A new XML-based competitive architecture for search engines

Iman Rasekh
Institute of Computer Science, University of the Philippines Los Baños
Los Baños, Laguna, Philippines
irasekh@up.edu.ph

Eliezer A. Albacea, PhD
Institute of Computer Science, University of the Philippines Los Baños
Los Baños, Laguna, Philippines
eaalbacea@uplb.edu.ph

***Abstract*— Nowadays, many architectures have been developed for search engines. The key concept of search engines is called Web Based Information Retrieval (WBIR), and it is involved in the retrieval of information. However, considering the large volume of data on the internet, returning values are mostly irrelevant to the searched queries. The current information retrieval system does not present an ideal behavior, thereby attempts for a new architecture are still going on. Our proposed search engine, which is called MISE, is able to improve the retrieval of information using XML search and it is capable of clustering the results using ICA algorithm.**

***Index Terms*— search engine, crawler, architecture, XML search, page rank, ICA**

I. Introduction

Due to the huge number of web pages that exist in World Wide Web, analyzing and clustering results of search queries is still the most important challenge in designing search engines. Today, more than half of all retrieved web pages in any search engine have been reported to be irrelevant [1].

The major problem with WBIR systems is that even though the search systems can acquire large amount of web pages reflecting users' preference from the internet, it is still unsatisfactory to analyze and cluster them because of the huge number of web pages. To obtain better search results from the massive number of web pages on the internet, a new architecture based on Imperialist Competitive Algorithm, XML, and regular expression, which is named Modified Imperialistic Search Engine (MISE), is now being proposed.

Regular Expression Search is the simplest method of search and it helps us to inspect different parts of the source code of the pages. In our proposed system, the retrieval performance will be improved by using XML Search (XML Information Retrieval System) that can find the relevancy of retrieved pages. Meanwhile, Imperialist Competitive Algorithm (ICA) is a socio-politically motivated global search strategy that has recently been introduced for dealing with different optimization tasks. This algorithm was used here for clustering the results.

The rest of the paper is organized as follows: Section 2 discusses the search engine architecture and its concepts. Section 3 introduces the Imperialist Competitive Algorithm that was used. Section 4 is talking about the related works. Section 5 introduces the proposed architecture and finally, the implementation of the proposed system is described in section 6.

II. Search Engine Core Architecture

In a standard search engine, websites that are more important should receive more links from other websites. This means that they should appear at the top of search engine results. This process is mostly done based on the concept which is called Page Rank. Page Rank is the algorithm developed to rank websites in their search engine results. Each page has a predefined default Page Rank as the initial value for each page [2].

On the other hand, Search Engine Optimization (SEO) is the process of maximizing the number of visitors of a certain website. The websites with high SEO are called SEO-friendly. Search engines like Google have a guideline for boosting SEO. Google uses its own crawling robots (Googlebots) to find new websites and they do the crawling and indexing of the pages of each website, which usually takes a long time. To facilitate the process of crawling and indexing, Google introduced the Google Webmaster[i] Tool that helps websites to be search-engine-friendly [3].

III. Competitive Intelligence

Imperialist Competitive Algorithm (ICA) is a socio-politically motivated global search strategy that has recently been introduced, and it is used in dealing with different optimization tasks [1]. ICA is an evolutionary global search strategy which is defined based on four terminological definitions of Countries, Imperialists, Colonies and Empires. Country (chromosome, population) is defined as the candidate solution. Best countries (countries with the least/most cost) are called Imperialists. Colonies are defined as the rest of countries, or countries that are not Imperialists. Meanwhile, Empires are formed from Imperialist states in addition to their relevant colonies [4]. ICA is used for finding the optimum solution based on five basic concepts of Assimilation, Revolution, Exchanging

Positions of the Imperialist and a Colony, Uniting similar Empires and making new empires, and Imperialistic Competition [1].

