A New Competitive Intelligence-based Strategy for Web Page Search

Iman Rasekh
Institute of Computer Science
University of Philippines at Los-Banos, Los-Banos, Laguna
Philippines
iman.rasekh@gmail.com

Abstract: *Semantic Web is known as next generation of web it is known as a new collaborative movement toward Web3.0 that led by the World Wide Web Consortium (W3C). It aims at converting the current web of unstructured documents into a "web of data". The proposed searching strategy for SEO in Semantic Web is a graph structured search (GSS). Search Engine Optimization (SEO) is defined as a collection of techniques and practices that allow a site to get more traffic from search engines and it is still one of the biggest challenge in search engines of Semantic Webs. In this paper, I proposed a new type of web page search which is based on the competitive intelligence. It use link-based ranking evolutionary scheme to accommodate users' preferences. I implemented the prototype system and demonstrate the feasibility of the proposed web page search scheme.*

Keywords: Linked based ICA Algorithm, Linked based Page Ranking, ICA, Folksonomy, Semantic webs

Received: 19 August 2014, Revised 25 September 2014, Accepted 29 September 2014

1. Introduction

Due to the huge number of web pages that exists in World Wide Web; *analyzing and clustering of the results* is still the most important challenges in design of search engines. Nowadays many web page search techniques have been developed and search engines became more complicated to process a large number of web pages. There are still some problems unresolved with the search engines. In terms of search quality, about one half of all retrieved web pages have been reported to be irrelevant. Even though the search systems can acquire large amount of web pages reflecting users' preference from Internet, it is still unsatisfactory to analyze and cluster them because of the huge number of web pages .To obtain better search results from massive web pages on the Internet, I propose a prototype linked-based search system based on Imperialist Competitive Algorithm and folksonomy strategy. Imperialist Competitive Algorithm (ICA) is a new socio-politically motivated global search strategy that has recently been introduced for dealing with different optimization tasks. Folksonomy is a new classification technique which attach tags or labels to each web page to suffice the practice and method of categorizing contents.

The proposed system implement as a linked based system based on page ranking algorithm.

PageRank calculates the probability that someone randomly clicking on links will arrive at a certain page and an architecture is proposed for the system.

The rest of the paper is organized as follows: Section 2 discusses the meaning of search engine optimization in Semantic Web in this section Page Rank is introduced and dynamic tree based folksonomy structure is discussed. Section 3 introduces the Imperialist Competitive Algorithm that was used. Section 4 is talking about my proposed architecture and redefined ICO algorithm and finally, the implementation of my proposed system described in section 5.

2. Search Engine Optimization In Semantic Web (SEO)

The Semantic Web is a major research initiative of the World Wide Web Consortium to create a metadata-rich web of resources that can describe themselves not only by how they should be displayed or syntactically , but also by the meaning of the metadata[1]. The main purpose of the Semantic Web is driving the evolution of the current Web by enabling users to find, share, and combine information more easily. A machines cannot accomplish search tasks without human direction, because web pages are designed to be read by people, not machines. The semantic web is a vision of information that can be readily interpreted by machines, so machines can perform more of the tedious work involved in finding, combining, and acting upon information on the web [2].

Search Engine Optimization (SEO) is a fundamental concept in Semantic Webs and refers to the collection of techniques and practices that allow a site to get more traffic from search engines1. The search engines usually are smart enough to award you that rank by default. Search engines have become more and more popular on the web, nearly anyone trying to get seen on the web can benefit from a little SEO loving [3].

Recently, folksonomy and ranking strategy based on links are getting popular in Semantic Web Search Engine Optimization and spreading widely. Folksonomy is one of the components of Web 2.0 which is extended to Semantic Web page ranking is also an important approach in web search strategy.

2.1 Page Rank - PR (E)

Page Rank is an algorithm to rank websites in their search engine results. It works by counting the number and quality of links to a page to determine how important the website is. More important websites receive more links from other websites. The numerical weight that it assigns to any given element E is referred to as the Page Rank of E and denoted by PR (E). The Page Rank of a page is the probability of arriving at that page after a large number of clicks. This happens to equal t-1 where t is the expectation of the number of clicks. One main disadvantage of Page Rank is that it favors older pages. A new page, even a very good one, will not have many links unless it is part of an existing site [4].

$$\text{Page Rank} \propto 1/(\text{The Number of clicks}) \qquad (1)$$

The Page Rank assume a probability distribution between 0 and 1 as the initial value for each page. The Page Rank value for a page *u* is dependent on the Page Rank values for each page *v* contained in the set *Bu* (the set containing all pages linking to page *u*), divided by the number *L*(*v*) of links from page *v*. In other words, the Page Rank is equal to the document's own Page Rank score divided by the number of outbound links *L*.

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)} \qquad (2)$$

Damping factor is defined as the probability, at any step, that the person will continue is a damping factor *d*. The damping factor will be set around 0.85[5]. The damping factor is subtracted from 1 and this term is then added to the product of the damping factor and the sum of the incoming Page Rank scores. The result is divided by the number of documents (*N*) that is:

$$PR(u) = \frac{1-d}{n} + d \sum_{v \in B_u} \frac{PR(v)}{L(v)} \qquad (3)$$

To more simplify it we can write the PageRank of the web page *i* as

$$PR_i = \frac{(1-d)}{n} + d \sum_{j=1}^{n} \frac{PR_j}{L_{ji}} \qquad (4)$$

In which L_{ji} is the number of out links from page *j* when page *j* is linked to page *i*.. *I* noticed that *I* should consider the historical behavior of all users. I calculate L_{ji} by divide the number of clicks from page *j* to page *i* when *j* is linked to *i*.

$$C_{ji} = \sum u \in U \; C_{ji}(u) \tag{5}$$

$$L_{ji} = \frac{\sum_{j=1}^{n} C_{ji}}{C_{ji}} \tag{6}$$

In which $C_{ji}(u)$ is the number of clicks of a user u from page j to i, while U is a set of all users.

