2013 International Conference on Advances in ICT for Emerging Regions (ICTer): 168 - 171

KAnt: Leveraging ant colony optimization for automatic knowledge acquisition from web documents

Rivindu Perera[#1], Udayangi Perera[#2]
[#]*99X Technology, 65, Walukarama Road, Colombo 04*
[*]*Informatics Institute of Technology, 57, Ramakrishna Road, Colombo 06*

[1]rivindu.perera@hotmail.com, [2]udayangi.p@iit.ac.lk

***Abstract*— This paper suggests a novel algorithm (KAnt) inspired by ant colony optimization strategies for knowledge acquisition. KAnt algorithm attempts to devise a unique solution for eminent knowledge acquisition problem of losing interest in content rich documents due to low familiarity. We utilize our solution to work with web based documents, considering documents as nodes in a graph problem. Locating content rich documents is achieved through intelligent ants that are equipped with numerical statistic for document identification. Documents are found via pheromones deposited by such ant colony. Experimental results acquired through domain expert evaluation show that our proposed approach has contributed for knowledge acquisition remarkably.**

Keywords— **Ant Colony Optimization, Knowledge acquisition, Web based documents, Swarm intelligence**

I. INTRODUCTION

Knowledge acquisition is a basic but influential cognitive task performed by humans. Later this cognitive task is modelled in a computational background with the name of automatic Knowledge Acquisition (KA), which then amalgamated with World Wide Web, instigating renowned automated KA from web documents.

Though diverse research attempts are taken towards KA from web documents, still it has not attained to the level of expectation. Recent researches which analyses the hidden reasons for these issues encompass that roots of this issue is associated with current web architecture. As current web architecture is based on topic or concept familiarity, documents which highlight the topic as a whole without rich content are appeared in the top of the search result list during knowledge acquisition from web [1] [2]. This issue is turned into a worst case when knowledge seekers are trying to access knowledge from the web targeting on particular topic based knowledge. Therefore, it is clear that this dilemma is one of the major drawbacks in current knowledge acquisition domain and needs long lasting and intelligent solution.

However, numerous research attempts taken towards this solution cannot handle the problem intelligently though they are equipped with advancements in the field. Amalgamation of semantic interpretation in knowledge acquisition proposed by Gomez and Segami [3] is one significant attempt made recently. In this research they mingle semantic interpreter based on English for knowledge acquisition. But major disadvantage noticed in this approach is that fundamental drawback is not addressed adequately. Simply, semantic interpreter can lead to inaccurate result if the web document is fulfilled with familiar terms that we try to eliminate. But methodology proposed by Jahiruddin et al. [4] has more value headed for drawback noticed. In this approach they totally depend on conceptual representation leading the process to biomedical text mining. Though it is in a different domain, it is vital to analyse this research as it adds further improvement to the intelligent knowledge acquisition. Nevertheless, concept mining technology introduced by them using Latent Semantic Analysis (LSA) inspired us to analyse its effect to our approach. But it is noted that still their approach is based on term extractor from biomedical documents before LSA. Though term extraction is mandatory for LSA the way they have performed it using shallow analysis and without considering the actual document content through semantic analysis is one factor that we noticed as weak.

During our investigation, we found several interesting strategies and tactics implemented to address the issue mentioned by our research as the problem statement. Among them collaborative knowledge acquisition approaches presented by [5], [6] and [7] can be considered as decent methodologies. Finally, as no adequate solution is provided for the issue brought up by this research, it inspired us to analyse and model our novel paradigm to fill the current gap in research.

This paper is structured as follows. Section II expresses Ant Colony Optimization (ACO) and intelligent ants considering how this strategy is applied in the current domain. Section III concentrated on analysis of the technique behind the KAnt algorithm which we present through this research. In section IV, we illustrate the empirical result which we achieved during evaluation giving inspiration to conclusion and future work, mentioned in section V.

II. ANT COLONY OPTIMIZATION AND INTELLIGENT ANTS

ACO is motivated by pheromone centred approaches of ant forging and listed under category of optimization algorithms [8]. Basic idea of ACO is to support an identified best solution attracting other agents as well to deal with the solution found [9]. However as this process is inspired form the biological environment, it has more practical value in a number of areas in computer science.

