similarity by mapping the semantic similarity landscape as a function of Content similarity and link similarity [16]. Averaging highlights of the expected semantic similarity values is akin to the precision measure used in IR. Summing captures of the relative mass of semantically similar pairs is akin to the recall measure in IR. Let us therefore define localized precision and recall for this purpose as follows:

$$R(s_c,s_l) = \frac{\sum_{p,q} d_c(p,q,s_c) d_l(p,q,s_l) s_s(p,q)}{\max_{s_c,s_l} \sum_{p,q} d_c(p,q,s_c') d_l(p,q,s_l') s_s(p,q)}$$

$$P(s_c,s_l) = \frac{\sum_{p,q} d_c(p,q,s_c) d_l(p,q,s_l) s_s(p,q)}{\sum_{p,q} d_c(p,q,s_c) d_l(p,q,s_l)}$$

6. MODELLING AGENTS

We can extract the Web pages from the internet by using many types of search engines [16] and this can give us the advantage of having high quality pages. We have chosen four search engines to cover all the possibilities for sorting out the pages that are characterized by criteria such as Content, Relevance, Popularity, etc.

Our Model makes an original agent that is known as abstract agent every time user opens the session. Their will be three types of agents for the successor concerning it's Content, Structure, and Usage, and each one will represent the space that explores according to similarity rules used by Menczer [14].

Three Agents Content, Structure, and Usage, roll the major activities in our model by the way of tagging the documents of the TDS, carrying back concerned page's tag and similarities resulted to the TDS. The abstract agent tagged by a tag and contains a similarity value. We can't find information about the existence of tags and similarities in the initial state, only the type of tag signed by the agent itself. For each sub-space, an agent created, contains certain

information of its space only when it has acquired enough resources to transfer its information, it becomes active. The agent fits its ability to produce offspring. Another sub-space agent is created and starts to specify their tags according to their properties and results of similarity process. After that an aggregation process holds in, with adhesion between agents represents their spaces forming multi-agent aggregates.

6.1. RESOURSES

We can lay the organization of the agent model if we specify a group of renewable resources of documents information that are dealt within an abstract way. We can also represent agent's resources by following Special characters concerning the type of their original correlated information. In the aggregation process, some mechanisms become active like tagging the different types of resources.

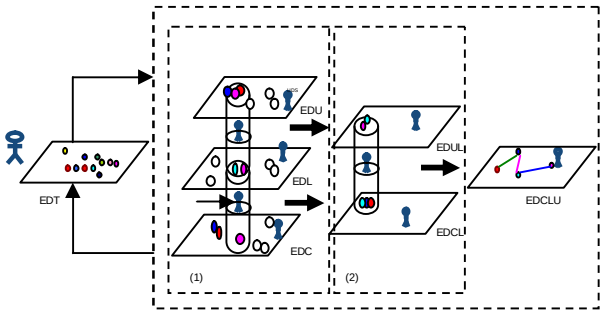

Figure 3: multi-scale space of the Web and life looping

6.2. TAGGING

The "tags" analyzed as word vectors of information that could be shared by different agents. They are the marks that distinguish types of common agents from others. In our model, there are three types of tags, space tags, document tags, and value tags; space tags are classified into four types related to the four sub spaces (ex. Content space tag is Tag_{Tc}), where the document tags are the word vector (frequency of terms > threshold) exists in its Content, links, and Usage information, finally the tags value which marks the pages with tag+ or tags – according to their similarity values where it is above or under the given threshold. The use of Menczer similarity σ_M over TDS in the beginning of our algorithm, will guarantee that the pages will be classified into three different spaces. The Web spaces are ordered using the decreasing values of their similarity. The degree of similarity is also used for the adhesion between agents. Using the tag+ could be considered a starting point in doing aggregation between two agents from the same space or from different spaces. We can determine the tag while we are calculating the similarity values within the same space. The similarity value helps to discover the similar area between two agents.