To consider *User's Preference* to a web page and support personalized search. It is necessary to count the number of a user's clicks in all pages that have linked to page i over the number of a user's clicks in web pages that have linked to all the pages. Equation (7) describes its definition.

$$PE_i(u) = \frac{\sum_{j=1}^{n} C_{ji}(u)}{\sum_{j=1}^{n} \sum_{j=1}^{n} C_{ji}(u)} \tag{7}$$

Finally, Equation (8) is my modified PageRank algorithm for the proposed system[6].

$$PR_i(u) = PR(u_i) = (1-d)\, PE_i + d \sum_{j=1}^{n} \frac{PR_j}{L_{ji}} \tag{8}$$

2.2 Semantic Web Folksonomy strategy

Folksonomy is a new classification technique which collaboratively creates and manages tags to categorize contents. Folksonomy may hold the key to developing a Semantic Web, in which every Web page contains machine-readable metadata that describes its content [7]. Such metadata would dramatically improve the precision (the percentage of relevant documents) in search engine retrieval lists [8]. It attach tags or labels to each web page to suffice the practice and method of categorizing contents. "*Tags*" are keywords that allotted by users to each page freely and subjectively, based on their meaning. Anyone can choose any word as tag and can put multiple tags to one page. Figure1 shows an example of tagging web pages by different users. We define the 'frequency' of a tag for a page as the number of users who used the tag for the page, and determine the category of a page using the frequency of each tag assigned to the page: the tag with the largest frequency becomes the category of the page. The most important advantage of folksonomy is that users can quickly search and easily classify related web pages. It is well-known that folksonomy provides a flat, non- hierarchical and shared terminology. Figure 1 shows the folksonomy tagging graph.

3. Competitive Intelligence

Imperialist Competitive Algorithm (ICA) is a socio-politically motivated global search strategy that has been introduced for dealing with different optimization tasks. This algorithm starts with an initial population. Each individual of the population is called a country. *Country* is an array that defines with the same meaning of *chromosome* in the GA terminology:

$$\text{Country} = (p_1, p_2, p_3, \ldots, p_{Nvar}) \tag{9}$$

The cost of a country is found by evaluation of the cost function f which is the same with metric function in *GA*,

$$cost = f(country) = f(p_1, p_2, p_3, \ldots, p_{Nvar}) \tag{10}$$

Some of the best countries are selected to be the imperialist states and the rest form the colonies of these imperialists. All the colonies of initial countries are divided among the mentioned imperialists based on their power. The imperialist states together with their colonies form empires.

After forming initial empires, the colonies in each of them start moving toward their relevant imperialist country (*Assimilation policy*). Assimilation policy is modelled by moving all the colonies toward the imperialist.

The total power of an empire depends on both the power of the imperialist country and the power of its colonies. This fact is modelled by defining the total power of an empire as the power of imperialist country plus a percentage of mean power of its colonies [9]. Figure 2 shows the levels of ICA algorithm.

Revolution can be define as a sudden change in socio-political characteristics of a colony that is, instead of being assimilated by an imperialist, the colony randomly changes its position in the socio-political axis.

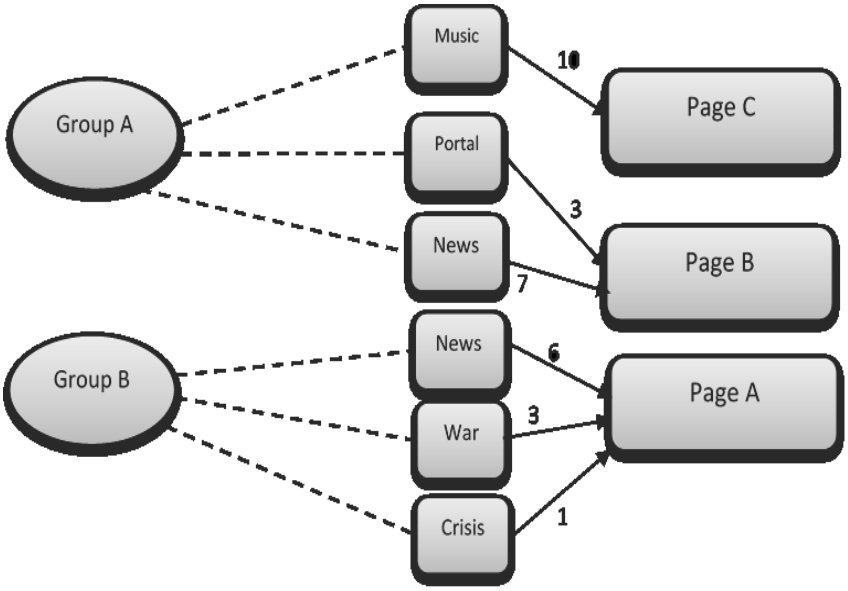

Figure 1. Folksonomy Tagging graph

While moving toward the imperialist, a colony might reach to a position with lower cost than the imperialist, *Exchanging Positions of the Imperialist and a Colony* is happen. Then the algorithm will continue by the imperialist in the new position and the colonies will be assimilated by the imperialist in its new position.

