Web Documents Ranked using Genetic Algorithm

Poonam Chahal
Research Scholar
YMCAUST,
Faridabad

Manjeet Singh
Associate Professor
YMCAUST
Faridabad

ABSTRACT

In recent years, semantic-based application has been using the Genetic Algorithm for solving the problem of information retrieval. The semantic search concept has been widely used in many fields like artificial intelligence, cognitive science, natural language processing, and psychology. In this paper we have proposed a ranking scheme for the semantic web documents by finding the semantic similarity between the documents and the query which is specified by the user using genetic algorithm. The novel approach proposed in this paper considers not only the conceptual level for finding the rank score of a document but also the descriptive level for finding the true semantics of the documents keeping the user view in mind. The combined use of conceptual and descriptive along with the genetic algorithm for providing the optimization for the ranking of the documents has significantly improved the performance of the proposed ranking scheme. We explore all the conceptual and descriptive scores using genetic algorithm for finding relevant relations between the keywords exploring the user's intention and then calculate the fraction of these relations on each web page to determine their relevance with respect to the query provided by the user.

Keywords

Semantic Web, Ranking, Parsing, Syntactic, Lexical, Semantic, Similarity.

1. INTRODUCTION

GA (Genetic Algorithm) has been widely used for solving the optimization problems in information retrieval which has been the topic of research for last few decades [2]. The World Wide Web (WWW) is large information resource centre in which information present in the form of web pages is interlinked with each other [1][9].

A GA is a variant of stochastic beam search in which successor states are generated by combining two parent states, rather than by modifying a single state.

Basic Algorithm:

Step 1. Initialization: GA begins with randomly generated states called, population. Each state or individual is represented as string over a finite alphabet.

Step 2: Selection: Select a pair of string from the population by calculating the fitness function.

Step 3: Reproduction:

The production of next generation of states is rated by fitness function.

a. Apply Crossover operation to the selected pair of strings to generate two offspring.

b. Apply Mutation to each offspring with small independent probability.

Step 4: Termination: If a pre specified condition is not satisfied, return to step 2. Otherwise terminate the algorithm.

Genetic Algorithm is inspired by the approach of selection is based on the "Survival of Fittest". With the increasing and enormous amount of information which is electronically present on web, the major area of research is in the field of information retrieval [11]. The users of the internet are facing difficulty in finding the relevant information because the result-set produced by the traditional search engine also includes irrelevant web pages. So, we have tried to propose a solution to ranking of the search engine. Our ranking model includes the approach of genetic algorithm in which the web page which is semantically more relevant to the query will be given to the users of the internet to increase the efficiency of the searching information from web. In section II, various approaches followed and proposed by various researchers is discussed, in section III the algorithm for our ranking approach using genetic algorithm is given. Finally in section IV performance analysis of our approach is given which is followed by conclusion and future scope in section V.

2. RELATED WORK

In fact, the process of information retrieval by a search engine is very crucial [10]. In semantic web various semantic search engines have been developed like Ontolook, Swoogle, Watson etc to help in performing the information retrieval from the semantic web[4][5].

Yuxin Mao [16] has given A Semantic-Based Genetic Algorithm for Sub Ontology Evolution. The author gave the formal semantics of ontology to improve genetic algorithm in several aspects for making it suitable for semantic-based problems. The author presented a semantic-based genetic algorithm to incorporate domain knowledge into the algorithm and perform evolution based on the ontology semantics.

Razib M. Othman et al. [9] has presented incorporating semantic similarity measure in Genetic Algorithm: An approach for searching Gene Ontology terms. The authors discussed that the drawbacks of the process of searching the terms using traditional approach of matching the keywords only. The traditional approach ignores the semantic relations between terms so for this they proposed a combined approach of using the semantic similarity and the genetic algorithm for giving better retrieval process in searching the terms which are semantically similar. The semantic similarity measure is used to compute similitude strength between two terms. Then, the genetic algorithm is employed to perform batch retrievals and to handle the situation of the large search space of the Gene Ontology graph.