IV. Related Works

Weighted Page Rank algorithm [5], a modification of the General Page Ranking Algorithm, uses graph theory to analyze the node and find the connection structure of a web site. WLRank algorithm [6] uses both Web Structure Mining and Web Content Mining techniques for Page Rank computing. Hyperlink-Induced Topic Search (HITS) algorithm is the oldest official Page Ranking. Web Structure Mining and Web Content Mining are the main techniques that are both used in HITS algorithm [7]. Modified HITS (Clever algorithm) [8] is a modification of HITS that provides a weight value to every link depending on the terms of queries and endpoints of the link. The same with HITS, Web Structure Mining and Web Content Mining are the main techniques that are used in Clever algorithm. Tag Rank algorithm (TR) is an algorithm for ranking the web pages based on social annotations. It calculates the heat of the tags by using time factor of the new data source tag and the annotations behavior of the web users [9]. Journal based Ranking [10] is a combination of General page ranking and HITS algorithms. Time Rank Algorithm works based on Web Usages Mining, which is a process of extracting useful information from server logs that is used for finding out what users are looking for on the Internet [11]. Query Dependent Page Ranking algorithm measures the similarities between the queries [12]. Distance Rank algorithm (DRA) is an intelligent ranking algorithm based on reinforcement learning algorithm [13]. In this algorithm, ranking is done based on the shortest logarithmic distance between two pages.

V. Proposed core Architecture for MISE search engine

In this part, the proposed architecture will be explained. Figure 1 shows the overall architecture of the proposed system.

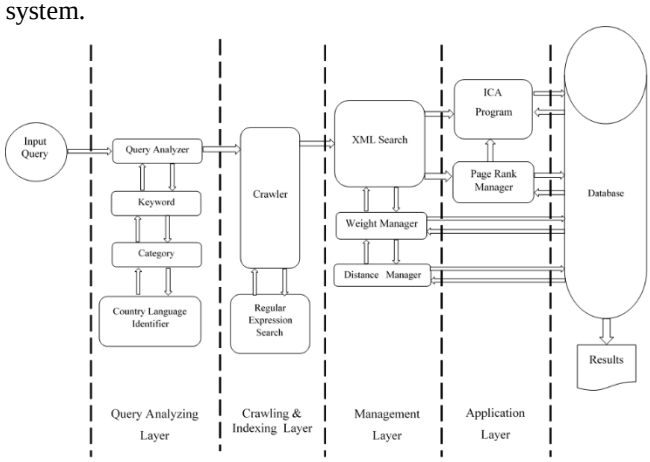

Fig. 1. The Proposed core for MISE Search Engine

The architecture is designed in four main layers: Query analyzing Layer, Crawling and Indexing Layer, Management Layer, and Application Layer.

Query Analyzing layer is the first layer that analyses the input query and it has three main parts: *Keyword Analyzer*, *Category Analyzer*, *and Country and Language Identifier*. *Keyword Analyzer* is the first part of this layer that takes into account keywords and it only returns page if all keywords are present within the page and they are related to the associated concept as described in the relational note associated with each page. *Category Analyzer* can find the real category of pages that are not only based on the category tags, but based on the whole HTML code. *Country and Language Identifier* can find country and language by checking Meta and Geo tags.

Crawling and Indexing Layer has two main parts. *Crawler* is the first part, and an open source crawler was used for this part. *Regular Expression Search* can find the elements of query in different parts of the source code.

Management layer consists of three parts: *XML search*, *Weight manager*, and *Distance manager*.

XML Information Retrieval System (XML Search) can improve the retrieval performance. The proposed XML Search consists of three major units. *Sitemap Search* can find the websites that have sitemaps and then use the sitemap information to update database fields. *Global Traffic Rank* shows the popularity of the website in the world. Lastly, *Country Traffic Rank* shows the popularity of the website in the target country. Figure. 2 shows the proposed structure for XML Search.

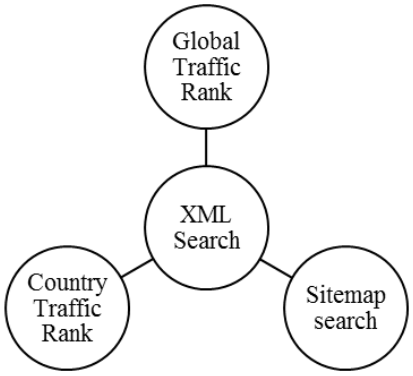

Fig. 2. The proposed structure for XML Search

Weight Manager calculates the weight of a web page based on the input and outgoing links, as well as the importance of the page.

Distance manager finds the shortest logarithmic distance between two pages.

Application Layer consists of *ICA program* that do the clustering of the results and the *Page Rank Manager* that calculates the page Rank.

The results are stored in a database, and the content of the database updates based on the page rank, weight, and distance. Finally, clustering of the contents of the database is done using ICA algorithm.

VI. Implementation and Analytic Comparison

This section presents the analytic comparison of MISE search engine and Google. Eight search queries were used to test each search engine. The first ten documents on outputs were evaluated for relevancy.