In the movement of colonies and imperialists toward the global minimum of the problem some imperialists might move to similar positions. If the distance between two imperialists becomes less than threshold distance, they will *Unite* and make a new empire which is a combination of former empires. All the colonies of two empires become the colonies of the new empire and the new imperialist will be in the position of one of the two imperialists[10].

Imperialistic competition is a process that brings a decrease in the power of weaker empires and an increase in the power of more powerful ones all empires try to take the possession of colonies of other empires and control them. It is modelled by just picking some of the weakest colonies of the weakest empire and making a competition among all empires to possess these colonies[9].

4. Proposed Architecture And Redefined Ica

Search Engine Optimization (SEO) refers to the collection of techniques and practices that allow a site to get more traffic from search engines. To improve the SEO using ICA; at First we need a strong web mining algorithm to determine the anchor nodes because we can't consider each separate word as an anchor, sometimes an anchor is made of some separate words for example for the phrase "*World Wide Web Consortium*" can be divided into only two phrases "*World Wide Web* "and "Consortium" because "*World Wide Web*" is a known phrase Figure 3 shows that how can it possible.

4.1 Overall Architecture (Protocol)

Introducing the most relevant web pages to users is my primary concern. I designed a new web page search system based on Imperialistic Competitive Algorithm combines folksonomy and link-based ranking strategy. Figure 4. shows the overall architecture of my proposed system. Both ends of the architecture are management and resource layers.

1. **Empire Initialization Layer:** Includes folksonomy database and page rank database. They store relevant information come from management layer.

2. **Data Bus Layer:** Represent mass volume of web page data.

3. **Application Layer:** Consist of search engine and QA engine. QA Engine is a computer program that can pull answers from an unstructured collection of natural language documents. This layer is responsible for processing the request of users and

returning the search result.

4. **Management Layer:** Includes ICA manager , ICA manager is used to analyze and classify mass data using ICA algorithm , all parts of Imperialistic competitive algorithm (Except for initialization of empires; which is done in Empire initialization layer) should be done in this layer.

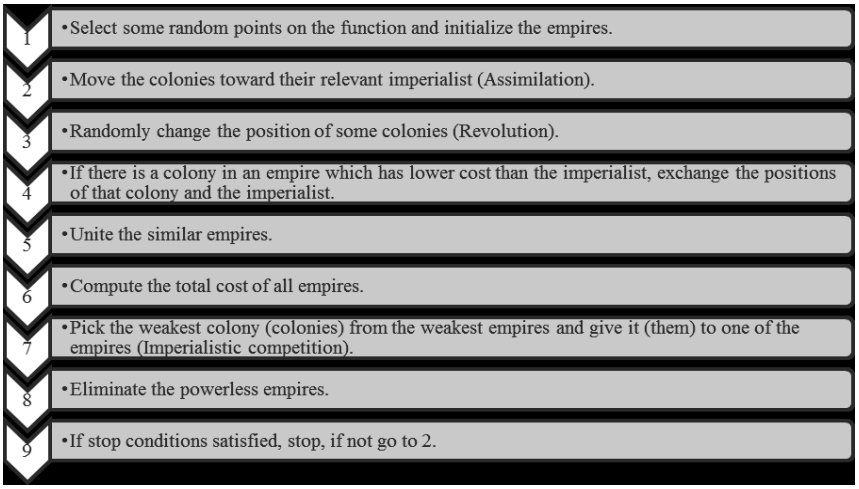

Figure 2. Steps of the Imperialist Competitive Algorithm

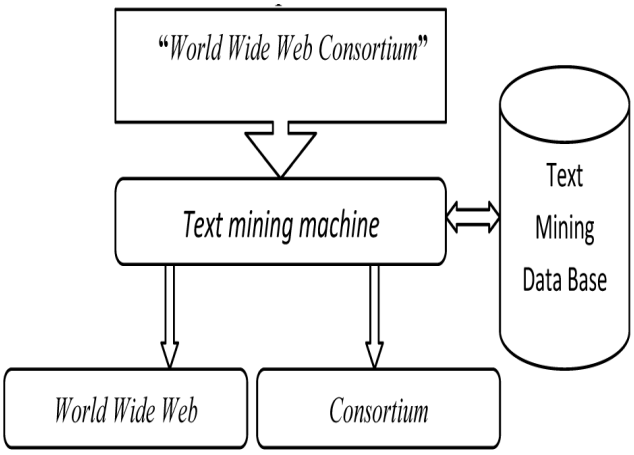

Figure 3. Text-Mining Approaches for World Wide Web Consortium

4.2 System Implementation Model

Now it's time to implement the system. My proposed model (as is shown in Figure 5 includes four layers:

1. **Transaction Analysis layer:** Collect the information from users, analyze them and transmit the result of analysis to resources management layer and it can display the search results for users.

2. **Resources management layer:** Manage and store the information about searching process.

3. **Transaction process layer:** Is responsible for deeply analyzing the requests of users and classifying them by the behaviors and interests of users.

