Revealing Paths of Relevant Information in Web Graphs

Georgios Kouzas[1], Vassileios Kolias[2], Ioannis Anagnostopoulos[1] and Eleftherios Kayafas[2]

[1] University of the Aegean

Department of Financial and Management Engineering

Department of Information and Communications Systems Engineering

{gkouzas, janag}@aegean.gr

[2] National Technical University of Athens

School of Electrical and Computer Engineering

vkolias@medialab.ntua.gr, kayafas@cs.ntua. gr

Abstract In this paper we propose a web search methodology based on the Ant Colony Optimization (ACO) algorithm, which aims to enhance the amount of the relevant information in respect to a user's query. The algorithm aims to trace routes between hyperlinks, which connect two or more relevant information nodes of a web graph, with the minimum possible cost. The methodology uses the Ant-Seeker algorithm, where agents in the web paradigm are considered as ants capable of generating routing paths of relevant information through a web graph. The paper provides the implementation details of the web search methodology proposed, along with its initial assessment, which presents with quite promising results.

1 Introduction

In this paper, a new web search methodology based on the ant colony algorithm, is proposed. In more detail we suggest an ant colony algorithm approach, which is capable of tracing relevant information in Internet. Based on [1][2], Ant Colony algorithm can be applied on a connected graph Gp = (P,L), where P are the nodes, and L the link between the nodes, which represents a problem definition. Every route in this graph represents a solution of the initial problem. ACO converges in an optimal solution tracing routes in the graph. In our approach, we consider the world-wide web as a graph G. Although our methodology maintains most of the ant colony algorithm characteristics [3][4], its uniqueness lies upon the fact that it applies in an environment, the structure of which is not pre-defined. In addition, some search techniques for locating and evaluating the relevant information are used based on web page similarity [5], [6],[7] as well as web page clustering [8].

2 Basic Concepts Used

The aim of this paper is to propose a methodology that is able to trace routes between hyperlinks, which connect two or more relevant information units (web pages) with the minimum possible cost. Initially we consider a web information unit (web page), which is relevant to the user requests. This web page is considered as the starting point of the search. The search is based on the principle that when some information relevant to the user's request exists on a point-node of the world-wide web, then another point-node in a "close distance" is highly possible to contain similar (and thus relevant) information [9]. We define the hyperlink as the basic distance unit in the web universe. The "distance" between two nodes, is defined as the number of subsequent hops needed in order to be transferred from one node to another and vice versa. The methodology consists of three phases. In the first phase the start point of the search are defined. These could be either user defined web pages, either result of search engines [10][11][12]. In phase two, the suggested search algorithm takes place. The algorithm procedure runs iteratively, and each time that converges to an information unit, it specifies a new starting point. The third and last phase is to group the results according to how relevant their contents are. During the pre-processing of the hypertext documents the textual information is extracted from HTML format. Depending on the tags, we consider three levels of importance (that is High, Medium and Low). Then, the outgoing links are extracted in order to construct the search graph. The final step is the web page similarities calculation according to [13][14], in which the document hyperlink structure is taken into account, while it consists of the pre-process phase and the similarity estimation phase. As document similarity, we consider that sentences and phrases carry also significant information regarding the textual content. According to this, we used a comparison measure based not only on the similarity of individual terms but also on the similarity of sentences and phrases [15]. The similarity between two documents, d_1 and d_2 is computed according to Equation 1. In Equation 1 $g(l_i)$ is a function that marks the length of the common phrase, while s_{j1} and s_{k2} represent the initial length of the d_1 and d_2 document sentences respectively. Function $g(l_i)$ is proportional to the ratio of the common sentence portion length to the total sentence length as defined in Equation 2, while $|ms_i|$ is the matching phrase length and γ indicates a sentence partitioning greater than or equal to 1. Parallel to this procedure, the term-based similarity of the tested web pages takes place using the Vector Space Model (VSM) [16],[17] as defined in Equation 3. However, the inverse document frequency between terms is not taken under consideration, since the estimated similarity value concerns two web pages and not a collection of web pages as required from the VSM. Therefore the final document similarity SIM_i value is given from Equation 4. Result grouping is used for presenting the search results better. The acquired web pages are analyzed and presented in clusters. Each cluster contains a set of high importance information units. The cluster creation is based on the methodology proposed in [13][14] and the similarity histogram analysis. For specifying similarity, the classic vector space model is used [16],[17].