To judge the search results for the purpose of relevancy, we had to consider the concept of familiarity in the selected queries. We selected eight queries within the domain of Travel to/in the Philippines and classified the results into four groups as is shown in Table 1. These queries were designed based on their popularity and number of monthly searches in Google[1]. We also tried to find the queries that do not create priority in Google results [14]. For this experience, we used a computer with Windows 7, 2 GHz Inter Core 2 Dual processor and 2 GB RAM as our Test Environment.

Table 1. Test queries and their popularity

	Query	Popularity	Group
1	Visa to the Philippines	4463160	Local searches(A)
2	Tours to the Philippines from Seoul	282200	
3	Visit the Philippines	135560	Descriptive searches(B)
4	Holidays in the Philippines	811110	
5	Hotels in the Philippines	5241990	Commercial searches (C)
6	Hotels in Manila	2475220	
7	Map of the Philippines	15497570	Informational searches (D)
8	Shopping in Makati	244200	

A. Response time

Response Time[2] of each search engines measured using Google Chrome Web Development Tools are shown in Table 2. Considering that our search engine is running from a simple host that supports up to 1GHz dedicated CPU power for each website and the CPU power of Google search engine is unlimited, our search engine has a better result

Query	Imperialistic Search	Google
1	0.55	0.41
2	0.51	0.48
3	0.57	1.34
4	1.3	0.46
5	2.3	0.61
6	2.6	0.55
7	2.25	0.65
8	1.11	0.51
Mean	**1.39875**	**0.62625**

B. Precision of the results

Precision of the search engine results is defined as the relevance of a search result to a search query. A precision score should be calculated based on the number of the results within the first ten retrieved deemed to be relevant [14]. For this purpose, we used SEOCentro Keyword Density Tool[3] which is a very famous keyword analyzing tool used by most SEO engineers around the world. The average results of Keyword Density Tool was assumed as the precision score for each query [15]. In order to assess the overall performance of each search engine, the average precision score was calculated based on all ten queries for each search engine.

The precision score for each query on each search engine is shown in Table 3. Although the ranking of the precision scores varied amongst MISE and Google depending on the query, MISE obtained a slightly higher mean precision score of 0.77875. These results show that MISE retrieves more relevant information for local searches and informational searches while Google, on average, shows better results for descriptive and commercial searches. The average precision scores for each group are shown in Table 4.

Table 3. Mean precision scores for each query

Query	MISE	Google
1	28.2	23.75
2	11	14.2
3	14.81	13.33
4	17.65	13.55
5	45.95	44.95
6	79.05	78.75
7	19.6	23.35
8	25.5	23.65

Table 4. Average precision scores for each group

	Average Precision	
Group	**MISE**	**Google**
A	19.6	18.975
B	16.23	13.44
C	62.5	61.85
D	22.55	23.5
Average	**30.22**	**29.44125**

VII. Conclusions

Nowadays, extensive research is being conducted on faster and more precise clustering algorithms to improve WBIR of search engines. In this paper, it is the aim of the authors to contribute to both the scientific and the practical trend in WBIR. We tried to propose and implement a new architecture whose

[1] Calculated using Google Keyword Planner

[2] Period between entering a query and the retrieval of the first search results

[3] Available at http://www.seocentro.com/

main priority would be to provide concise and meaningful group labels for the retrieved information. By comparing the results of Google with our proposed method, the advantages of our proposed method was demonstrated.

In this research, we aimed to implement a new architecture based on Imperialist Competitive Algorithm, XML search, and regular expression. Our second directive was to redefine the ICA clustering algorithm, so that it can be compatible with large-scale search in search engines.

A. Scientific contributions

Design of a content-oriented clustering algorithm for search engines. In this research we used a combination of XML search and ICA that can explore the content of the page.

Covering large amount of data. Using ICA clustering, the search space is divided into smaller knowledge graphs (Empires). The process starts separately in each empire, therefore this algorithm is able to cover large amount of data. However, considering that all the empires are united in the end, this algorithm leads to a unique answer.

B. Future works

Future works are still needed to make the system fit for real world applications. Word ambiguity should be considered in future works. This applies when a single word can take on multiple meanings. Since we also did not consider multiple tags in this research, its quality can be unsatisfactory.