Mehrnoush Shamsfard et. al. [12] have given a method of ORank: An Ontology Based System for Ranking Documents. In their paper the authors proposed the new method of ranking by determining semantic similarity based on structure-knowledge extracted from ontology. The approach given by the authors exploits natural language processing techniques for extracting phrases and stemming words. Then the authors used the ontology based conceptual method to incorporate the semantics by annotating the documents and also expand the query. The spread activation algorithm is used and improved for the expansion of the query so that it can be done in various aspects. Finally, the annotated documents and the query which is expanded can be used for processing to compute the relevance degree by exploiting the statistical methods.

Thiagarajan R. et. al [14] proposed computing semantic similarity using ontology's. In this paper authors represented the web page can be represented either Bag of Concepts (BOC). In BOC the concepts are taken from the web page to represents the web page more semantically. Now for computation of semantic similarity between the web pages the authors used process of spreading, which means including additional related terms to an entity by referring to ontology such as Word Net, Wikipedia. For the spreading process two schemes are used one is set spreading and other is semantic network.

B. Hajian et. al. [6] proposed a method of measuring semantic similarity using a multi-tree model measuring semantic similarity using a multi-tree model. In this paper the authors proposed the new method for semantic similarity based on the knowledge that is extracted from ontology and taxonomy. The technique described uses multi-tree similarity algorithm to measure similarity of two multi-tree constructed from taxonomic relations between entities in ontology. The similarity comparison is done by comparing the feature list representing the concept, i.e. each concept in the approach given is represented by the features describing its properties. Though the authors considered the edges representing the relationship between the concepts but they did not considered the number and the type of relationship which can exist between two concepts in the ontology or the document.

Lamberti F. et. al. [7] used relation-based page rank algorithm for semantic web search engines to focus on the concepts that exist within the document and also on the relationship that exists between the concepts. The authors fully explored the number of relationships that exists between the concepts in a document with respect to the number of relations that are presented in the given ontology between same concepts. They proposed a technique to exploit the relevance feedback and post process result-set to develop a ranking strategy which considers relation between keywords which is given in web page. A page rank algorithm which is based on relations that exists between the concepts is given which can be used in conjunction with the semantic web search engine. The approach is to construct the graph of underlying ontology, query, page annotation and page sub graph. Then they computed probability for a page to be selected by taking the factors of number of relation in ontology, query and page annotation and sub graph. This considers the ontology graph, query graph, and annotation of page and its sub graph. Now, some concepts can be possible which are not related to any other concept in the annotations but that can be of user interest. So, the probability that each concept is related to other concept is modeled using graph theory i.e each concept related to at least another concept in the query is equivalent to considering all possible spanning trees.

Wang Wei et. al. [15] has given a paper on Search with Meanings: An Overview of Semantic Search Systems. In this paper the authors discussed the comprehensive survey on overall view of current research trends in the field of semantic search. The authors report their study and findings based on which a generalized semantic search framework is given. Further, they describe issues with regards to future research in this area.

In all these contributions given by various researchers the focus is on introducing the semantics for finding the semantic relatedness either by taking ontology [3][8] or relationship that exists between the concepts using genetic algorithm to optimize the information retrieval process [13]. To make ranking of the documents by the search engine having only the relevant pages in the result-set it is necessary to compare the complete semantic similarity between the documents and the query given by the user. We have attempted this and have achieved very good results.

3. PROPOSED APPROACH FOR FINDING SEMANTIC SIMILARITY USING GENETIC ALGORITHM.

Normally, the search engine analyses the text of web pages by extracting keywords/concepts to determine the relevance of the page with respect to the query given by the user. To find the semantic similarity of the web page with the query by considering user view, we have analyzed the document both at conceptual level and as well as at descriptive level to determine the true semantic relevance of the page to the query. Both these aspects are important but their relative importance may vary from sentence to sentence. We have therefore introduced the concept of assigning weights to concepts and as well as description aspect of the document.