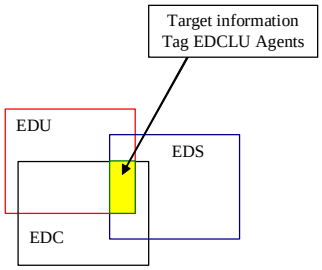

Figure 4: commons spaces zone where agents could find documents that are shared by their different characteristics

6.3. MODEL'S FUNCTION

Our model integrates two types of users, authorized and invited. The first is related to user name and password for the documents treated or used by the user himself. Usage information will be the Usage space information that Usage agents will work in. Similarity, values and tags are found randomly in the sub space pages by the Agents in this model, which have been treated by the tags and similarity agents, follow mining process enriching pages in the TDS. These agents add the positive values and tags to a reserve concerning pi, where these information are considered as input resources for the TDS agents in their space. From collected values, an agent goes out from one page to another, looking for values that match its information. TDS agents could have the three types of information including their total tags. In Figure 6, we show how our agents travel from sub spaces to TDS and communicate between each other in one space. In phase (1) one agent enter to a page randomly, trying to recognize the space that this page belongs to. This information is taken from the tagT, why start with this information? To avoid mixing the space's information while it tries to specialize in their activities, also, to mark the pages that have a high value in their spaces. The agent which is considered as a Content agent when it finds a page with a high value of Content similarity, starts to travel from page to page with the values that this page has until it finds a Tagv in this page that has value above lambda while its Tagv,c is positive, and the energy tag is also

positive. At this moment the agent will change the space to see the page that has a power value from another space. Next the agent from the page value, the agent will continue in exploring the space until there are no more pages to be visited, and then it will die.

6.4. EXPLORER AGENTS

Explorer agents are created to collect information from each page visited by one of the sub space tags and similarities agents, and send it to the TDS. The pages in the TDS have four reserve boxes related to their similarities values, and marked by four different types of signals with an independent reserved tag. If page pi frequently visited by the sub space's agents, recording its new values of similarities, and fulfilling the threshold condition; there exists an agent's trace that will stay sending signals to the TDS agents. Matching signals between agents depends on the strength of the signal itself according to the similarities values. Those who have high signal will do an adhesion, constructing a multi agent's level; otherwise, they will stay individuals.

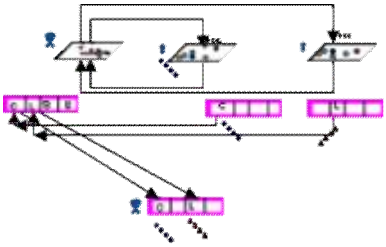

Figure 5: Explorer agents enrich the TDS page by collecting the Tags and similarities values from virtual space.

6.5. AGENT'S ORGANISATION:

In our model, we use the conceptual Agent/Group/Role, where many types of agents are created.

Abstract, simple, global, generator, tags, space explorers, similarities, and information collectors' agents are the main variety of agents who works on documents Content, link, and Usage information. Space information considered as Food for our agents of type search and collector, or tag, where their values of similarity are used also as collected food, but with different assess, and for each agent we have a brain, that generally derives from the abstract agent. These agents roll the major activities in our model by tagging the documents of the TDS, calculate the similarities between pages in each document space, explore the spaces to find which pages have been treated, and collect the tags and similarities values resulted from each space and register it into the TDS.

The Global agent contains two components, the tags information and similarities values.

The generator agent is the agent who generates the agent similarity. Many types of tags exist in our model, as Content tags; and these tags tagged the different sort of information. Links tags, which contains tags tagged the information in the links and tagged the Structure of the pages, and finally Usage tags, where Usage tags are divided into two types; Usage text tags and Usage links tags. To understand this behavior we added another types of tags explaining the matching tags in the different steps of aggregation property.

6.6. AGENT'S CONSTRUCTION

The agent's similarities are three types in Similarity group where each agent has many simple agents:

1. **Agent Content similarity**

This agent calculates the Contents similarity between pages Pi, Pj, where P is the set of the pages of the Total document space EDT as follows:

AgentSimilarity (Agent/Group/Role)
Input: Page P= {P}
$\sigma_c(p,q)=\Sigma(TF.IDF_{t,p} * TF.IDF_{t,q} / ((\Sigma(TF.IDF_{t,p})^2 * \Sigma (TF.IDF_{t,q})^2)^{1/}$
Output: Space (S_c)
Where
Space (S_c) = CDS

