Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2015) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

The 2015 International Conference on Soft Computing and Software Engineering (SCSE 2015)

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

Abstract

Search Engine Optimization (SEO) is a collection of techniques that allow a site to get more traffic from search engines. Page Ranking is the fundamental concept of SEO and defines as a weighted number that represent the relative importance of the page based on the number of inbound and outbound links. 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.

Peer-review under responsibility of organizing committee of The 2015 International Conference on Soft Computing and Software Engineering (SCSE 2015).

Keywords: Linked based ICA Algorithm, linked based page ranking, ICA, folksonomy, semantic webs

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 challenge in design of search engines and still more than half of all retrieved web pages in any search engine have been reported to be irrelevant. So many issues should be considered to design an efficient WBIR[†] and I listed the most important factors as follows: first of all, the word ambiguity should be considered where a single word can take on multiple meanings and the typographical errors contained within web information should be found. Secondly a WBIR system should cover different types of media, search applications and tasks. Last and foremost the

* Corresponding author. Tel.: +63-999-732-2070; .
E-mail address: iman.rasekh@gmail.com

[†] Web Based Information Retrieval; defined as searching for relevant documents or information among the large

Peer-review under responsibility of organizing committee of The 2015 International Conference on Soft Computing and Software Engineering (SCSE 2015).

feedback given by the information retrieval system should be evaluated but retrieving too much information is not necessary [1]. 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 SEO 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 Webs

Search Engine Optimization (SEO) is a fundamental concept in Semantic Webs and refers to the collection of techniques to make websites appear in the search engine's results pages (SERPS). Each page has a default Page Rank which is specified by the search engine [2].

2.1. Page Rank - PR (E)

Websites which are more important should receive more links from other websites .Page Rank is the algorithm developed to rank websites in their search engine results [5].Each page has a predefined default Page Rank as the initial value for each page. The current page rank is defined based on the *binary link variable* which is defined as follows:

$$L_{ij} = \begin{cases} 1 & \text{If page j points to page i} \\ 0 & o.w \end{cases} \tag{1}$$

In the nest step the *total number of pages (c_j)* is computed based on the L_{ij}

$$c_j = \sum_{i=1}^{N} L_{ij} \tag{2}$$

And finally *the recursive page rank formula* is defined as follows

$$PR_i = (1-d) + d\sum_{i=1}^{N} (\frac{L_{ij}}{c_j}) PR_j \tag{3}$$

I which d *(damping factor)* is the probability, at any step, that the person will continue (mostly 0.85) and p_j is the

iInitial values of page rank[7] .

2.2. Semantic Web Folksonomy strategy

Folksonomy is a new classification technique in Semantic Web which creates and manages tags to categorize contents, in which every Web page contains machine-readable metadata that describes its content [6] it helps users to do the search quickly and easily classify related web pages. It provides a flat, non- hierarchical and shared terminology for the search engines. 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. Tag can be chosen by both users and programmer and it is possible to put multiple tags to one page. The tag with the largest frequency is chosen as the category of the page [7].

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. Like the other evolutionary algorithms this algorithm also starts with an initial population which is called a country. Some of the best countries are selected to be the imperialist states and the rest form the colonies which are divided among 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*). *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 [8]. During the *Revolution events*, the colony randomly changes its position in the socio-political axis. 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. 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. *Imperialistic competition* which is the most important part of the 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. Related Works

