International Journal on Applications in Information and Communication Engineering
Volume 1: Issue 4: April 2015, pp 5-8. www.aetsjournal.com ISSN (Online) : 2394-6237

Ontology Based Semantic Information Retrieval Using Particle Swarm Optimization

Gomathi A, Jayapriya J, Nishanthi G, Pranav K S,Praveen Kumar G

Abstract - In the field of Information retrieval, user expects to get the very accurate result. Even though more approaches are effective but sometimes they provide less accurate information to the user specified query. With the development of technology, our project aims to improve the accuracy of the result with exact information. In this paper ontology based approach for query expansion and Particle Swarm Optimization (PSO) algorithm is used for data clustering. The PSO algorithm is also used to retrieve the nearest optimal solution with the help of the global best and local best variables.

Keywords - Information Retrieval (IR), Ontology, Particle Swarm Optimization, Query Expansion.

I. INTRODUCTION

Internet is being a vast collection of information. Users tend to search for the particular data which they are in need but the World Wide Web provides less relevant information from a query processed by the user. Therefore the efforts should be made to create such methodology which generates quality of results, instead of just storing mass of document. In today's scenario, the relevancy of web document is evaluated by matching each keyword of query with information on web. Semantic web [1] are becoming a good example establishing a relationship among the documents.

The Semantic web as "a web of data that can be processed directly and indirectly by machines". Ontology [2] defines accepted concepts and a relationship in some specific domain. Ontology capture the structure of the domain, i.e. conceptualization. The conceptualization describes knowledge about the domain, not about the particular state of affairs in the domain. Ontology consist of formal description is normally written in a language like RDF or OWL, so that "detailed, accurate, sound and meaningful distinctions can be made among the classes, properties, and relation". Ontology has the concepts and properties to define the knowledge about the domain.

Swarm Intelligence (SI) is an innovative distributed intelligent paradigm for solving optimization problem that originally took is information from the biological examples by swarming, flocking and herding phenomena in the vertebrates. Particle Swarm Optimization [3] incorporates swarming behaviors observed in flocks of birds, schools of fishes or worms of bees, and even human social behavior, from which the idea is emerged. PSO is a population based-based optimization algorithm, which could be implemented and applied easily to solve various function optimization problems. As an algorithm, the main strength of PSO is its fast convergence, which compares favorably with many global optimization algorithms like Genetic Algorithms (GA), simulated (SA), and other global optimization algorithms [4].

This paper presents an approach of finding relevant documents by using the semantic similarity and query expansion. Semantic similarity between concepts is determined by applying PSO. The theme of proposed technique is based on searching and extracting precise information from user defined knowledge base.

Rest of the paper is organized as: Section 2 provides the related work; Section 3 presents the proposed system consists of query expansion, Calculating semantic similarity using PSO. Finally section 4 concludes the paper.

II. RELATED WORK

Many approaches have been proposed by a number of researches in information retrieval. In the proposed approach the information retrieval can be manipulated by semantic based on ontology. An ontology based Information retrieval model, as in large repository of documents the semantic search can be supported by domain specific knowledge base. Xing Jiang et al. have proposed user ontology, it is an ontology based model which provides personalized information services which is used in semantic web. In this ontology concepts, taxonomic relationship and non-taxonomic relationship for a given domain ontology is used to assume the user interest [5]. Another research done by A. Kiryakov et.al is based on producing architecture for Indexing, semantic annotation, and extracting documents with respect to repositories based on semantic [6]. A survey on existing research activities in this field have shown various applications for information retrieval such as: query expansion used in graph based approach focusing on multi-document summarization [7]. The search can be improved by Query Expansion (QE). QE can be simply classified into three groups: interactive QE, semantic dictionary QE and the QE method based on document set.

Gomathi A is Associate Professor/Computer Science and Engineering, , Narasu's Sarathy Institute of Technology, Salem-636 305 (email : sengomsuji@gmail.com)

Jayapriya J, Nishanthi G, Pranav K S, Praveen Kumar G are UG Scholars (B.E - Computer Science and Engineering) , Narasu's Sarathy Institute of Technology, Salem-636 305 (email : jayapriyaj@hotmail.com, spranavcs@gmail.com)

III. PROPOSED WORK

In this paper we have suggested a technique to get a nearest optimal solution based on the user query using ontology based query expansion and PSO.