$$S_p(d_1,d_2) = \frac{\sqrt{\sum [g(l_i)\cdot(f_{i1}w_{i1} + f_{i12}w_{i2})]^2}}{\sum |s_{j1}|\cdot w_{j1} + \sum |s_{k2}|\cdot w_{k2}} \tag{1}$$

$$g(l_i) = \left(\frac{|ms_i|}{|s_i|}\right)^{\gamma} \tag{2}$$

$$sim(d_i,d_j) = \frac{d_i \cdot d_j}{|d_i| \times |d_j|} = \frac{\sum(w_{k,i} \times w_{k,j})}{\sqrt{\sum w_{k,i}^2} \times \sqrt{\sum w_{k,j}^2}} \tag{3}$$

$$SIM_i = S(d_1,d_2) = 0.5\mathrm{S}_p(d_1,d_2) + 0.5\mathrm{S}_t(d_1,d_2) \tag{4}$$

4 Ant Seeker

In this section we present the proposed search algorithm, which is based on the theoretical model of the ant colony algorithm [1]. The proposed algorithm deals with the world-wide web information search problem. The structure of the world-wide web consists of a set of information units (web pages) and a set of links (hyperlinks) between them. Thus, assuming a graph Gp = (P,L), where P are the nodes, and L the link set between the nodes, we can consider the world-wide web as a graph G with infinite dimensions.

4.1 An Overview of the Proposed Algorithm

The proposed algorithm is a slight modification of the ant colony algorithm [9], [18] and therefore it adopts most of the basic characteristics of the colony algorithms family [1]. However, several modifications were made in order to apply the proposed algorithm in the particular World Wide Web paradigm. Thus:

- Each artificial ant can visit a predefined maximum set of nodes.

- All artificial ants start from the start-node. In addition, when the algorithm starts there is no further information for the graph structure in which search will be applied. In this way, the harvesting procedure of real ants is simulated.
- The process of nodes recognition, which they have relevant content with the initial node, is based on the web page similarity mechanism as described in the previous section.

The search begins from the initial start node which is given by the user and in each step of the algorithm, each ant-agent moves from node i to node j. Assuming that in node j the pheromone value at time t is τj(t), then the ant visits node j through node i according to the pheromone function. The process is repeated until each ant visits the predefined maximum number of nodes. After the creation of the candidate routes, the best route is extracted and the node pheromone values are updated. This process is then repeated, while it ends when a convergence to a specific route is found. The node, which has the route with the maximum similarity value in comparison with the starting node, is assigned as the starting node for the next search. During the initialization phase, three variables are defined: the total ant-agent number NoA, the initial pheromone value IPV is defined in each new node and the number (Nmax) of the maximum nodes an ant can visit.

4.2 Heuristic Function-pheromone Model

For every new node is added, the content similarity to the initial one is checked. Thus, each node is characterized from a similarity value, which implies the quality of a node (Equation 4). In order to specify the quality of a node, apart from its content similarity to the initial node, a second value that defines its ability to leads to a node with high quality content is also specified. Therefore, points (nodes) with low similarity values increase their significance when they lead to points of high frequency. The calculation of a quality value is given by the heuristic function according to Equation 5, where d is the path of an ant-agent where node i is included in ($0 < d < NoA$), SIM_i is the similarity function of node j as defined in Equation 4, and SIM_j^d stands for the similarity function of node j, which belong to the route d right after its previous visit in the node i ($i < j < N_{\max}$). As mentioned before, in the initial phase, each node inserted in the graph, has a pheromone value IPV. Every time, a complete iteration of algorithm occurs, the pheromone is updated. More specifically, the nodes used as intermediate or final points in the ant routes are updated. This procedure is given from Equations 6 and 7. In Equation 6, h_i is the heuristic function given from Equation 5, while k is the number of ants that used node i for the route creation. According to Equation 7, the nodes that lead to high frequency routes increase their pheromone values substantially over the algorithm iterations. To avoid infinite assignment of pheromone values in certain nodes, the pheromone value is normalized between zero and one as defined by Equation 8, where $\tau_{\max}(t+1)$, is the maximum pheromone value in the current iteration. Each time an ant is in a node *i*, it must choose the next node *j*. The nodes, which are considering to be visited, are the directly connected nodes. This defines the accessibility value of each web page given by Equation 9. In order to avoid endless loops, accessibility excludes the nodes, which contributed in the past in the creation of the route. The algorithm utilizes the classic probability model of the ant colony algorithms given from Equation 10. Whenever an algorithmic iteration occurs, ants make a route, based on the pheromone value and the quality of the nodes in the search graph. The nodes with the highest quality values increase their pheromone values, and thus they have higher probabilities to be chosen. A solution consists of a chosen route (set of nodes) and not a single node. As solution we define the node of the final route, which presents the larger similarity value. This node is then added to the list of solutions.