978-1-4799-1276-6/13/$31.00 ©2013 IEEE

When analysing past research attempts, we noticed that ACO is used in data mining [10], sequence alignment [11] and heterogeneous information handling [12]. These mentioned approaches belong to a subset of applications based on ACO, which closely related to the approach we discussed in this paper. Specially, among these decisive researches, methodology mentioned in [10] can be considered as important and influential.

Concept of intelligent ants is one unfamiliar term that we introduce in this paper. In our approach we consider all ants are equipped with ability to determine the quality of the document they find and deposit pheromones proportional to the quality. Therefore, we name all ants as intelligent as they contribute to the process intelligently supporting the main task of knowledge discovery.

In next sections we closely examine term frequency-inverse document frequency (tf-idf) scoring, the concept of ACO and its applicability in this research with intelligent ants.

III. TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY

Text mining concept we employ for our knowledge acquisition is tf-idf scoring which determine the basic quality of document being considered in relation to the collection. Formula 1 and Formula 2 expresses tf and idf scoring strategies respectively.

$$tf\ (t, d) = \frac{f(t,d)}{\max\{f(w,d) : w \in d\}} \quad (1)$$

$$idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \quad (2)$$

where, f(t,d) denote the raw frequency of term t in the document d and w stands for any term in the document. D represents the total number of documents considered for knowledge acquisition.

IV. SCIENCE BEHIND KANT

A. Scope of Kant

KAnt only considers web documents in Hypertext Mark-up Language (HTML) pages and Portable Document Format (PDF) files hosted in World Wide Web. There-fore, search functionality which is not in the scope of KAnt is responsible to return appropriate set of documents for KAnt for knowledge acquisition from identified documents with rich content. Furthermore, algorithm used for the initial document search is totally independent from the process performed by KAnt. Because of that implementation can incorporate any searching algorithm for initial web document search as long as it can return set of documents. KAnt will then try to analyse documents and will identify documents with rich content for knowledge acquisition. In KAnt implementation, we incorporated a link analyser which can navigate through links mentioned in web documents. Sole purpose of this is to maximize the search of rich content rather depending on the result list returned by initial we search.

B. Overview of Kant

One main concept that we have to emphasize is that KAnt assumes all web documents can be represented in as a node in a graph problem. Therefore, regardless of whether document is hosted in a web server or File Transfer Protocol (FTP) server, KAnt model all documents are in an environment that it can access in a similar manner. As a result of this consideration underlying document search engine is loaded with an additional responsibility to compile such common representation. But KAnt implementation is equipped with a link analyser which can navigate from document to document. Fig. 1 depicts an example of such representation of document collection.

Furthermore, it is important to notice that in this research ACO is not employed for shortest path navigation and therefore it shows considerable difference from the earlier approaches. Instead of this traditional approach, we have incorporated the Ant nest where ants live and the place which they store knowledge. Therefore, each ant after knowledge acquisition should return to its nest using the same path it travelled searching document. In addition, if ant selects an edge then it can move to the end of the path through passing several nodes. According to the graph shown in Figure 1, particular ant can move in following sample paths, nest→G→C→B, nest→C, nest→C→B, nest→D etc. In addition, it can be identified that document A (Doc A) has two links directing to Doc I and Doc D.

In addition path like nest→G→C→B will be considered as one document collection as it contributes for the knowledge as a whole, but not individually. Best path will be selected according to the sum of all pheromones deposited, showing a greedy behaviour of favour for pheromones.