4. **Semantic layer:** A Semantic classifier than can act as a heuristic classifier that implement ICA algorithm.

The details for system implementation would be as follows:

1. Users enter the search keywords as the tags of web pages.

2. Stored information in the storage server (Resources Management layer)

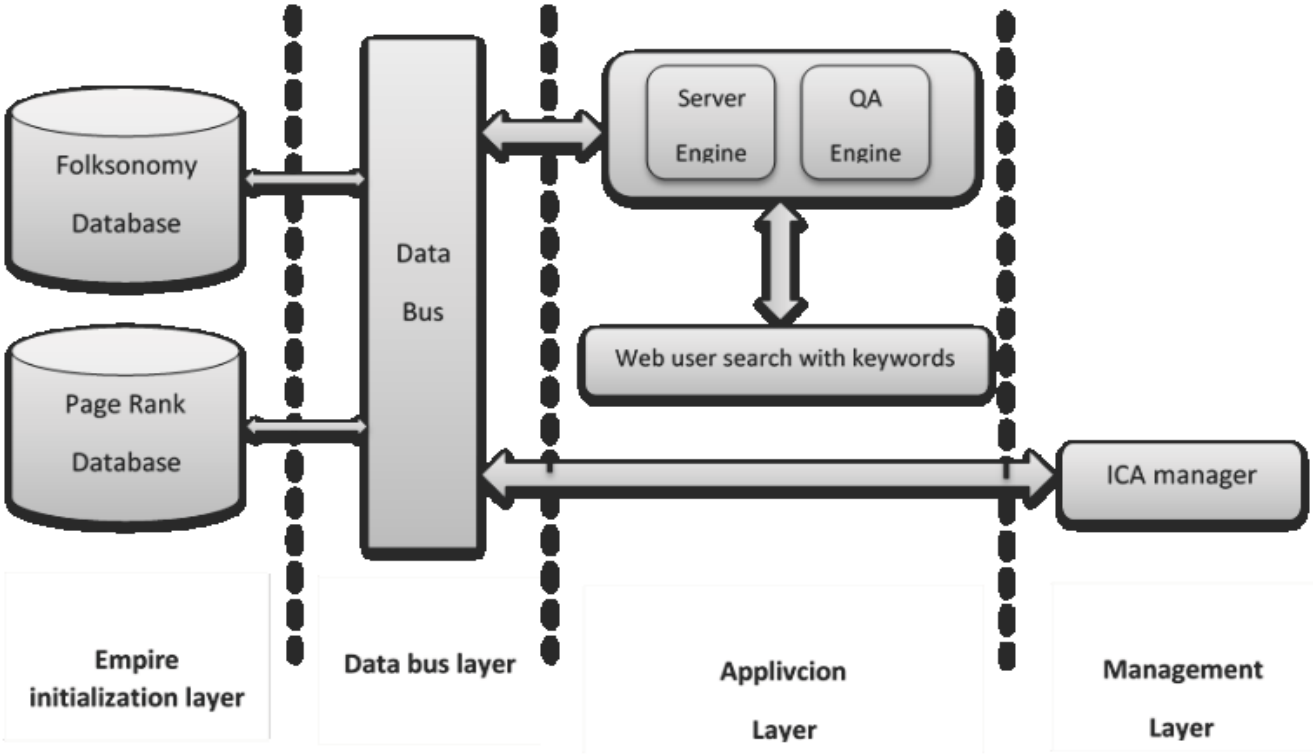

Figure 4. ICA_based Search Engine architecture

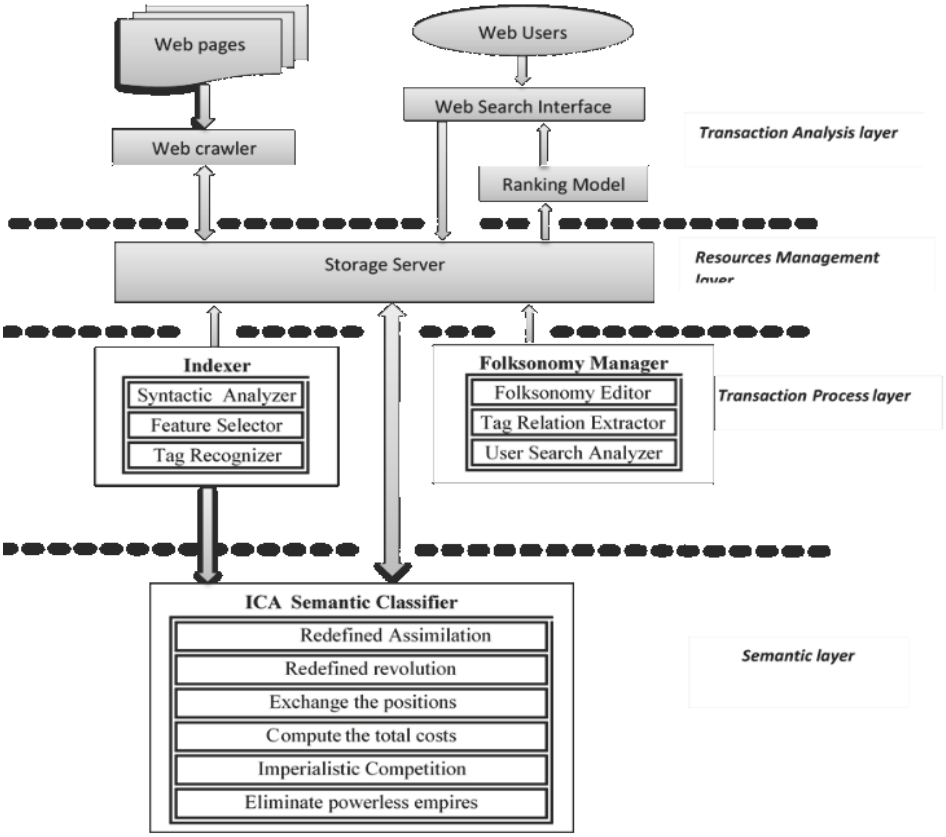

Figure 5. System implementation model

3. Transmitted user tags information to the Transaction process layer.

4. Folksonomy manager do the following processes

a. Edit the tags

b. Extract the relationships among the tags

c. Analyze users' behavior.

5. Indexer module classify web pages based on tags (Transaction Process Layer).

6. The output of indexing delivered to the ICA semantic classifier to do the semantic search .Web pages which are labeled by the same tags belong to the same category.

7. Sorted web pages with the same category by using the link-based ICA algorithm

8. Finally, the results are displayed to the users.

4.3 Redefined Initialization of Empires

To make the countries at first we should assign a numerical weighting to each element of a hyperlinked with the purpose of "*measuring*" its relative importance it can be computed by using Folksonomy strategy which is discussed in section 2 part A. Consider all pages as nodes and hyperlinks as edges, Compute the Page Rank of each page recursively and based on incoming links. A page that is linked to by many pages with high Page Rank receives a high rank itself. A Country can be defined as a randomly arranged integer sequence (1 to n) in which each integer represents the page rank of a node in a local knowledge graph. To define the cost of a country at first we should define the cost of each node of the local knowledge graph so need to use folksonomy strategy to define that. In our proposed system, folksonomy is used to classify the searched pages by analyzing tags and user's behavior. The algorithm can be define as follows; Record tags that entered by user and capture user' behavior by counting clicking web pages. In the second step tag trees will be formed by connecting the tag (Ti) and page (Pi) in a tree, third the system will calculate the number of users that use the same tags. And at last the tag of the highest frequency which is calculated by the loop program will be inserted into tag database as the final category of the web page. If the categories of these web pages are the same, these web pages belong to a category.

4.4 Redefine the cost of countries and imperialists based on page ranking algorithm

Based on *Imperialistic Competitive Algorithm* we know that the normalized cost of an imperialist is defined by

$$Cost_n = Cost_n - max_i \{c_i\} \tag{11}$$

And the normalized power of each imperialist is defined by