For each page Pi we calculate the summation of all the similarities between this page and the rest of the pages in the space, divided by n-1 as follows:

$$s\ (c, \mathrm{Pi}) = \sum s\ \mathrm{cP}/\ \mathrm{n}\ -1$$

2. **Agent link similarity**

This agent works with respect to the following

$$s_{l}(p,q) = \left|\mathbf{U}_{p} \cap \mathbf{U}_{q}\right| \Big/ \left|\mathbf{U}_{p} \cup \mathbf{U}_{q}\right| \text{rule}$$

on EDT and the output will be a virtual space called Link document space: we also calculate for each document the similarity value according to the other documents as:

$$s\ (\mathrm{L},\ \mathrm{Pi})\ =\ \sum\ s\ \mathrm{L},\ \mathrm{P}/\ \mathrm{n}\ -\ 1$$

Input: Page P= {P}
Output: Space S where
SpaceL=SDL

3. Popular Link similarity:

The popular link similarities used as tool to refine or to adjust the results we have from the classical measure in the link similarity tools. We can form the popular link similarity as:

$$s_l(p,q) = \left|\mathbf{U}_p \cap \mathbf{U}_q\right| / \left|\mathbf{U}_p \cup \mathbf{U}_q\right| * \text{Pop}$$

Where Pop is the popularity measured between two inlink factors taken from the concerned two pages p and q. Pop calculated as:

Pop = Max(Pinlink,Qinlink)/ Min (Pinlink,Qinlink)
If Pop is over threshold done by the user then the pop=1
Else Pop = Pop
End

4. Agent Usage similarity

In the initial state, no information about tags and similarities exist, only the tag signed by the topic itself.

Input: Page P= {P}/ P is the historic of the pages registered when we open a session.
Output: Space S where
SpaceG=SDU
Page P= {P}/ P is the historic of the pages registered when we open a session.
For SDUC
Do
AgentContentSimilarity
Input: Page P= {P}
Output: Space S where
Space= SDUC
For SDUL
Do
AgentLinkSimilarity
Input: Page P= {P}
Output: Space S where
Space= SDUL
End

The Tags agents groups are started by the following types of agents tags:

- TagCG

CCT - Tag Content title: it tagged the information existing in the title and subtitles; C= T/min fw where T is the threshold of frequency of words
CCP - Tag Content paragraph: it tagged the information existing in the context with fw>T
ILC - Tag Content information Link: it tagged the information existing in the link information text with fw>T

- TagLG

TLW: it tagged the words which are hyperlinked to out-links
TLL: it tagged words existing in the out-links or the in-links
TLI: it tagged the information existing in the links information text with fw>T
TLP: it tagged the information existing in the hyperlinks over a paragraph.

- TagUG

TagAgentUsag(Agent/Group/Role)
<OpenSession>
Opp: get Usage space tag
Do
For space = DUS
Get TUS Auth
Get TUS inv
Get TUSc
Get TUSl
End

6.7. WebComp Agents Composition:

In the WebComp approach the agents are created in a dynamic way. We start by creating an agent called "generator agent" which controls the entire agent creation mechanism. The WebComp agents are sensitive to some notion, of the process order, where the similarity agent works before the collector agent, and the tagging agent works before the aggregation one. Activity of agents depends on web resources (pages in the spaces), where the higher the pages exist in the space; the more the activity happens in the space.

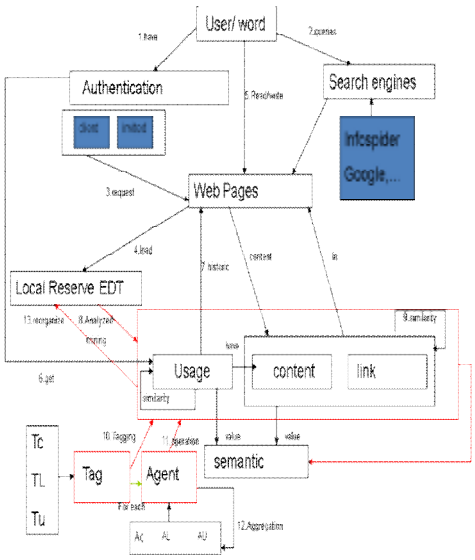

Figure 6: Webcomp architecture

6.8. ENFORCING THE SELF ORGANIZATION BY USING THE USAGE SPACE INFORMATION

Usage information has a great effect of recovering the relevant pages to the user's demand. Who is looking forward to finding what they are looking for Usage information includes two factors that could possibly affect the efficiency of page recover from the web. These factors are: First, using the same topic frequently through search engine process. Second, the time spent by users to browse the page. Furthermore, there are various factors that vary dynamically influencing the suitable data process, add, update, and delete. Of any page or site i.e. this change occurs in a non-linear manner.