A. *Overall Architecture*

In the proposed architecture, when the user enters the query, if the query term is in ontology then QE module expands it by listing its equivalent classes by using reasoner.

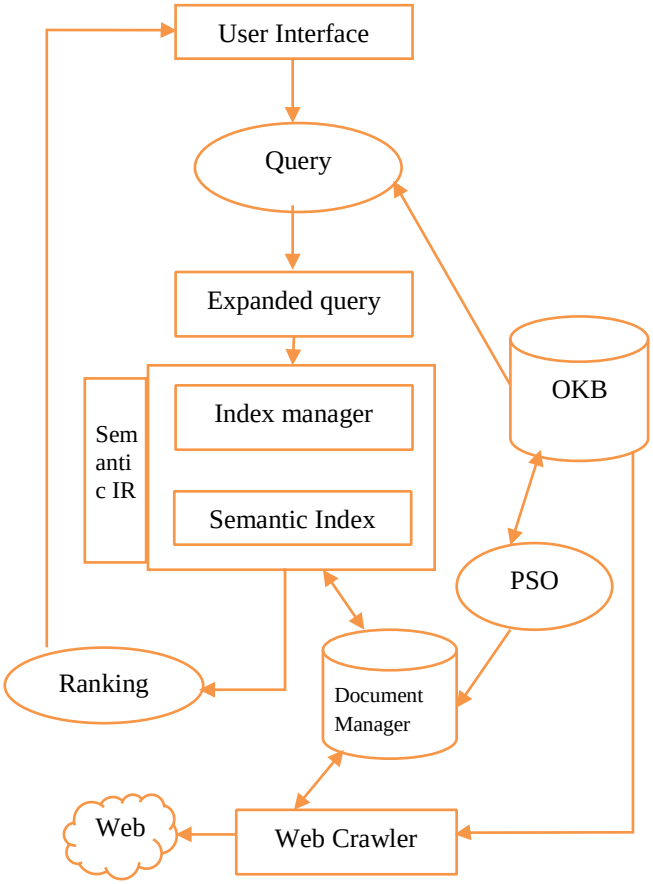

Fig 1: Overall Architecture for the proposed system

Then expanded query is used to extract the information through semantic IR. The semantic index is created with the help of Particle Swarm Optimization and domain ontology.

The steps involved in this approach are:

i. Create or reuse an existing ontology.
ii. Web documents are extracted using web crawler for the concepts in the Ontology.
iii. The Semantic terms are extracted by applying PSO clustering algorithm to the classes found in fetched document and ontology.
iv. If a user enters the query, the query is expanded and the results are retrieved with the nearest optimal solution.

B. *Query Expansion*

The QE module expands the query based on domain ontology. If the concept is a class or instance in ontology then it expands by the following steps:

i. Query from user: UQ.
ii. If the UQ is a class in ontology then subclasses, equivalent classes and sub instances are retrieved.
iii. Else if UQ is an instance in ontology then list the sub instances of the class to which the instance belongs.
iv. Else delivers UQ to direct retrieval.

When the user enters a query in a search box which was present in the user interface. If the user entered query is a class in ontology then the subclasses, equivalent classes and instances of the particular class will be retrieved. Suppose, if the user entered query is an instance in ontology then list the sub instances of the class are retrieved which was belongs to the particular instance. Otherwise user query will be delivers to retrieval.

C. *Particle Swarm Optimization*

Document clustering is widely used in information retrieval as well as text mining. PSO is one of the document clustering algorithm to find nearest optimal solution [11]. It is a population based stochastic optimization technique which was introduced by Kennedy and Eberhart, it was based on the swarming behavior of animals and human social behavior [12]. A particle's location in the multidimensional problem space corresponds one solution for a particular problem. When a particle moves from one location to other location, a different solution is generated for a problem. This solution is examined by a fitness function which provides a quantitative factor of the solution that is called as fitness value. A particle will remember its velocity and its current coordinates which shows the speed of its mobility among the dimensions in a problem space.

