KNAPSACK BASED ACCS INFORMATION RETRIEVAL FRAMEWORK FOR BIO-MEDICAL LITERATURE USING SIMILARITY BASED CLUSTERING APPROACH.

[1]K.Latha [2]S.Archana [2]R.John Regies [3]Dr. Rajaram

[1]Lecturer of Information Technology Department,
[2]Students of Information Technology Department,
[3]Head of Computer Science and Engineering Department,

Thiagarajar College of Engineering,
Madurai - 15, Tamil Nadu,
India (South).

ABSTRACT

This paper investigates the effect of convergence of the documents in the same cluster which degrades the performance of the SA and Proposed knapsack. For this purpose, we have introduced a dynamic cluster selection mechanism of ACCS which enhances the performance.

KEY WORDS: Medline, Heuristics, Artificial Intelligence

1. INTRODUCTION

A vast amount of new biological information is available in electronic form on a regular basis. Medline [4] and Pubmed[10],[11] contains over 10 million abstracts, and approximately 40,000 new abstracts are added each month. Information retrieval is concerned with identifying, within a large document content which is most relevant to a user's need. The task has to be performed accurately and efficiently. The objective of this paper is to build a text mining framework, which consists of four distinct stages such as 1.Text Gathering 2.Text Preprocessing

3.Clustering based on similarity approach and 4.Information retrieval process using three different methodologies such as 1. Proposed Knapsack based Ant colony Cluster Selection (ACCS) 2.Proposed Knapsack and 3.Simulated Annealing Approach. The above three methodologies are presented to resolve Retrieval problems in IR. Experimental results on standard Retrieval benchmarks demonstrate the ACCS is more direct, easy to implement, and more efficient than previous methods**.** The goal of this paper is to enable effective document retrieval using various optimization techniques, to extract knowledge from such documents and to maximize the average fitness. Initially the documents are clustered using similarity based approach [3]. Our Proposed ACCS has been embedded with the Proposed Knapsack [3] which enhances the performance.

2. MATERIALS AND METHODS:

2.1 Ant Colony optimization:

A finite size colony of artificial ants searches collectively for good-quality solutions to the optimization problem. Each ant builds a solution or a component of solution, starting from initial state. Prabhaharan [9] proposed a method with the objective of minimizing the total cell load variation and the total intercellular moves. Oscar cordon [5] proposed the ideas of ACO approach that lead from Biological Inspiration to the ACO Meta Heuristics gives a set of rules of how to apply ACO Algorithms to the Challenging Combinatorial Problems. Rafael S. Parpinelli [12] proposed an algorithm for data mining called Ant-Miner (ant-colony-based data miner). The goal of Ant-Miner is to extract classification rules from data. The algorithm is inspired by both research on the behavior of real ant colonies and some data mining concepts.

2.2. Simulated Annealing:

This is a technique to find a good solution to an optimization problem by trying random variable variations of the current solution. A worse variation is accepted as the

new solution with a probability that decreases as the computation proceeds. The slower the cooling schedule, or rate of decrease, the more likely the algorithm is to find an optimal or near optimal solution. Arindam K. Das [1] mentioned that broadcasting in wireless networks, unlike wired networks, inherently reaches several nodes with a single transmission. For omni directional wireless broadcast to a node, all nodes closer will also be reached. This property can be used to compute routing trees which minimize the sum of the transmitter powers. This paper presents a mixed integer programming formulation and a simulated annealing algorithm. E. P. Zafiropoulos [13] discusses about reliability and cost optimization of power electronic devices, considering the component failure rate uncertainty. This paper presents an efficient methodology to solve the nonlinear reliability and cost optimization problem of power electronic devices, with complex structures under reliability and cost constraints.

2.3. Proposed Knapsack:

Latha [3] proposed a new algorithm with additional constraints. This paper applies text mining to extract knowledge from documents, and to minimize the number of documents to be checked. Initially the document is selected randomly, and then the fitness value is computed. This fitness value is compared with the user defined value and the document is accepted if and only if the fitness value is greater than or equal to the user defined value, otherwise rejected. This process is repeated until the number of documents is equal to the total number of documents to be displayed. For each iteration the average fitness value is calculated and will be replaced by the old fitness value and is considered for further evaluations.