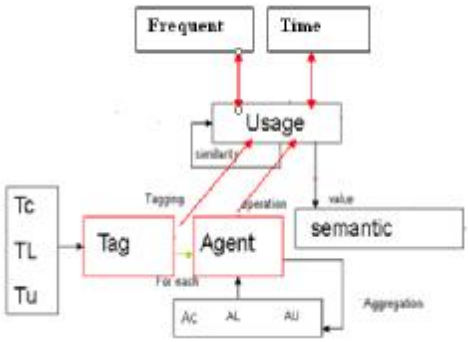

Figure7: Usage space

6.9. Definitions for Usage information:

- The Usage Space represents the information extracted from the session's history for each user invited or authorized.
- **Space** (Su) represent the environment where agent is living as a member in a set (Group(Gu)).the Web Pages represent the target for agent to get their information.
- **Usage Space**: Agent, Page, Operation, Tag, Time, Frequent. Where the *agents* are navigating and exploiting the entities searching for similar information to their tag. The agent defined as a set of (Group(G)) of agents that navigate in his space, search and collect specific information according to its speciality. *Page* is a set of target pages, Operation is the role that the agent will do, and, aggregation, *Tag* is the kind of information that illustrates each agent. *Time* is the time that the user spends in using Web Page. And *Frequent* is the summation of term that user used it.
- **Usage Agents**: UsPace, Operation, Tag, Life time cycle, Number of agents exist in agent's group, and type where the link, time, and frequent information represented types of information.
- **Role "operation"**: is an abstract representation of an Agent function service or identification within a group. Each agent can handle a role, and each role handled by an agent is local to a group.
- **Group**: a set of Agent Aggregation. Each agent is part of one or more group formed by three spaces (Content, Structure, and Link)

The Usage space is construct from those URLs that the user used concerning a topic, knowledge, and technique in exploiting the information while he is exploring the Web. Let U = {u1, u2... uw} be the set of pages in the user's historic and P = {p_1, p_2... p_v} the set of pages in the EDT where EDT is the set of all the pages collected by the search engines for a keyword. Using similarity rules between the user site U and the pages collected by the search engines P could be useful.

Here, the user is interested in finding some pages in P that are similar to a page in U. For example, the user reads an article on a particular topic and may want to know what has been published on similar sites on the same topic.

Our comparison is done as follows: Given a U page uj, we use the cosine measure to compute the similarity between u_j and each page in P. After the comparison, the pages in P are ranked according to their similarities in a descending order. Example: In this example, we have 4 pages in U and 3 pages in P, which are shown bellow (the number in each pair is the frequency of the keyword in the page).

U pages	P pages
Upage $_1$: (office, 1), (home, 1)	Ppage $_1$: (office, 2), (home, 2), (Web, 3)
Upage $_2$: (information, 2), (mining, 1), (office, 2)	Ppage$_2$: (association, 3), (mine, 2), (rule, 1)
Upage$_3$: (Web, 2), (probability, 2)	Ppage$_3$: (clustering, 3), (segment, 2), (office, 2)
Upage $_4$: (clustering, 2), (segment, 1)	

If we want to find the corresponding page(s) of Upage 1, we obtain the following ranking:
Rank 1: Ppage 1 Rank 2: Ppage 3

Ppage 2 is not shown in the ranking as its similarity value with Upage 1 is 0.

6.10. COMMUNICATION AGENTS AND AGGREGATION PROPERTY

By using the terms in the space for Content terms agents can communicate with each other. And it is straightforward to them to communicate if it is clear that they point at the same set of terms. If we link these spaces, the agents can commit the terms consistently with the Usage mandated in that space. But if they are not using the same space, they may still be able to communicate. Any agent can communicate with another if there is a common document space and if all mapping where perfect.

7. THE RESULTS OF WEBCOMP MODEL

In this section we will illustrate how our approach can reduce the complexity and reorganizing effort. That includes three steps: step1: Collecting Process and creating TDS. This is the first step of our algorithm. The URLs are stored in the TDS by using keywords as a topic table and are ready to get processed by the mining algorithms and their basic similarities. Step2 TDS is categorized into two sub spaces (Content, Link) according to it information type. Content Similarity calculated by TFIDF among all pair of pages. And Link similarity calculated using similarity rules as shown in table 1, Person Correlation coefficient applied over the results between the two majors for TDS. As shown in figure 8 At time t=0 just Content and Link spaces are formed. Step3: in t>2 user start using the system. As a result, Usage space is created and Usage agent's algorithms start working. As a result, we can find that when a page is added from the Usage space to the TDS, another is brought from the edge of TDS to the center. So, a correlation starts emerging.