$$Power_n = \left| \frac{Cost_n}{\sum_{j=1}^{N\ imp} Cost_i} \right| \tag{12}$$

From section 2 we know that PageRank score for a web page *i* is calculated as

$$PR_i = PR(u_i) = (1-d)\, PE_i + d \quad \sum_{j=1}^{n} \frac{PR_j}{L_{ji}} \tag{13}$$

In which V is the set of nodes in the local knowledge graph, hence the country of k^{th} country can be redefined as:

$$cost_k = f(country_k) = f(PR_1, PR_2, PR_3, ..., PR_n) \tag{14}$$

The metric function for cost; we can define it as the maximum of costs of the nodes in knowledge graph and conclusively empire will be the country with the highest cost.

4.5 Redefined Assimilation algorithm using Random Substitution

During learning process the colonies obtain information from the imperialist and adjust themselves to keep consistent with the relevant imperialist. Redefined assimilation can be implement using *Random Substitution* algorithm; at first a subsequence is randomly chosen from the relevant imperialist, and a position is randomly chosen from the colony. In the next step; the mentioned subsequence is inserted to the mentioned position; at last the imperialist which are included in the subsequence

are deleted from the part coming from the previous colony. For example (5, 1, 7, 9, 3, 2) is the cost of an imperialistic country of a knowledge graph of 5 nodes the subsequence (1, 7, 9) is chosen from the imperialist. The position between city 5 and city 3 is chosen from the colony. After the assimilation process: (1, 7, 9) transferred from the imperialist to its colony.

4.6 The Modified Revolution Process based on Random Crawling Strategy Modified by 2-Opt Algorithm

The Revolution Process replace the randomly selected colonies with an equal number of new randomly generated countries. To optimize and modify the revolution process I used a combination of Random Crawler and 2opt algorithms, Random crawler is a simple random algorithm for scanning a knowledge graph in semantic webs and 2-opt algorithm is a local search approach. The proposed method can be implemented as follows:

1. Model the network of knowledge graphs as a Markov chain1 in which the states (nodes) are knowledge graphs, and the transitions are the links between them.

2. A node with no links to other nodes (sink), terminates the random crawling process. If the random crawler arrives at a sink page, it picks another URL at random and continues crawling again.

3. If two local knowledge graphs cross, use the 2 opt algorithm to find the shortest path between knowledge graphs.

As an example, consider the network (A-B-F-E-C-D-H-I-G-A) a network of knowledge graphs in a semantic web. In the first step Links A-B and C-D are selected, then a new network is generated by linking A and C, B and D and finally if the new network (A-C-E-F-B-D-H-I-G-A) has better cost then, the new network is accepted.