Initially, the PSO algorithm chooses solutions randomly within the search space [10]. The fig 2 indicates initial state of a four particle which was having global maximum in a one dimensional search space. Here the curve denotes the objective function. The objective function which was used to examine its candidate solution, and thus will operates upon the resultant values.

Candidate solution means fitness value for a particular

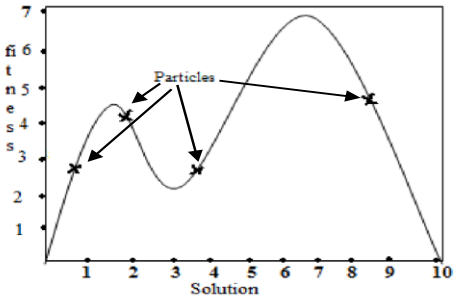

Fig 2: Initial PSO State

particle. Each and every particle maintains its candidate solution and its corresponding fitness values, its position, and its velocity. In additionally, it remembers the best fitness value it has achieved throughout the process of the algorithm. This is generally called as individual best fitness. The Candidate solution that attained this fitness evaluation is called as individual best solution or individual best candidate solution. At lastly, the PSO algorithm which maintains the best fitness value achieved between all particles in the swarm called as global best fitness and the candidate solution which was achieved this fitness called as global best position or global best candidate solution [10].

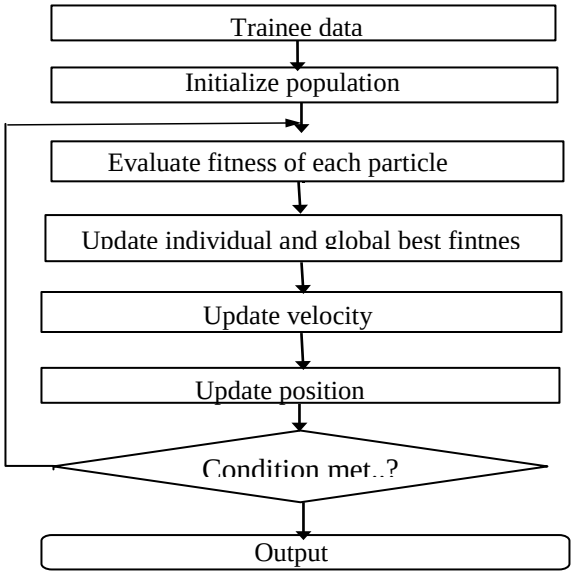

Fig 3: Flowchart of PSO

Fig 3 shows the flow of a Particle Swarm Optimization algorithm in a step by step process.

Initially, it will initialize a population of particles with their random positions and velocities in search space. Secondly, find the fitness value for the each particle in the swarm. For the each particle update the individual and global fitness value. Then update the velocity and its position according to the fitness value which was evaluated earlier, repeat this process until the termination conditions are met. Finally, this will produce the nearest optimal solution for the particular input [12].

1) Document Representation

Each text document can represented by Vector Space Model (VSM) [8]. In this model, the document is considered as a dot in multi-dimensional space and represented by a vector x, then $x = \{w_1, w_2, \ldots, w_t\}$ where w_i is the term weight of the term t_i in one document and it is used to represent the importance of the term in that document. In order to calculate the term weight, the term occurrence frequency within the document and the entire document set is considered. Term frequency and Inverse document frequency (TF-IDF) [8] is commonly used to calculate the term weight. The weight of term i in the j document is given by

$$x\,ji = tfji * \log 2\,(n/dfji) = tf_{ji} * idf_{ji} \qquad (1)$$

Where,

tf_{ji} is the number of occurrence of term i in the document j, idf_{ji} represents the inverse document frequency and it is calculated as $\log_2 (n/df_{ji})$ where n is the total number of documents and df_{ji} represents the term frequency in the document set.