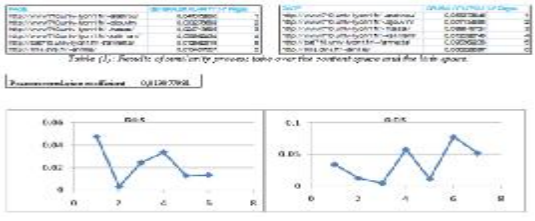

Figure 8: The general link and Content similarity distribution for the TDS at t=0.

Pages added to Usage space are compared if they found in TDS or not. If not added to TDS and recalculating and re- ranking is followed by it as shown in figure9. As a result a new TDS is created. In Usage space, Content and Link similarity is calculated in the same way as it is in TDS.Tabel2 illustrate that:-

If another page is added from Usage space to TDS, which is different from the first page, we can see that the correlation is becoming clearer figure10.

By adding another page at t=4. It is found that the system has transformed into a new form of aggregation and that correlation becomes higher and higher as shown in figure 11 and 12.

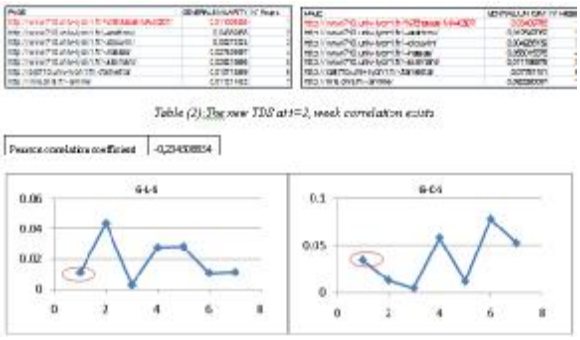

Figure 9: The general link and Content similarity distribution for the TDS at t=2, we can see that emergence of new changes in the distribution take place.

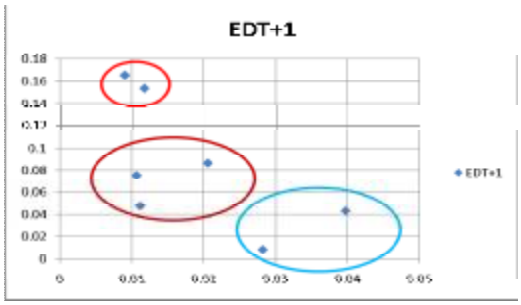

Figure10: The σc, σL , distribution for new TDS at t=2, good correlation emerges, and aggregation can happen.

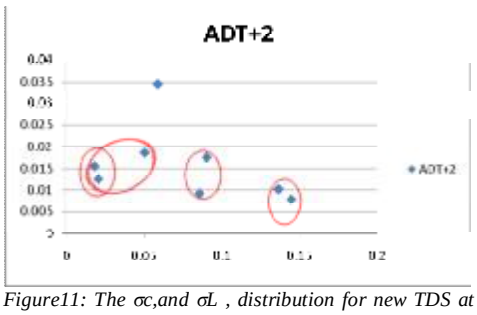

Figure11: The σc,and σL , distribution for new TDS at t=4, good correlation emerges, and aggregation can happen between three or more agents.

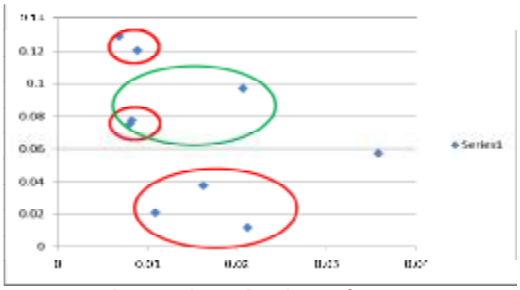

Figure 12: The σc,and σL , distribution for new TDS at t=6, good correlation emerges, and aggregation can happen between four groups of agents.