Weighted Page Rank algorithm (WPR) [10] is a modified Page Rank Algorithm based on use *Web Structure Mining*. In this algorithm every out-link page is given a primitive rank value and decides the rank score based on the popularity of the pages. *WLRank algorithm* [11] provides weight value to the links based on three parameters; Length of the anchor text; Tag of the link and *relative position* of the page which reveal that physical position does not always in synchronism with logical position is not so result oriented. *HITS* [12] which is one of the oldest official Page Ranking algorithm; divides pages into two categories, *Authority pages*; which are the page which is pointed by many hyperlinks and *HUBs* that points to various hyperlinks. In this algorithm, ranking of the web page is decided by analyzing their textual contents against a given query. *Modified HITS* (PHITS) is a modification of HITS which provides a weight value to every link depending on the terms of queries and endpoints of the link [13] a probabilistic explanation of relationship of term document is provided by PHITS. TagRank Algorithm (TR) [14] is the most common Web Content Mining algorithm for page ranking; this algorithm is a comparison based approach and based on social annotations which calculates the heat of the tags by using time factor of the new data source tag and the annotations behavior of the web users. In Time Rank algorithm (TIR) [15] the default page rank is calculated based on the visit time of the page and visiting time considered as a factor that shows the degree of importance to the users. Finally the visiting time is added to the computational score of the original page rank of that page .EigenRumor Algorithm (ER)[16] is proposed for ranking the blogs. This algorithm provides a rank score to every blog by weighting the scores of the hub and authority of the bloggers depending on the calculation of eigen vector. Relation based algorithm [17] which is known as the most accurate page ranking algorithm among those that use Web Content Mining proposes a relation based page rank algorithm for semantic web search engine that depends on information extracted from the queries of the users and annotated resources. Query Dependent Page Ranking (QDR) [18] is a powerful semantic search engine that take into account keywords and return page only if both keywords are present within the page and they are related to the associated concept as described in to the relational note associated with each page.In Distance Ranking Algorithm (DRA) [19]; ranking is done base on the shortest logarithmic distance between two pages.

5. Proposed Architecture and Redefined ICA

SEO architecture can be improved by using ICA; at First a strong web mining algorithm is needed to determine the anchor nodes; each separate word as an anchor cannot be considered as a separate anchor node, sometimes an anchor is a combination of separate words. I introduced the architecture in the next section

5.1. Overall Architecture (Protocol)

Introducing the most relevant web pages to users is my primary concern. Figure.1. shows the overall architecture of my proposed system. First layer is *Empire initialization layer* which includes folksonomy and page rank databases. They store relevant information come from management layer. Second layer which is called *Data bus layer*; represents mass volume of web page data. *Application layer* is the third layer and consists 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. The last layer is *Management layer* that includes ICA manager , ICA manager is used to analyse 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.

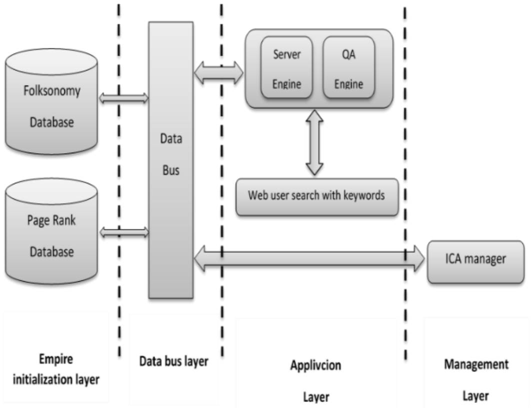


Figure 1. ICA_based Search Engine architecture.

5.2. Redefined Initialization of Empires

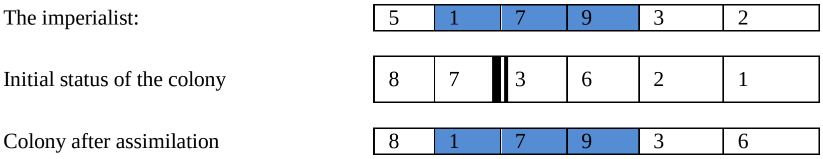
To initialize the countries at first a numerical weight should be assigned to each element of a country with the purpose of "measuring" its relative importance. It can be computed by using Folksonomy strategy which is discussed in section II. A Country can be defined as the page rank of a node in a local 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 analysing tags and user’s behaviour.

5.3. Redefined Assimilation algorithm using Random Substitution.

Redefined assimilation is implemented using *Random Substitution* approach; 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. Figure 2. Shows my Redefined Assimilation.

5.4. The Modified Revolution Process using Random Crawling strategy modified by 2-opt algorithm.

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: at first; the network of knowledge graphs should be modelled as a Markov chain[‡] in which the states (nodes) are knowledge graphs, and the transitions are the links between them. In the second step; 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. At last if two local knowledge graphs cross, use the 2 opt algorithm to find the shortest path between knowledge graphs. Figure 3 shows an example in which a network of knowledge graph is defined as (A-B-F-E-C-D-H-I-G-A) . 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.