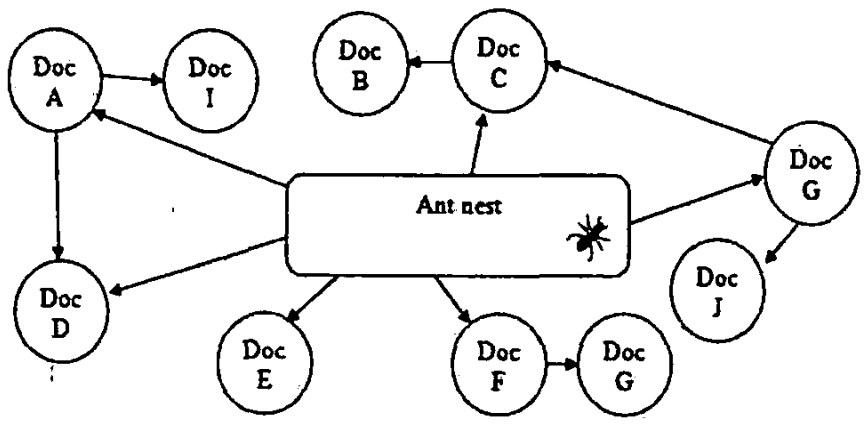

Figure 1 - Example of web document representation

Entire algorithm can be considered with 6 steps as follows:

Initial step: Documents are represented in a graph according to the search result list acquired for specified terms and these are stored in a First-In-First-Out (FIFO) queue for later usage (to analyse tf-idf scores).

Step 1: Ants move randomly in the graph covering all paths (number of ants propositional to the number of paths)

Step 2: While moving ants deposit pheromones in the edge as a weighted factor propositional to the tf-idf score of each document (in this context entire document collection is incorporated to measure inverse document frequency value.)

Step 3: Once all ants have finished their job evaporation happens by a small quantity to reduce the partiality.

Step 4: Terms are replaced with synonyms using Bag-of-Word model [13] and ants restart the job.

Step 5: Once all terms are fmished and if there no terms coming from bag-of-word model process is stopped.

Step 5: Once all terms are finished and if there no terms coming from bag-of-word model process is stopped.

Step 7: Knowledge acquisition is performed through extraction of relations from the documents discovered using conceptual graph structures.

C. Edge selection and virtual pheromone deposit

Generic edge selection is employed for this research with customization for our proposed solution and formula is given in (3) (with customizations for the original representation taken from [8]):

$$p_{ij}^{k} = \frac{\tau_{ij}^{a}\eta_{ij}^{a}}{\sum_{h \in J^{k}}^{H} \tau_{ih}^{a}\eta_{ih}^{a}} \quad (3)$$

Where, τ_{ij} represents pheromones deposited on the edge, η_{ij} is expressed as visibility of the node worked out by using tf-idf score of currently directed node, a and b represent weighted factors to determine the importance of previously mentioned criteria. If a = 0, selection strategy of ants is based on tf-idf scores and if b = 0 then selection strategy is based on pheromone amount deposited. Furthermore, H shows all available nodes and J^{k} represent nodes t that ant k has not visited.

We simulate the virtual pheromone deposit process according to the tf-idf score computed for the next document to be visited. For an example, if ant expects to visit from node A (Document A) to node B (Document B), then pheromones deposited in the edge A→B, is proportional to the tf-idf score of Document B. Therefore, we can show our pheromone deposit strategy using formula (4) given below:

$$\Delta\tau_{ij}^{k} = Q / T^{k} \quad (4)$$

Where, Q represent a constant which is set to tf-idf score computed for a document which is found as content rich using greedy term searching method, a simple heuristic which analyses all available documents for terms. T^{k} shows the tf-idf score of next document to be visited by ant k using the current edge.

D. Shaping daemon actions and updating pheromones

Once the ants have finished depositing pheromones using search based on actual terms which are not in the set of terms coming from Bag-of-word model [14] [16], we execute the daemon action to tag paths which are with high density of pheromones.

Basic idea behind this daemon action is to maximize the opportunity for actual terms before they are replaced by synonyms or semantically related concepts through Bag-of-word model. Reason behind this type of a implementation is that past researches in the semantic analysis have shown that Bag-of-word models can easily be subjected to erroneous results with incorrect term replacements [14]. Specially, in [14], Wu and Hoi empirically investigate the Bag-of-word model and coming up with semantics preserving model. Though this research investigates a slightly different domain ideas mentioned are applicable for our approach as well. Chau et al. [15] bring up another approach towards bag-of-word model considering the clustering of conceptually related concepts which is similar to methodology presented in [16]. But in our approach we consider bag-of-word model as a shallow processing technique which is employed with its original form of inspiration.