$$h_i(t+1) = \max(\underset{i<j<N_{MAX}}{SIM_j^d}, SIM_i, h_i(t)) \tag{5}$$

$$\Delta\tau_i = kh_i \tag{6}$$

$$\tau_i(t+1)' = \tau_i(t) + \Delta\tau_i \tag{7}$$

$$\tau_i(t+1) = \frac{\tau_i(t+1)'}{\tau_{\max}(t+1)} \tag{8}$$

$$\eta_{ij} = \begin{cases} 1 & \text{if node j is directly linked from node i} \\ 0 & otherwise \end{cases} \tag{9}$$

$$P_{ij} = \frac{\tau_j \cdot \eta_{ij}}{\sum_{k \in allowed_k} \tau_k \cdot \eta_{ik}} \tag{10}$$

5 Results - Evaluation

This section presents the results of the proposed methodology. The evaluation took place in two experiment phases. In the first we evaluate the performance of the proposed Ant-Seeker algorithm [9], [18] by applying the algorithm in three different queries. In the second phase we evaluate the introduction of some clustering techniques to the methodology for grouping the results. The search procedure was examined by querying different parts of the World Wide Web three times. During this procedure we used only web-pages with English content. In order to apply and evaluate the algorithm, we followed three steps. The first step involves the preprocessing of the WebPages; the second involves the balancing of variables NoA, Nmax and IPV while the third step involves the algorithm execution. For all experiments the variable values where chosen to be NoA=10, Nmax=3, NC=100 and IPV=0.4. For each search the set of returning results is equal to the number of algorithm iterations. This reduces the result quality, but this is important especially in cases where the algorithm doesn't manage to converge in a relative to the query solution during the initial search stages. The results of clustering to the three algorithm evaluation sets appear in table 6. Applying clustering methods to the returning results, the percentage of related document retrieval is decreasing (between 2% to 5%) but at the same time their quality increases (from 30-50% to 80-90%). This is an expected behavior, because, the nodes-pages used for the search continuation only, are cut off due to the low similarity value with the initial document. However, a small part of the correct results, are not ranked correctly during the clustering. The explanation is that the similarity calculation model during algorithm execution is given by Equation 4 and on the other hand the similarity calculation model, during clustering is the vector space model [17]. We use a different function to calculate the similarity because the final collection of the web pages is unknown during the search, so we use equation 4 that defines similarity between a pair of pages. On the contrary during clustering the total returning results virtually defines the collection. In table 2 the results of applying the methodology for 6 random queries in the world-wide web appear. The proposed algorithm's ability to seek and extract information relative to the query from the world-wide web is outlined in the experimental results. However during the experiments we created a set of constraints.

The most important constraint is the scale of the world-wide web which did not allow applying the algorithm in a wider scale search. The sample size was of value 200.000. The proper evaluation of the algorithm requires full definition of the samples as concerning their informational relativity to the reference node. The classification of all the samples based on similarity gives an estimation of the relation of the documents and therefore a classification measure but still remains a mechanical classification method which cannot replace the human factor. For the evaluation some machine learning techniques could be used as in neural networks [10], [11] but a set of already classified documents is required in order to extract the relative documents.

The second constraint that virtually is a result of the previous constraint is the overlapping between searches. The algorithm includes an overlapping search prevention mechanism in order to avoid creating cyclic search routes. However, in a limited portion of the world-wide web forbidding backtracking would result in a search termination in only a few steps (2 to 5 searches per sample). For this reason, the only constraint assigned for route creation was blocking adding a node, which belongs to the current set of solutions of the algorithm.

6 Conclusion

By evaluating the behavior of the algorithm we observe that it enables search in real time. The ability to choose the direction of the search autonomously, allowing the search of unknown territories at the same time is noteworthy. An important advantage in contrast to other classic search and classification methods is the fact that it does not require covering the whole search space. In the experiments the coverage percentage was at 40%, taking all the constraints under consideration. However the retrieved document percentage is limited to 80% as depicted in table 1. The document structure in which the search takes place doesn't seem to affect the search. In addition as the structure tends to that of a fully linked document the algorithm performance is increasing. In contrast to the most search techniques the ability to use a textual query allows quality searches. Search with queries of type of a set of terms has the advantage of being short as long as the query is accurate. But when the query is inaccurate then the results are confusing. The use of a reference document for the search improves the quality of the returning results. Of course, the documents relative to the informational need, but with low similarity value with the reference document cannot be retrieved. According to the experiment results the covering percentage is about 40% with an average near 90%. Although the download percentage for the specific covering percentage is satisfying, applying the algorithm to large scale searches is prohibitive. Applying the algorithm to the world-wide web for the second set of experiments is quite time-consuming. Additionally, adding solutions unrelated to the query produces low quality results. Adding clustering techniques to the retrieved documents improves the quality of the results, but for each irrelevant to the query document return, an analogical search cost is added. As an alternative method for cutting off bad solutions could be an additional variable to the algorithm solving rule. This value would define a minimum limit value to the similarity of the candidate solution to the reference document.