The imperialist:	5	1	7	9	3	2
Initial status of the colony	8	7	3	6	2	1
Colony after assimilation	8	1	7	9	3	6

Figure 2. Random Substituting strategy in assimilation process

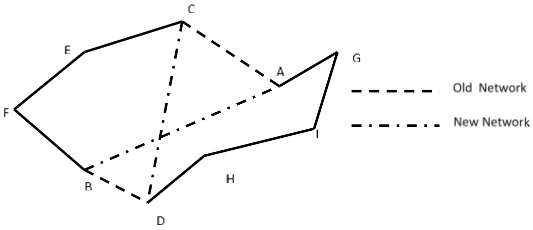


Figure .3. Random crawling in revolution step

5.5. My proposed System

Now it’s time to implement the system. In My proposed model when users enter the search keywords as the tags of web pages then the collected Information are analysed and stored in the storage server. In the next step folksonomy procedure should edit the tag to find the relationships among them, based on this relationship and users' behaviour the web pages are classified. In the next step the Semantic Search should be done using an *ICA Semantic Classifier* that is implement the ICA algorithm on Semantic Webs, before delivering the final results to the user the web pages with the same tags should be set into the same category. Finally, the results are displayed to the users.

6. Implementation and Experiments

Tagging pages is the first step in system implementation. To do this I defined 300 pages(P_1 to P_{300}) and tagged 100 of them (pages $P_{201}$to P_{300}) based on some common tags like: news, groups, social networks Iran, Philippines, Japan, portal and forums. To simplify the computation I categorized all tagged pages with the tag ‘portal’. Then I assigned default Page Ranks (A Random number between 0.1 to 0.2) to the last 100 pages (from P_{201} to P_{300}) .For the first 200 pages (P_1 to P_{200}) I considered link relationship with the next 100 pages, they were involved in the calculation. I calculated the modified PageRank score for them using modified PageRank algorithm. For computing PR_i I computed $\frac{PR_j}{L_{ji}}$ for each $1 \leq j \leq 200$ and $201 \leq i \leq 300$ the results for the first 20 pages are shown in the Table 1.

Table 1. PageRank calculation for pages between P1 to P20

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

6. Conclusions

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 . ICA is defined based on the definition of Empires which is an association of countries (basic elements) so it can accelerate the clustering of information and also avoid retrieving too much information in the results therefore it can make a considerable improvement in SEO systems. The redefined assimilation policy improves the clustering of the results and also categorizing of the web pages by using random substitution. Finally, my modified revolution process which is a combination of a scanning algorithm (Random crawler) and a local search approach (2-opt algorithm) covers more spaces in a search space by trying different paths in a knowledge graph. Future work is needed to make the system fit to the real World Wild Web. However, future work is needed to make the system fit to the real world applications. Word ambiguity should be considered in future works where a single word can take on multiple meanings. Since I do not consider multiple tags in my research its quality can be unsatisfactory. My proposed approach in this dissertation is a sample of ICA with Tuned parameters; I modified the assimilation and revolution processes, the parameters of a general ICA algorithm can be tuned better by using neural network, machine learning algorithms, and specifically fuzzy adaptive approach. I also compared may proposed method with most well-known modified Page ranking Algorithms which are introduced in section 4 and the results are shown in table 2.

Table 2. Compare with the most well-known modified Page Ranking Algorithms

Algorithm	Disadvantages	Compare with ICA Page Ranking
WPR	This algorithm totally ignores the concept of Relevancy	Relevant pages are categorized in the same empire
WLRank	The logical default positions which are considered by this algorithm does not always matches the physical position.	Predefined initial empires are exists that shows the default relation between the tags so there is no need to define the initial relative position.
HITS and PHITS	Topic drift and efficiency problem are the most obvious problems with these methods	The high quality clustering ability of ICA algorithm can eliminate the problem of topic drift
TR	Comparison based and requires more site as input	Can do the categorization with any number of countries
TIR	Important pages are mostly ignored because it increases the rank of those web pages which are opened for long time.	By using ICA algorithm important countries can be defined in an independent emperor with higher cost so they can never be ignored.
ERA	It requires a large number of characteristics to calculate the similarity.	Initial similarities is defined based on the initial empires and will be improved based on Uniting of Empires and imperialistic competition
QDR	Every page is to be annotated with respect to some ontologies and not practical for large scale data.	My proposed method only considers tags and default page ranks
DRA	If a newer page would be more interested than an old page with the same category then the crawler should perform a large calculation to calculate the distance vector	In my approach since categorization is done during the initialization of empires then all new pages are considered in categorization.