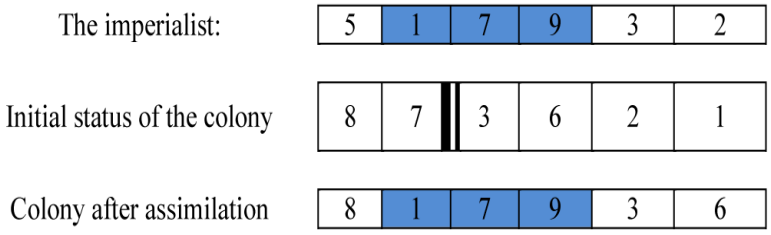

Figure 6. Random Substituting strategy in assimilation process

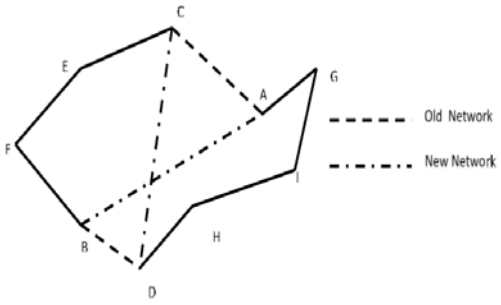

Figure 7. Random crawling in revolution step

4.7 The Modified ICA algorithm for competitive search

Based on what have discussed above my revised ICA algorithm for semantic search summarized as follows:

5. Implementation and Experiments

Tagging pages is the first step to design the system. To implement this I defined 300 pages and tagged 100 pages (pages 201-300) based on some common tags like: news, groups, social networks Iran, Philippines, Japan, portal, forums.

To simplify the computation I categorized all of them with the tag 'portal' (as the most tagged label)

1. **Redefined Initialization of empires**: Generate population based on the local knowledge graphs of Semantic Webs. Computes the cost of empires using Page Rank algorithm.
2. **Redefined Assimilation:** Move the colonies toward their relevant imperialist using *Random Substitution*
3. **Redefined Revolution:** Select out a part of colonies and apply the Random crawling algorithm ; modify the results by using 2-OPT algorithm on them
4. **Exchange the positions** : If a colony is better than the relevant imperialist, exchange the role of the colony and the imperialist
5. **Compute the total cost of all empires.**
6. **Imperialistic competition:** Pick the weakest colony (colonies) from the weakest empires and give it (them) to one of the empires.
7. **Eliminate the powerless empires:** *If an empire losses all colonies*
8. If only one empire is existed stop, if not go to 2.

Figure 8. Redefined ICA Algorithm

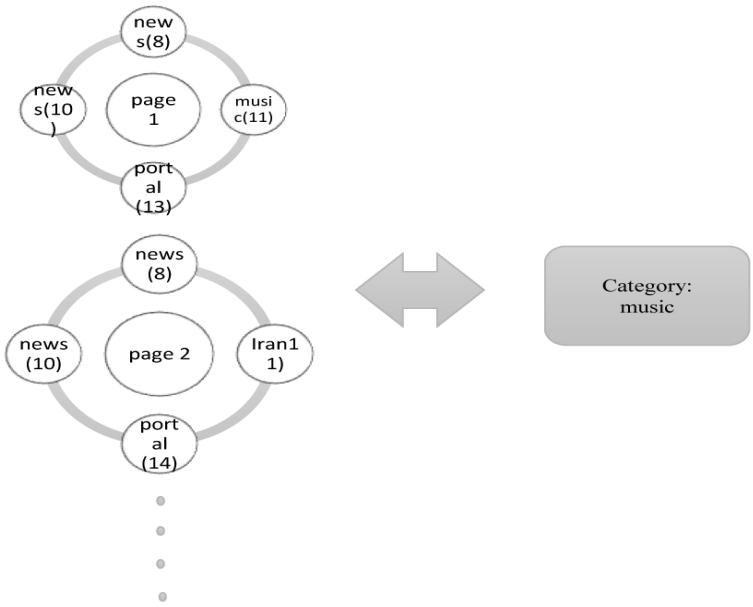

Figure 9. Categorizing pages based on Folksonomy strategy

I considered default Page Ranks (A Random number between 0.1 to 0.2) for the 10 pages (P_{201} to P_{300}) .For the first 200 pages (P1 to P200) I considered link relationship with the next 10 pages, pages P1 to P200 are involved in the calculation. we can get the modified PageRank score of pages P_1 to P_{200} by using modified PageRank algorithm. For computing PR_i *I* computed $\frac{PR_j}{L_{ji}}$ for each $1 \le j \le 200$ and $201 \le i \le 300$ the results for the first 20 pages are shown in the Table 1.

In this step we should group a set of web pages in such a way that pages in the same group would be more similar to each other than to those in other groups. *I* used the imperialistic competitive algorithm to grouping the pages with the same page rank. The point is the countries in each group. At first we should find the pages with the most known and most searched tags and the highest default Page Ranks and consider them as the center of the semantic knowledge graph.