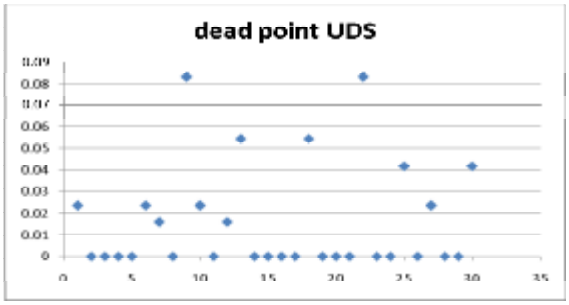

Figure13: Dead point in the UDS where the similarity value is equal to zero. These values isolate the concerned pages from the tagging collecting and aggregation later on

Finally: In this work we present a model that uses new web usage information to see the effects on the semantic values, and how it will help us to achieve a powerful and well self-organized Web.

8. CONCLUSION

In this paper we have presented an analysis of the Web as a complex adaptive system (CAS), and have proposed its modeling using the seven characteristics and mechanisms, which are proposed by J. Holland, to overcome its evolving complexity. It focused on the association of a semantic, to information contained on the web, through the combination of the web Content, Usage, Structure, Time. And Frequent information that Exist in Web Page and their interrelated multi-scale spaces are enriched by the Usage. The proposed models open a new approach in the web-searching domain using complex adaptive systems, properties and mechanisms especially Non- linear, and Multi-agent system paradigm in order to reduce the complexity of the complex web. The Usage Space is opened for researches and several scholars are investigating this matter.

9. REFERENCES

[1] J. Holland, *Hidden Order*: Addison-Wesley, 1995.

[2] S. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution*: Oxford University Press, 1993.

[3] W. B. Arthur, S. Durlauf, and 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," presented at British Academy of Management Conference, 1997.

[5] S. Levin, "Ecosystems and the Biosphere as Complex Adaptive Systems," *Ecosystems*, vol. 1, pp. 431-436, 1998.

[6] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee, "Self-Organization of the Web and Identification of Communities," *IEEE Computer*, vol. 35, pp. 66--71, 2002.

[7] J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," presented at ACM-SIAM Symposium on Discrete Algorithms, 1998.

[8] J. Liu, S. Zhang, and Y. Ye, "Understanding emergent Web regularities with information foraging agents," presented at Proceedings of the first international joint conference on Autonomous agents and multiagent systems, 2002.

[9] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks.," *Science*, vol. 298, pp. 824-827, 2002.

[10] M. Kirtland, "The Programmable Web: Web Services Provides Building Blocks for the Microsoft .NET Framework," in *MSDN Magazine*, vol. 15, 2000.

[11] S. Hassas, "Systèmes complexes à base de multi-agents situés," University Claude Bernard Lyon, 2003.

[12] S. Hassas, "Using swarm intelligence for dynamic Web Content organization," presented at IEEE Swarm Intelligence Symposium, Indianapolis, IN, USA, 2003.

[13] F. Menczer, A.E. Monge: Scalable Web Search by Adaptive Online Agents: An InfoSpiders Case Study . In M. Klusch, ed., Intelligent Information Agents , Springer, 1999

[14] F. Menczer, R.K. Belew: Adaptive Retrieval Agents: Internalizing Local Context and Scaling up to the Web .Machine Learning Journal 39 (2/3): 203-242, 2000.

[15] G. Pant, F. Menczer: MySpiders: Evolve your own intelligent Web crawlers .Autonomous Agents and Multi-Agent Systems 5(2): 221-229, 2002 Systems 5(2): 221-229, 2002.
[16] F. Menczer, G. Pant, P. Srinivasan: Topical Web Crawlers Evaluating Adaptive Algorithms, ACM TOIT 4(4): 378-419, 2004
[17] F. Menczer: Lexical and Semantic Clustering by Web Links .JASIST 55(14): 1261-1269, 2004.
[18] A.Rattrout, M.Rupert, S.Hassas," Reorganizing Dynamic Information Resources in the ComplexWeb,"ACIT2005, Jordanian. pp. 103-110. ISSN ISSN: 1812-0857. 2005.
[19] M. Rupert, A. Rattrout, S. Hassas, "The web from a complex adaptive systems perspective," (Journal of Computer and System Sciences) Journal ELSEVIER 2007
[20] M. Rupert, S. Hassas, A. Rattrout,"The Web and Complex Adaptive Systems," AINA (2) 2006: 200-204 IEEE.