3. PHASES OF TEXT MINING FRAMEWORK

3.1. Text Gathering

The process of text gathering in Biomedical literature involves PubMed Open Access Initiative [10], [11], MedLine [4], National Library of Medicine [4], MeSH databases, MDB etc. All the above databases contain more than 12,000,000 references of biomedical publications. We have used PubMed [10],[11] and other Biological databases to obtain large cluster of full text documents, all containing relevant results. The basic step is to download the documents and to prepare them for further stages. We have downloaded 1000, 750 and 500 sample set of documents from three biological domains such as Tuberculosis, Cancer, and AIDS. All the documents are stored in a separate centralized directory from where the documents can be processed and retrieved. The Biologists and the researchers who have administrative rights can upload their documents by using our interface. Most of the journals and research publications are encoded in PDF format [7]. It is necessary to convert these PDF documents into the textual documents for preprocessing and other phases

3.2. Text Preprocessing

Text Preprocessing transforms text into an information rich document matrix. This method identifies the frequency of occurrence of the keyword in the document collection. "80-20% rule" (i.e.) 80% of the work is done by the preprocessing stage and 20% of the work is done by the remaining stages. Text preprocessing is to prepare the data for the Clustering and Information retrieval process. It is the most important phase because if the data is not properly preprocessed it will be reflected in the remaining phases.

The basic steps [2], [6], [8] are

- Tokenization
- Data cleaning
- Stop word removal
- Stemming (Paice/Husk)
- Identification of most interesting terms.

3.3. Clustering (Similarity Based Approach)

Similarity based approach [3] for clustering is to classify or to group the objects based on attributes or features into K number of groups. K is a positive integer number. The grouping is done by maximizing the similarity between the document and the cluster centroid. Here the centroid is computed by the similarity measure not based on the distance metrics. The output of this phase is a set of clusters which is given as the input to the fourth phase (Information retrieval process).

The number of documents for each cluster is identified by similarity based clustering approach and the results are shown below.

Table 1: Similarity based clustering results

No of docs	No of Iterations	Cluster Id	No of Docs for each cluster	No of Clusters
1000	3	1	198	3
		2	371	
		3	431	
750	3	1	88	3
		2	288	
		3	374	
500	4	1	63	3
		2	194	
		3	243	

3.4 INFORMATION RETRIEVAL PROCESS

In this phase we have applied three methodologies such as (i) Proposed Knapsack based ACCS (ii) Proposed Knapsack and (iii) Simulated Annealing. The results comparing the Fitness and Cluster Selection are reported in Table 2.The documents are clustered by Similarity based approach and is given as the input to all the above three phases.

3.4.1. Knapsack based ACCS:

Ant Colony Optimization is based on the behavior of real ants searching for food. Real ants communicate with each other using an aromatic essence called pheromone, which they leave on the paths they traverse. If ants sense pheromone in their vicinity, they are likely to follow that pheromone, thus reinforcing this path. The pheromone trails reflect the 'memory' of the ant population.

The quantity of the pheromone deposited on paths depends on both, the length of the paths as well as the quality of the food source found. Initially the Pheromone trail is fixed as 0.1. The probability of choosing a cluster P_{ij} is computed for each cluster and this is based on the parameters τ_{ij} and η_{ij} and is given in equation (1).

$$P_{ij} = \begin{cases} \dfrac{\tau_{ij} \cdot \eta_{ij}}{\Sigma [\tau_{ij}].[\eta_{ij}]} & \text{, if } j \in \\ \end{cases} \quad (1)$$

Where $\eta_{ij} = \text{Max } d_{ij} - d_{ij}$ and $\in$ is the feasible neighborhood of ant k for the Keyword i. A Random number is generated and the path is selected based on the probabilities. To increase the density of pheromone trail for the selected path is computed by equation (2)

$\tau_{ij} \longleftarrow \tau_{ij} + \Delta \tau^{k}$ where $\Delta \tau^{k}$ is a constant. (2)

To evaporate the pheromone trails for all the paths.

$\tau_{ij} \longleftarrow$ **(1-ρ)** τ_{ij} where ρ $\in \{0,0.01,0.1\}$ (3)

if ρ =0 no pheromone evaporation takes place. An evaporation rate of ρ = 0.1 is large because evaporation takes place at each iteration of the ACCS. ACCS and proposed knapsack [3] are applied simultaneously to retrieve the documents. For each iteration a set of documents are retrieved from the cluster. Then ACCS is applied only for the remaining set of documents in the cluster. This process is repeated until the required documents are retrieved.

3.4.2. Proposed Knapsack with additional constraints:

In this paper we propose our own algorithm (Modified Knapsack with additional constraints). The documents are selected randomly from the appropriate cluster in such a way that the weight of the keyword is maximum. Fitness value is computed for each document. The document is selected if the fitness value is greater than the threshold limit. This threshold value is dynamically changed by the Tuning Coefficient (TC) [3].