For grouping the page ranks and find the best page Rank for each page the best method is to minimize sum of distances from the center of each knowledge graph If we consider di as the distance of each node from the central node, then we should minimize *D* that define as follows:

$$D = \sum_{i=1}^{n} di \quad (15)$$

In which d_i is the distance between $\overrightarrow{x_i}$ and the closest central point (c_1, c_2,…c_n) will be the same with:

$$d_i = \| x_i - c_j \|_2 \quad (16)$$

We need to minimize all the distances in a knowledge graph so the formula that we should compute is like this:

$$min \sum_{\forall i} min \sum_{\forall j} \| x_i - c_j \| \quad (17)$$

Page	Page Rank	Page	Page Rank
P1	0.08599	P11	0.08991
P2	0.09998	P12	0.09590
P3	0.09321	P13	0.09498
P4	0.09775	P14	0.06693
P5	0.08898	P15	0.09597
P6	0.07598	P16	0.07798
P7	0.09004	P17	0.09596
P8	0.09590	P18	0.08898
P9	0.09693	P19	0.09594
P10	0.09448	P20	0.09397

Table 1. Pagerank Calculation For Pages Between P1 To P20

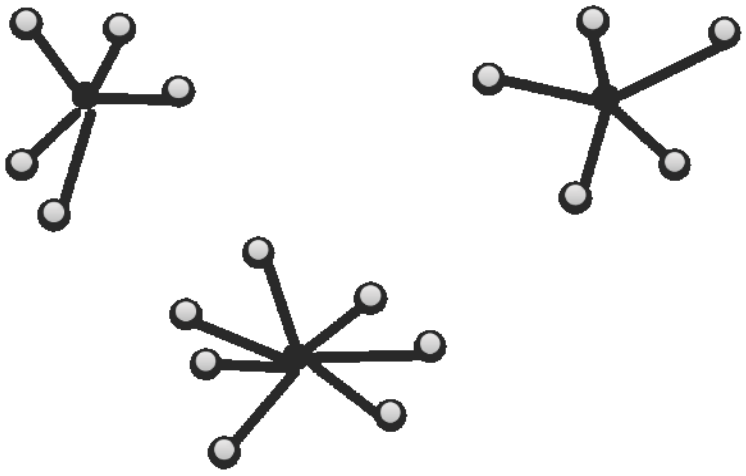

Figure 10. Grouping of the page ranks based on the most default page rank in the knowledge graph

I divided space of Page Ranks into *three knowledge* graphs. The plot is showed in Figure 11. to define this at first the program starts with random central point of knowledge graphs (The maximum Page Rank in each group) In each iteration the program choose some Page Ranks which are close to the central page rank and group them as a knowledge graph, in each step the knowledge graphs change and finally the best page rank foe each group will be selected. Figure 11 shows different groups of Page Ranks per different iterations, I repeated the ICA algorithm for 100 times and Figure 12 shows the diagram of iteration.