The proposed approach in this dissertation is a sample of ICA with Tuned parameters. We modified the assimilation and revolution by using some algorithms like Random Substitution, 2-OPT algorithm, and Random Crawler. The parameters of a general ICA algorithm can be tuned better by using neural network, machine learning algorithms, and specifically fuzzy adaptive approach.

A new page ranking algorithm can be designed so that it considers the location of pages. Combined with the location of users, the new page ranking algorithm is expected to greatly increase the quality of the search result. Typographical errors are still one of the biggest challenges in current search engines, and it is needed that even those that are contained within web information should be found. This problem should also be considered in future works.

REFERENCES

[1] Rasekh I "A new Competitive Intelligence-based strategy for Web Page Search", 2015 International Conference on Soft Computing and Software Engineering (SCSE'15), University of Clifornia , Berkely, March 2015 . Procedia Computer Science 62 (1877-0509), 450-456 , 2015

[2] Sharma, D.K., & Sharma, A. K., (2010) A Comparative Analysis of Web Page Ranking Algorithms, International Journal on Computer Science and Engineering (IJCSE), 02(08) 2670-2676.

[3] Eric Enge, Stephan Spencer, Jessie Stricchiola, Rand Fishkin , The Art of SEO (Theory in Practice) Second Edition Edition, O'Reilly Media, March 2012

[4] Atashpaz-Gargari, E., Hashemzadeh, F., Rajabioun, R., & Lucas, C. (2008). Colonial competitive algorithm: A novel approach for PID controller design in MIMO distillation column process. International Journal of Intelligent Computing and Cybernetics, 1(3), 337-355. doi:10.1108/17563780810893446

[5] Xing, W., & Ghorbani, A., (2004). Weighted PageRank Algorithm. In proceedings of the 2rd Annual Conference on Communication Networks & Services Research, 305-314.

[6] Baeza-Yates, R., & Davis, E., (2004). Web page ranking using link attributes. In proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, 328-329

[7] Kleinberg, J. M. (1999). Mining the Web's link structure. IEEE Computer,32(8), 60-67. doi:10.1109/2.781636

[8] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. New York: Cambridge University Press.

[9] Jie, S., Chen, C., Hui, Z., Rong-Shuang, S., Yan, Z., & Kun, H. (2008). TagRank: A New Rank Algorithm for Webpage Based on Social Web. In the proceedings of the International Conference on Computer Science and InformationTechnology. doi:10.1109/ICCSIT.2008.45

[10] Cheng et al., (2009). PageRank, HITS and impact factor for journal ranking. *Proceedings of the World Congress on Computer Science, 6*, 285-290. doi:10.1109/CSIE.2009.351

[11] Jiang, H., Ge, Y., Zuo, D., & Han, B. (2008). TimeRank: A method of improving ranking scores by visited time. In the 7th International Conference on Machine Learning and Cybernetics. doi:10.1109/ICMLC.2008.4620671

[12] Lee, L., Jiang, J., Wu, C., & Lee, S. (2009). A Query-Dependent Ranking Approach for Search Engines. *In the proceedings of the 2nd International Workshop on Computer Science and Engineering, 1*, 259-263. doi:10.1109/WCSE.2009.666

[13] Bidoki, A. M., & Yazdani, N. (2008). DistanceRank: An intelligent ranking algorithm for web pages. Information Processing and Management. doi:10.1016/j.ipm.2007.06.004

[14] Google Web Master Tools Guideline "A guide to building successful AdWords campaigns – Google" ,2015 from https://static.googleusercontent.com/media/www.google.com/en/ph/adwords/pdf/step_by_step.pdf

[15] Manendra ,"7 Online Free Keyword Density Calculator Tools" , 2015, from http://www.shoutmeloud.com/7-online-tools-to-analyze-keyword-density-seo.html

Gmail **Iman Rasekh <iman.rasekh@gmail.com>**

PCSC2016 notification for paper 90

PCSC2016 <pcsc2016@easychair.org> Mon, Feb 15, 2016 at 12:19 PM
To: Iman Rasekh <iman.rasekh@gmail.com>

Dear Iman,

We are pleased to inform you that your paper

Paper 90
A new XML- based competitive architecture for search engines

is accepted for poster display during the the upcoming Philippine Computing Science Congress (PCSC2016).

Your paper will not appear in the PCSC2016 Proceedings. However, you will be provided an opportunity to present your poster during the schedule for poster viewing.