E. Document discovery and knowledge extraction

Document discovery, initial step in our knowledge acquisition approach is based on density of pheromones. Therefore, the path which has the highest density of pheromones is selected for knowledge extraction process.

Though we know the documents which are equipped with rich content of knowledge, it is not enough for the knowledge acquisition task.

To extract knowledge we employ, Conceptual Graph (CG) ([17]) based knowledge extraction where relations found in the document are modelled to CGs. To extract relations from the document, methodology presented in [18]. We noticed that approach mentioned in [18] has more practical value as well as can direct to better accuracy when comparing with other similar approaches, [19]. It is also noted that though some of the approaches we considered here can bring excellent accuracy in one domain, it cannot reach to that highest accuracy in a general text mining task. Sample conceptual graph after relation extraction of a web document is shown in Figure 2 below.

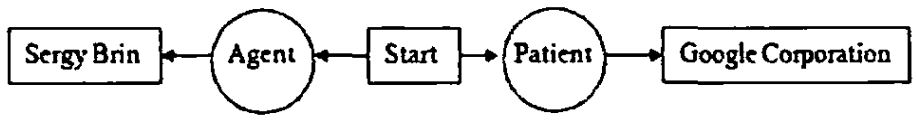

Figure 2 - Sample CG after knowledge acquisition

V. RESULTS AND DISCUSSION

To evaluate the novel approach presented by this research, we used analysis per-formed by domain experts. Group of domain experts who tested the approach consisted of 11 members representing 8 different domains: weather, sports, health, business, science/technology, culture/ politics, law and computer/internet.

During evaluation we restricted the number of search result list to 50, therefore we obtained more than 50 documents as our approach can find URIs in documents and acquire them as well. For the entire evaluation, 14 phrases are considered and these are listed in Table 1 below.

We analysed the tf-idf value that is associated with document collection selected by KAnt over decision taken by domain experts after deep analysis of entire document collection. Final result we attained is shown in Table 1, with tf-idf score for KAnt best case (S_k) and domain expert vale for KAnt best case (S_d). We consider the set of documents selected by KAnt (d_k) and set documents selected by domain experts (d_e) as documents having rich content for knowledge acquisition for particular query.

According to the empirical evaluation carried out, we noticed that our approach has achieved better accuracy when comparing with domain expert decision for the same set of documents considered.

Phrase	S_k	S_d (%)	$\#d_k$	$\#d_e$	$\#(d_k \cap d_e)$
FIFA World Cup 2010	0.68	85	8	5	4
World War I	0.03	23	3	6	3
Space Shuttle Discovery	0.74	65	12	17	6
Economy of Sri Lanka	0.62	72	4	4	1
Hurricane Katrina 2005	0.78	82	6	4	3
Food safety	0.03	13	5	4	1
Local Government Finance Act	0.45	61	2	5	3
Agile software development	0.04	53	9	9	0
Democratic National Committee	0.56	76	11	3	9
Scale-free networks	0.23	47	4	10	7
Health and Social Care Act 2012	0.07	23	7	3	5
Biosafety	0.21	45	8	7	2
Gulf of Mexico	0.05	19	11	9	3
Internet Protocol Television	0.75	51	9	4	7

Table 1- Evaluation result of KAnt

But in some scenarios several gaps are also noticed. However, we tried to get the accuracy considering the intersection of document sets output by two approaches according to the formula 5 given below.

$$Relative\ accuracy = \frac{\#d_k}{\#(d_k \cup d_e)} \qquad (5)$$

In this analysis, we noticed that our approach has achieved 76.28% average accuracy value for the considered 14 phrases. When considering other approaches in the same domain, this can be considered as an excellent accuracy level for knowledge acquisition problem.

VI. Conclusion and Future Work

This paper presented a novel approach towards knowledge acquisition inspired by ant colony optimization strategies. In addition to these main technologies we amalgamated some text mining approaches to model our new algorithm to improve the knowledge acquisition quality by addressing the problem of losing interest in content rich document due to familiarity. This novel approach is empirically evaluated and we attained excellent accuracy which inspired us to carry our research further enhancing the quality. But it should be emphasized that limitations we noticed in this research should also be considered into account during enhancement. As our approach is based on bag-of-word model for solution search, it should be noted that semantic value of the latter stages can reach low values. We tried to address this by introducing daemon function, but our belief is that it is not satisfactory. However, augmentation for solution search and tf-idf measurements are considered as future works, which we believe as essential future enhancements.