References

1. Maryam Hourali and Gholam Ali Montazer, "An Intelligent Information Retrieval Approach Based on Two Degrees of Uncertainty Fuzzy Ontology", Hindawi Publishing Corporation Advances in Fuzzy Systems Volume 2011
2. Gibbons, Kevin. "Do, Know, Go: How to Create Content at Each Stage of the Buying Cycle". Search Engine Watch. Retrieved 24 May 2014.
3. Gyöngyi, Zoltán; Berkhin, Pavel; Garcia-Molina, Hector; Pedersen, Jan (2006), "Link spam detection based on mass estimation", Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB '06, Seoul, Korea), pp. 439–450.
4. Page, Larry, "PageRank: Bringing Order to the Web" at the Wayback Machine (archived May 6, 2002), Stanford Digital Library Project, talk. August 18, 1997 (archived 2002).
5. 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.
6. Mohamed, Khaled A.F. (2006) "The impact of metadata in web resources discovering"
7. Tao Zhang, Byungjeong Lee , Hanjoon Kim, Sooyong Kang, Jinseog Kim "Collective Intelligence-Based Web Page Search: Combining Folksonomy and Link-Based Ranking Strategy" , Ninth IEEE International Conference on Computer and Information Technology, 2009.
8. Atashpaz-Gargari, E., Lucas, C. (2007). Imperialist Competitive Algorithm: An algorithm for optimization inspired by imperialistic competition, IEEE Congress on Evolutionary Computation, 4661–4667.

9. 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

10. Wenpu Xing and Ali Ghorbani, “Weighted PageRank Algorithm”, In proceedings of the 2rd Annual Conference on Communication Networks & Services Research, PP. 305-314, 2004.

11. Dilip Kumar Sharma, A. K. Sharma, A Comparative Analysis of Web Page Ranking Algorithms, International Journal on Computer Science and Engineering (IJCSE) Vol. 02, No. 08, 2010, 2670-2676 , 2010

12. Jon Kleinberg, “Authoritative Sources in a Hyperlinked Environment”, In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1998.

13. Cohn, H. Chang, “Learning to Probabilistically Identify Authoritative Documents”,. In Proceedings of 17th International Conference on Machine Learning, PP. 167–174.Morgan Kaufmann, San Francisco, CA, 2000.

14. Shen Jie,Chen Chen,Zhang Hui,Sun Rong-Shuang,Zhu Yan and He Kun, "TagRank: A New Rank Algorithm for Webpage Based on Social Web" In proceedings of the International Conference on Computer Science and Information Technology,2008.

15. H Jiang et al., "TIMERANK: A Method of Improving Ranking Scores by Visited Time", Seventh International Conference on Machine Learning and Cybernetics, Kunming, 12-15 , July 2008.)

16. Ko Fujimura, Takafumi Inoue, Masayuki Sugisaki, “The EigenRumor Algorithm for Ranking Blogs”, In WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem,

17. Fabrizio Lamberti, Andrea Sanna, Claudio Demartini , “A Relation-Based Page Rank Algorithm for. Semantic Web Search Engines”, In IEEE Transaction of KDE, Vol. 21, No. 1, Jan 2009.

18. Lian-Wang Lee, Jung-Yi Jiang, ChunDer Wu, Shie-Jue Lee, "A Query-Dependent Ranking Approach for Search Engines", Second International Workshop on Computer Science and Engineering, Vol. 1, PP. 259-263, 2009.

19. Ali Mohammad Zareh Bidoki , Nasser Yazdani, “DistanceRank: An Iintelligent Ranking Algorithm for Web Pages”, Information Processing and Management, 2007.