3.4.3. Simulated Annealing Optimization:

A technique to find a good solution to an optimization problem by trying random variable variations of the current solution. A worse variation is accepted as the new solution with a probability that decreases as the computation proceeds. The slower the cooling schedule, or rate of decrease, the more likely the algorithm is to find an optimal or near optimal solution.

An annealing algorithm contains four basic components:
1. Configurations,
2. Move set,
3. Cost function,
4. Cooling schedule.

Pseudo-code: Simulated Annealing Optimization

Select an initial state i € S;
Select an initial temperature T > 0;
Set temperature change counter t = 0;
Repeat
Set repetition counter n = 0;
Repeat
Generate state j, a neighbour of i;
Calculate δ = *f (j)* - f (i) ;
If δ < *0 then i* := j
else if random(0, 1) < exp(- δ */T) then i* := j;
n:=n+l;
until n = N(t);
t : = t + l ;
T:= T(t);
until stopping criterion true.

---

4. BENEFIT OF THE CHOSEN METHODOLOGIES:

4.1. SAO:

- SAO [13], [1] is a computational optimization search technique which produces an optimal or near optimal solution with less computational time.
- Heuristic search techniques are reliable and effective for solving combinatorial optimization problems which are known to be NP-hard.

4.2. Proposed Knapsack vs. Existing Knapsack:

- The dynamically changing behaviour is attained by the specific coefficient TC (Tuning Coefficient).
- Number of iterations is less compared to Existing Knapsack.
- Optimal feasible solutions with maximum relevancy are obtained.

4.3. Knapsack based ACCS:

- Clusters are selected dynamically.
- The value of the pheromone trail decreases when the number of relevant documents decreases.
- Cluster selection has been done, instead of path selection in existing ACO.
- Proposed Knapsack is efficient for the retrieval of documents and ACCS is efficient for cluster selection. Document retrieval is optimized by the combination of two distinct methodologies (Knapsack and ACCS).

Table 2: Comparison of SAO, Proposed Knapsack and Knapsack based ACCS

Key	No of Docs	No of Doc to be displayed	Cluster id			Fitness		
			SA	Knap	ACCS	SA	Knap	ACCS
Cell	500	25	3	3	3	98.12	100.00	100.00
AIDS	500	25	2	2	1,2,3	71.83	73.00	81.319
Tuberculosis	500	25	3	3	2,3	71.27	96.00	98.031
Cell	750	38	3	3	3	96.71	100.00	99.645
AIDS	750	38	2	2	1,2,3	73.10	73.00	76.757
Tuberculosis	750	38	3	3	2,3	73.62	95.00	95.455
Cell	1000	50	3	3	3	91.13	100.00	99.694
AIDS	1000	50	2	2	1,2,3	73.20	72.00	76.205
Tuberculosis	1000	50	3	3	1,2,3	76.4	96.00	96.018

EXPERIMENTAL RESULTS:

Results of three algorithms for the 1000,750 and 500 sample data set are shown in Table 2. It can easily seen from Table 2 that the quality of cluster selection in ACCS is better than SA and Proposed Knapsack. Among the 500 sample set of documents, for the keyword tuberculosis the fitness values are 71.27, 96.00, and 98.031 for SA, Proposed Knapsack and ACCS respectively. Significant improvements were observed for the ACCS. In ACCS the documents are selected globally among the clusters whereas in SA and Proposed Knapsack, document selection is localized within a cluster. For keyword tuberculosis the documents are retrieved from cluster number 3 in SA and Proposed knapsack, but in ACCS the retrieval is from the clusters 2 and 3 in fig 2, 3 and 1 respectively.

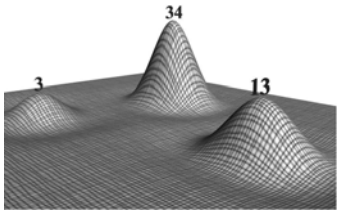

Figure 1: Knapsack based ACCS

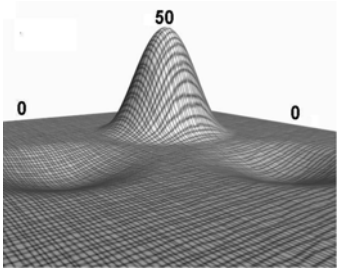

Figure 2: Simulated Annealing

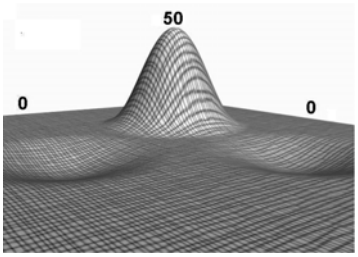

Figure 3: Proposed Knapsack

Fitness for chosen methodologies:

Figure 4 shows the results comparing three algorithms based on their fitness rate. We see that introduction of ACCS improves the performance of fitness on the whole. Optimal solution was found over a wide range of random selection rate and best solutions among three algorithms were obtained which shows the stability of the extended ACCS approach.