References

[1] G. Kumaran, R. Jones, and O. Madani, Biasing web search results for topic familiarity, in *Proc. 14th ACM international conference on Information and knowledge management*, 2005, ACM: Bremen, Germany. p. 271-272.

[2] O. Hoeber and X.D. Yang, A Comparative User Study of Web Search Interfaces: HotMap, Concept Highlighter, and Google, in *Proc. of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, 2006, IEEE Computer Society. p. 866-874.

[3] F. Gomez and C. Segami, Semantic interpretation and knowledge extraction. *Knowledge Based Systems*, 2007. 20(1): p. 51-60.

[4] M. Jahiruddin, Abulaish, and L. Dey, A concept-driven biomedical knowledge extraction and visualization framework for conceptualization of text corpora. *Journal of Biomedical Informatics*, 2010. 43(6): p. 1020-1035.

[5] T. Chklovski and Y. Gil, Improving the design of intelligent acquisition interfaces for collecting world knowledge from web contributors, in *Proc. of the 3rd international conference on Knowledge capture*, 2005, ACM: Banff, Alberta, Canada. p. 35-42.

[6] D. R. Liu and C.W. Lin, Modeling the knowledge-flow view for collaborative knowledge support. *Knowledge Based Systems.*, 2012. 31: p. 41-54.

[7] T.M.T. Mwebesa, V. Baryamureeba, and D. Williams, Collaborative framework for supporting indigenous knowledge management, in *Proc. of the ACM first Ph.D. workshop in CIKM*, 2007, ACM: Lisbon, Portugal. p. 163-170.

[8] D. Floreano and C. Mattiussi, *Bio-Inspired Artificial Intelligence: Theories, Methods and Technologies*, 2008: The MIT Press. 544.

[9] J. Yang and Y. Zhuang, An improved ant colony optimization algorithm for solving a complex combinatorial optimization problem. *Applications of Soft Computing*, 2010. 10(2): p. 653-660.

[10] R.S. Parpinelli, H.S. Lopes, and A.A. Freitas, Data mining with an ant colony optimization algorithm. *Transaction of Evolving Computing*, 2002. 6(4): p. 321-332.

[11] Z.J. Lee, Genetic algorithm with ant colony optimization (GA-ACO) for multiple sequence alignment. *Application of Soft Computing*, 2008. 8(1): p. 55-78.

[12] S. Schockaert, P.D. Smart, and F.A. Twaroch, Generating approximate region boundaries from heterogeneous spatial information: An evolutionary approach. *Jornal of Informatics* , 2011, vol. 181, p. 257-283.

[13] Z.S. Harris, Distibutional structure. Word, 1954, vol.10, p.62-146.

[14] L. Wu. and S.C.H. Hoi, Enhancing Bag-of-Words Models with Semantics Preserving Metric Learning. *IEEE MultiMedia*, 2011, vol.18. p. 24-37.

[15] R. Chau, A ConceptLink graph for text structure mining, in *Proc. of the Thirty-Second Australasian Conference on Computer Science* - Volume 912009, Australian Computer Society, Inc.: Wellington, New Zealand. p. 141-150.

[16] P. Ordonez, Using Modified Multivariate Bag-of-Words Models to Classify Physiological Data, in *Proc. of the 2011 IEEE 11th International Conference on Data Mining Workshops*, 2011, IEEE Computer Society. p. 534-539.

[17] J.F. Sowa, *Conceptual Graphs*. 1984.

[18] T. Wang, Automatic extraction of hierarchical relations from text, in *Proc. of the 3rd European conference on The Semantic Web: research and applications*, 2006, Springer-Verlag: Budva, Montenegro. p. 215-229.

[19] V. Nebot and R. Berlanga, Semantics-aware open information extraction in the biomedical domain, in *Proc. of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences*2012, ACM: London, United Kingdom. p. 84-91.