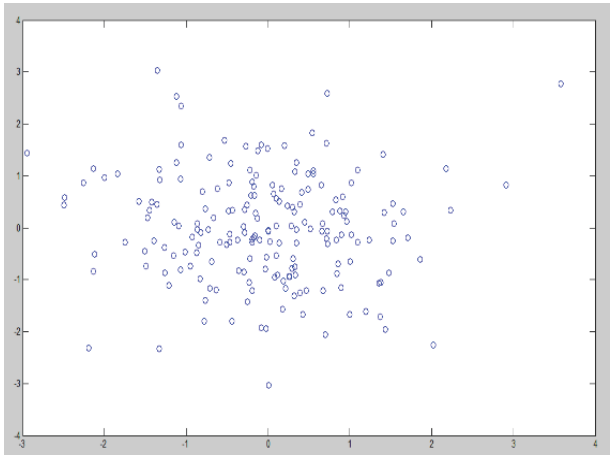

Figure 11. The space of Random Page Ranks

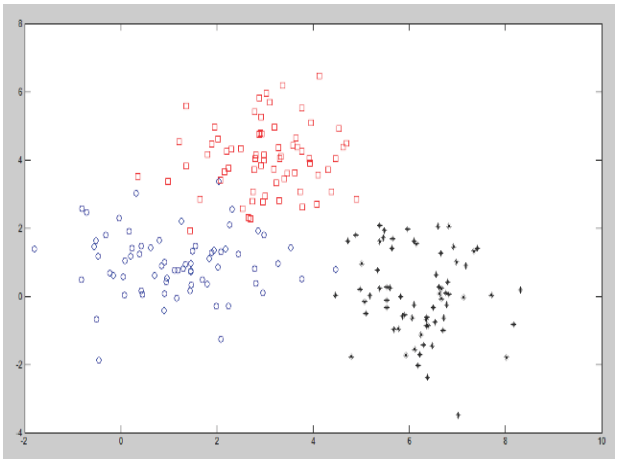

Figure 12. Dividing the space of Page Ranks into three knowledge graph

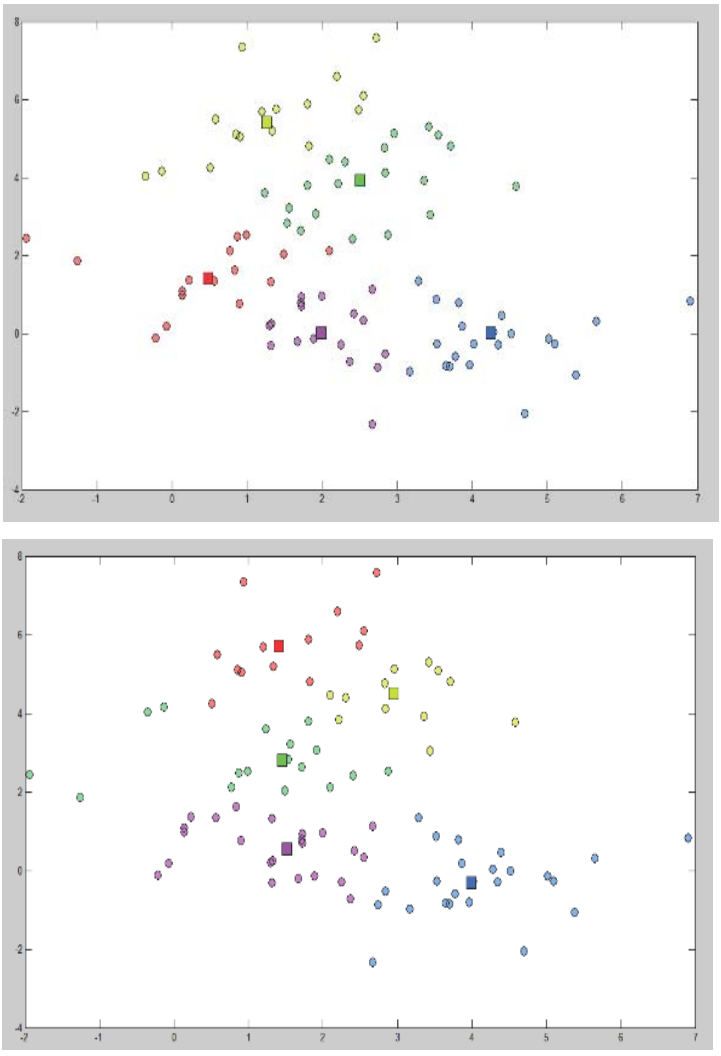

Figure 13. Different groups of Page Ranks per different iterations

The following diagram shows the best cost in each of 100 iterations:

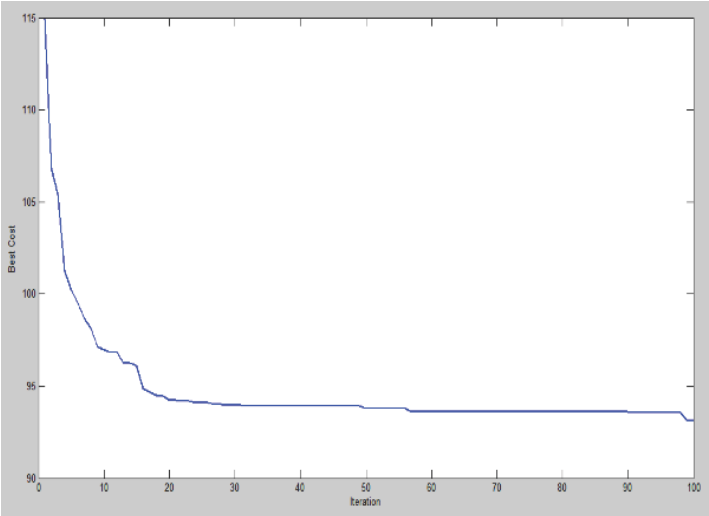

Figure 14. The diagram of costs per iteration

6. Conclusion

In this paper, I proposed a novel search system based on competitive intelligence that implement a high quality web search engines. The proposed system combines ICA algorithm and link based ranking scheme. My goal was to redefines the assimilation and revolution so that they can be compatible with large scale search in semantic webs. "*Enhancing the performance*" of searching and Applying it on larger-scale data and combinational optimizations was my other goals in this research . Future work is needed to make the system fit to the real World Wild Web. A new page ranking algorithm can be defined by mixing ICA and other heuristic algorithms that would be able to consider the location of pages and the location of users, the new page ranking algorithm should be able to greatly increase the quality of the search result.

References

[1] Manola, F., Miller, E. (2003). (eds). RDF Primer. W3C Working Draft 23 January 2003.

[2] Berners-Lee, Tim., Hendler, James., Lassila,Ora. (May 17, 2001). The Semantic Web, *Scientific American Magazine*. Retrieved March 26, 2008.

[3] Kevin, Gibbons. (2014). Do, Know, Go: How to Create Content at Each Stage of the Buying Cycle. *Search Engine Watch*. Retrieved 24 May.

[4] Moore, Ross. (2014). Connectivity servers. Cambridge University Press. Retrieved 24 May.

[5] Gyöngyi, Zoltán., Berkhin, Pavel., Garcia-Molina, Hector., Pedersen, Jan. (2006). Link spam detection based on mass estimation, *In*: Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB '06, Seoul, Korea), pp. 439–450.

[6] Page, Larry., (1997). PageRank: Bringing Order to the Web, at the Wayback Machine (archived May 6, 2002), Stanford Digital Library Project, talk. August 18, (archived 2002) .

[7] Fields, Kenneth. (2007). Ontologies, categories, folksonomies: an organised language of sound. Cambridge.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] Mohamed, Khaled, A. F. (2006). The impact of metadata in web resources discovering.

[9] Atashpaz-Gargari, E., Lucas, C. (2007). Imperialist Competitive Algorithm: An algorithm for optimization inspired by imperialistic competition, *IEEE Congress on Evolutionary Computation*, 4661–4667.

[10] Biabangard-Oskouyi., Atashpaz-Gargari, E., Soltani, N., Lucas, C. (2008). Application of Imperialist Competitive Algorithm for materials property characterization from sharp indentation test. To be appeared in the *International Journal of Engineering Simulation* (In Print)