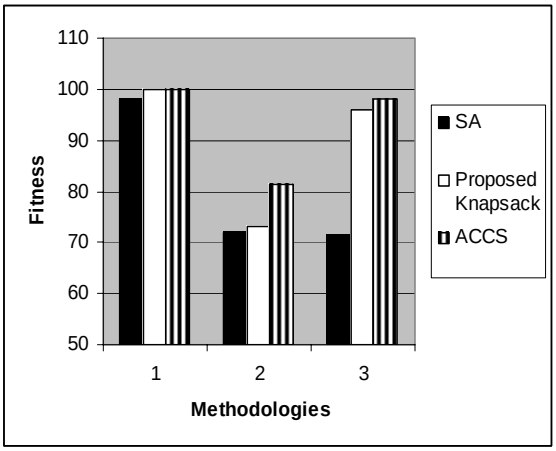

Figure 4: Comparison of SAO, Proposed Knapsack and Knapsack based ACCS

6. CONCLUSION AND FUTURE WORK:

Result from the above data sets show the superiority of our approach over the other methodologies. Two important directions for future research are as follows. First, the methodologies (SAO, Proposed Knapsack and Knapsack based ACCS) are based on random number generation, it is not guaranteed to produce better results all the time. Second, in ACCS methodology cluster selection is purely based on the probability of choosing a cluster, so there may be a chance of selecting a wrong cluster in the initial stage. The related work can be done to overcome the above issues in future. The work can be extended by applying other heuristic algorithms like genetic algorithms, evolutionary programming and simulated annealing with ACCS.

7. REFERENCES:

[1]Arindam K Das, Roberto Montemanni, Luca Maria Gambardella."The Minimum Power Broadcast Problem in Wireless Networks: a Simulated Annealing Approach." In *Proceedings of AIRO 2004: Annual Conference of the Italian Operations Research Society* Lecce, Italy, September 2004.

[2] Lancaster Paice/Husk Stemming algorithms www. lancs. ac.uk /ug/ oneillc1/ stemmer.

[3] Latha K, John Regies R, Archana.S, Rajaram R, "Knapsack based IR System for Bio-medical Literature using Similarity based Clustering Approach", *International Journal of Information Technology*, Faisalabad, Pakistan, 2007.

[4] Nelson S, 2004. "MedicalSubject Headings FactSheet",http://www.nlm.nih.gov/pubs/factsheets/mesh.html",U.S. National Library of Medicine.

[5] Oscar cordon, Fransisco Herrera,Thomas "A Review on Ant Colony Optimization Meta Heuristic: Basis, Models, and new Trends ",*Mathware &Soft computing 9*,2002.

[6] Paice C,Hooper R (2005), The Lancaster Stemming Algorithm. Available at: http:// www. comp.lancs.ac.uk/computing/research/ stemming/

[7] PDF Specification, http://partners.adobe. com/ asn/acrobat/sdk/public/docs/PDFReference15_v6.pdf,Adobe, 2004.

[8] Porter M.F, "An algorithm for suffix stripping (reprint)", in Readings in Information Retrieval, MorganKaufman, http:// www. tattarus. org/ ~martin/ PorterStemmer

[9] Prabhaharan G, Muruganandam A, Asoka P,Girish B.S."Machine Cell Formation for cellular Manufacturing Systems using an ant colony system approach", *The International Journal of Advanced Manufacturing Technology,* Volume 25, Numbers 9-10 / May, 2005

[10]PubMedCentralOpenAccessInitiative, http://www.pubmedcentral.nih.gov/about/openftplist.html,2004.

[11] PubMed, http://www.ncbi.nlm. nih.gov/ PubMed/, 2004.

[12] Rafael S. Parpinelli, Heitor S. Lopes, Member, IEEE, and Alex A. Freitas, "Data Mining With an Ant Colony Optimization Algorithm", *Sixth International Conference on Intelligent Systems Design and Applications (ISDA'06)* pp. 400-403,2002

[13] Zafiropoulos E. P and Dialynas E. N," Reliability and cost optimization of electronic devices considering the component failure rate uncertainty using Simulated Annealing", *Reliability Engineering & System Safety* , Volume 84, Issue 3, Pages 271-284 , June 2004