2) Similarity measure

The similarity between the documents is find using the cosine similarity, which is to be used in clustering of documents. The Euclidean distance is one of the method to find the similarity metric between two documents m_a and m_b, by considering that the distance will vary according to the dimension number. In this approach normalized Euclidean distance function is used to manipulate the equivalent threshold distance:

$$d(m_a, m_b) = \sqrt{\sum_{k=1}^{d_m} (m_{ak} - m_{bk})^2 / d_m} \qquad (2)$$

Where,

m_a and m_b are the two document vectors, d_m is the dimension number of vector space, $m_{ak}$and m_{bk} are weight values in the dimension k for the documents m_a and m_b.

3) PSO Clustering

The aim of PSO clustering is to find the centroid of cluster, because it minimizes the inter-cluster distance and also it increases the distance between the clusters. A swarm

represents the number of solution for the candidate clustering. To evaluate the solution for each particle the fitness value is used, and it is the average distance between a cluster center and the document. The fitness value is measured by the following equation.

$$f = \frac{\sum_{i=1}^{N_c} \left\{ \frac{\sum_{j=1}^{p_i} d(o_i, m_{ij})}{p_i} \right\}}{N_c} \qquad (3)$$

Where,

o_i denotes the centroid of the i^{th} cluster, m_{ij} is the j^{th} document vector belongs to cluster i, p_i is the document number belongs to the cluster C_i, N_c stands for cluster number, $d(o_i, m_{ij})$ is the distance between document m_{ij} and the cluster centroid O_i.

The PSO clustering can be summarized as:

1) Initially n number of document from document vector is taken as a centroid for a cluster.
2) For each document:
   i. Each document vector in the document set is assigned to the closest centroid.
   ii. Fitness value is find using equation 3.
   iii. Then the velocity and particle position are updated for each document (particle).
3) Step 2 is repeated until one of the following termination condition occurs:
   i. Until the maximum iteration exceeds.
   ii. The predefined value of centroid cluster is greater than the average value.

IV. CONCLUSION

In this paper we have presented an approach for PSO based clustering of documents for semantic information retrieval. We have provided that the optimal information can be retrieved with ontology based approach using semantic index. We have presented an approach for domain specific. In future the approach may be generalized for various domains combined together.

REFERENCES

T. B. Lee, J. Hendler and O. Lassila, "The Semantic web", Sci. Am., vol. 284, pp. 34-43, 2001.

P. Paggio, B .S. Pedersen, and D. Haltrup, Applying Language Technology to Ontology-Based Querying: the Ontoquery Project, Applied Artificial Intelligence. Volume 17, Number 8 & 9, pp.817-883, 2003.

R. Eberhart and J. Kennedy, "Particle swarm optimization", Proc. of the IEEE Int. Conf. on Neural Networks, Piscataway, NJ., 1995, pp.1942-1948.

K. O. Jones, "comparison of genetic algorithm and particle swarm optimization for fermentation feed profile determination", Int. Conf. on Computer Systems and Technologies, 2006.

Xing Jiang and A.H. Tan, "Learning and inferencing in user ontology for personalized Semantic Web search", Information Science, vol.179(16), pp.2794-2808, 2009.

A. Kiryakov, B. Popov, I. Terziev , D.

Manov and D. ognyanoff, " Semantic annotation, indexing, and retrieval," Web Semantics: Science, Services and Agents on the World Wide Web, vol.2(1), pp.49-79, 2004.

L. Zhao, L. Wu, X. Huang, "Using query expansion in graph-based approach for query focused multi-document summarization", Information Processing and Management Journal, Vol.45, pp. 35-41, 2009.

Everitt, B., 1980. Cluster Analysis. 2nd Edition. Halsted Press, New York.

Kajal Joshi, Ashish Verma, Ankita Kandpal, Shalini Garg, "Ontology based Fuzzy Classification of Web Documents for Secmantic Information Retrieval", 2013.

James Blondin, "Particle Swarm Optimization: A Tutorial", September 4, 2009.

Xiaohui Cui, Thomas E. Potok, "Document Clustering Using Particle Swarm Optimization". 12. Pritesh Vora, Bhavesh Oza, "A survey on K-mean Clustering and Particle Swarm Optimization", ISSN: 23196386, Volume-1, Issue-3, February 2013.