The poster should be printed on a tarpaulin with dimensions 2 ft by 5 ft (portrait orientation). The poster should contain the following:

a. Title of the Work
b. Authors (and affiliation of each author)
c. Abstract
d. Objectives
e. Methodology
f. Main Results and Conclusion
g. References

You are required to send a copy of your poster in pdf to easychair by February 20, 2016.

PCSC2016 is hosted by the Palawan State University, Puerto Princesa, Palawan, on March 16-18, 2016. It is organized by Palawan State University and the Computing Society of the Philippines (CSP).

Please find below the reviews of your paper. We hope that the reviews will be of help to let you strengthen your paper.

Details of the conference are available in this webpage:

https://sites.google.com/site/2016pcsc/

In order for your brief poster presentation be included in the PCSC2016 program of activities, at least one of the authors should be available to present it during the poster viewing.

There will be a free pre-conference on Women in Computing in the morning of March 16, 2016. PCSC2016 will start in the afternoon of the same day.

We hope to see you at the Palawan State University, Puerto Princesa, Palawan.

Best Regards,

Jon Fernandez and Allan Sioson
PCSC2016 Program Committee Co-Chairs

Didith Rodrigo, Jade Pabico, Rachel Roxas, Henry Adorna
PCSC2016 Track Chairs

Henry Adorna

PCSC2016 Conference Chair

---------- REVIEW 1 ----------
PAPER: 90
TITLE: A new XML- based competitive architecture for search engines
AUTHORS: Eliezer A. Albacea and Iman Rasekh

OVERALL EVALUATION: 0 (borderline paper)
In general, how well is the paper written and formatted?: 4 (good)
Clearly presents ideas and arguments: 3 (fair)
Provides sufficient theoretical basis: 3 (fair)
Adequately supports claims and results: 3 (fair)
Ensures precision and correctness of tables and/or figures: 4 (good)
Provides relevant references: 4 (good)
Properly cites relevant references: 4 (good)

---------- REVIEW ----------
make the abstract relevant to what was done and discussed in the paper

clarify the introduction, as it is unclear how sections 2 relates to the research with sections 2 and 3 seem detached from the paper.

elaborate the discussion on how comparison/evaluation is done. The reader can only guess it relates to sections 2 and 3, but this is not clearly elaborated.

improve the clarity of the conclusion. The scientific conclusion seems out of place based on the evaluation performed. Furthermore gthe abstract states this is an improvement

---------- REVIEW 2 ----------
PAPER: 90
TITLE: A new XML- based competitive architecture for search engines
AUTHORS: Eliezer A. Albacea and Iman Rasekh

OVERALL EVALUATION: -2 (reject)
In general, how well is the paper written and formatted?: 3 (fair)
Clearly presents ideas and arguments: 1 (very poor)
Provides sufficient theoretical basis: 1 (very poor)
Adequately supports claims and results: 1 (very poor)
Ensures precision and correctness of tables and/or figures: 2 (poor)
Provides relevant references: 3 (fair)
Properly cites relevant references: 2 (poor)

---------- REVIEW ----------
The authors did not present exactly how the ICA was implemented and integrated in the search engine.

The comparison of the performance of the system the authors developed with that of Google's is not convincing. Exactly how those were done?

The table showing numbers for the response time for ICA vs that of Google's is confusing. What are those numbers? Are those supposed to be in milliseconds? seconds? ICA's number is greater than that of Google's, so why is ICA better here? The same comment is true for the other tables.

While the formatting of the paper is fine, the exposition of the paper is problematic. Assuming that there are real results in the paper, the authors will have to write the paper better and provide sufficient details to support the claims they will make.

---------- REVIEW 3 ----------
PAPER: 90
TITLE: A new XML- based competitive architecture for search engines
AUTHORS: Eliezer A. Albacea and Iman Rasekh

OVERALL EVALUATION: 2 (accept)

In general, how well is the paper written and formatted?: 4 (good)
Clearly presents ideas and arguments: 4 (good)
Provides sufficient theoretical basis: 4 (good)
Adequately supports claims and results: 2 (poor)
Ensures precision and correctness of tables and/or figures: 2 (poor)
Provides relevant references: 4 (good)
Properly cites relevant references: 4 (good)

---------- REVIEW ----------
Results might need to be re-evaluated as there are some inconsistencies. For example, Response Time is better the lower the value, but it was claimed that their results, having higher values, are better. IN addition, it should be cited that there may be other factors that contribute to these values, e.g., network latency.