Table 1 System performance using clustering techniques

Experiment	Relative Pages	Algorithm		Clustering		Download (%)	Accuracy (%)
		Total	Found	Relative	Correct		
1	43	80	37	42	35	81,40	83,33
2	34	85	29	32	26	76,47	81,25
3	72	135	65	70	65	90,28	92,86

Table 2 System Performance for random internet queries

Experiment	Relative Pages	System Relative	Correct	Download (%)	Accuracy (%)
1	32	37	28	87,50	75,68
2	40	56	32	80,00	57,14
3	21	25	19	90,48	76,00
4	10	8	8	80,00	100,00
5	17	17	15	88,24	88,24
6	36	42	33	91,67	78,57

References

1 M. Dorigo and T. St¨utzle. Ant Colony Optimization. The MIT Press, 2004.

2 Dorigo M., and Caro G.D., 1999, “Ant Algorithms Optimization. Artificial Life”, 5(3):137-172.

3 Dorigo M., and Maniezzo V., 1996, “The ant system: optimization by a colony of cooperating agents”. IEEE Transactions on Systems, Man and Cybernetics, 26(1):1-13.

4 Dorigo M. and Caro G.D., 1999, “The Ant Colony Optimization Meta-heuristic” in New Ideas in Optimization, D. Corne, M. Dorigo, and F. Glover (Eds.), London: McGraw-Hill, pp. 11-32

5 Pokorny J (2004) Web searching and information retrieval. Computing in Science & Engineering. 6(4):43-48.

6 Oyama S, Kokubo T, Ishida T (2004) Domain-specific Web search with keyword spices. IEEE Transactions on Knowledge and Data Engineering. 16(1):17-27.

7 Pokorny J (2004) Web searching and information retrieval. Computing in Science & Engineering. 6(4):43-48.

8 Broder A, Glassman S, Manasse M, Zweig G. Syntactic clustering of the Web. Proceedings 6th International World Wide Web Conference, April 1997; 391-404.

9 G. Kouzas, E. Kayafas, V. Loumos: “Ant Seeker: An algorithm for enhanced web search”, Proceedings 3rd IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI) 2006, June 2006, Athens, Greece. IFIP 204 Springer 2006, pp 649-656.

10 I. Anagnostopoulos, C. Anagnostopoulos, G. Kouzas and D. Vergados, “A Generalised Regression algorithm for web page categorisation”, Neural Computing & Applications journal, Springer-Verlag, 13(3):229-236, 2004.

11 I. Anagnostopoulos, C. Anagnostopoulos, Vassili Loumos, Eleftherios Kayafas, “Classifying Web Pages employing a Probabilistic Neural Network Classifier”, IEE Proceedings – Software, 151(03):139-150, March 2004.

12 Anagnostopoulos I., Psoroulas I., Loumos V. and Kayafas E., “Implementing a customized meta-search interface for user query personalization”, Proceedings 24th International Conference on Information Technology Interfaces (ITI’2002), pp. 79-84, June 2002, Cavtat/Dubrovnik, Croatia.

13 K.M. Hammouda, M. S. Kamel,”Phrase-based Document Similarity Based on an Index Graph Model”, Proceedings IEEE International Conference on Data Mining (ICDM’2002), December 2002, Maebashi City, Japan. IEEE Computer Society 2002, pp. 203-210.

14 K.M. Hammouda, M. S. Kamel, “Incremental Document Clustering Using Cluster Similarity Histograms”, Proceedings WIC International Conference on Web Intelligence (WI 2003), October 2003, Halifax, Canada. IEEE Computer Society 2003, pp. 597-601

15 J. D. Isaacs and J. A. Aslam. “Investigating measures for pairwise document similarity. Technical Report PCS-TR99-357, Dartmouth College, Computer Science, Hanover, NH, June 1999

16 G. Salton, M. E. Lesk. Computer evaluation of indexing and text processing, Journal of the ACM, 15(1):8-36, 1968.

17 G. Salton. The SMART Retrieval System – Experiments in Automatic Document Processing. Prentice Hall Inc., 1971.

18 Kouzas G., E. Kayafas, V. Loumos “Web Similarity Measurements using Ant – Based Search Algorithm”, Proceedings XVIII IMEKO WORLD CONGRESS Metrology for a Sustainable Development September 2006, Rio de Janeiro, Brazil.