In the proposed technique each sentence in the document is analyzed deeply at both the levels. At conceptual level the weight is assigned to each sentence with the help of the words which are present in the sentence and also in the query. The similarity value at conceptual level is computed using the cosine similarity function as

$$\mathrm{sim}_{\mathrm{conce\ ptual}}\,(\mathrm{e_j}, \mathrm{e_k}) = \frac{\vec{\mathrm{V}}(\mathrm{e_j}).\vec{\mathrm{V}}(\mathrm{e_k})}{|\vec{\mathrm{V}}(\mathrm{e_j})||\vec{\mathrm{V}}(\mathrm{e_k})|}$$

At descriptive level the weights are assigned depending upon the description i.e. considering the relation between the concepts identified at the conceptual level which are present in the sentence and user query. Usually, the description of any concept/word/sentence can be given in several ways, mainly by the type and number of relations that exists between the concepts. It means at conceptual level, lexical matching approach is followed which is accompanied by the semantic approach which is added during descriptive level. The algorithm for our approach is given in Algorithm 1.

The working of our approach is done by assigning the weights during the conceptual and descriptive level which are manipulated using two weights w1 and w2. The value of these weights is in the range between 0 and 1. These values are determined using the Genetic Algorithm which helps in finding the optimal solution to the problem. The use of genetic algorithm helps in improving the ranking of the web documents providing meaningful documents in the result-set and giving high correlation to the human rating.

In our approach initially the weights w1 and w2 are taken at equal level i.e 0.5 and the constants helps in finding the actual semantic score of the document which is found by multiplying the w1 with the conceptual level weight for the sentence and w2 with the descriptive level weight for the same. Finally both the values obtained in the previous steps are added for finding the sentence score. The paragraph value is obtained by using the statistical approach and similar approach is used to find the semantic score of the document giving the relevancy of the document with respect to the query. After getting the score for each of the document they are ranked accordingly.

```
Algorithm 1:
Input:Set of documents (S), Query (Q) & Ontology (O).
Output: Sim (D,Q) // where D is the Document present in set S and Q is the user query.
Begin Process
1. For each sentence $S_i$ in document $D_i$
   a. Extract all the concepts.
   b. Extract all the relations between the concepts that are extracted in 1(a) using ontology.
   c. Compute $Wt(r_{ij})=Dis(C_i, C_j)/DP(C_i)+DP(C_j)$
      // Where $Dis(C_i, C_j)$ is the shortest path between the concepts in the ontology and, $DP(C_i)$, $DP(C_j)$ are depth of the concepts in the ontology.
   d. Compute $Sim_{Descriptive}(S_i,Q)=\sum Wt(r_{ij})$
   e. Compute $Sim_{Descriptive}(D,Q)=\sum Sim_{Descriptive}(S_i,Q)/n$
      // where n is total number of sentences in the document.
   f. Then, compute final similarity
      $Sim(D,Q)=w_1*sim_{conceptual}(e_j, e_k)+w_2*Sim_{Descriptive}(D,Q)$.
```

Further, we will explain the detailed working of our model with the help of examples. We have taken set of seven documents from the domain education and the user query "what is education". These documents were first analyzed at conceptual level and the score were assigned at this level using lexical matching with respect to query. These documents were then followed by descriptive level analysis in which scores were assigned by extracting the relationship between concepts with respect to the user query. These scores obtained at conceptual and descriptive level were modified using two weights w1 and w2 using genetic algorithm. For example for the document D1 each sentence of the document are analyzed and the conceptual and descriptive weights computed are .7, .6 for s1, .8, .5 for s2, .9, .5 for s3 and so on. The weights are modified by setting the w1 and w2 (value between 0 and 1) 0.5 initially and then performing the number of iterations by modifying the w1 and w2 value using genetic algorithm to obtain the optimal semantic score for the document. After performing the iterations and following the statistical approach the final score for the document is obtained. Similar procedure is done for all the remaining set of documents and the final semantic score for each document provides the ranking of the document with respect to the user query. Table 1 shows the results of the set of documents ranked by using the proposed genetic approach.

Table 1: Results of Ranking of set of documents.

SNO	Weight score w1	Weight score w2	Rank for the set of documents with respect to user query	Variance from Actual Rank(D2,D1,D5,D4, D3,D6,D7)
1.	.5	.5	D1,D2,D5,D3,D4,D7,D6	6
2.	.7	.3	D1,D2,D5,D3,D4,D6,D7	4
3.	.6	.4	D1,D2,D5,D7,D3,D4,D6	16
4.	.2	.8	D1,D5,D3,D2,D4,D6,D7	26
5.	.8	.2	D1,D2,D3,D5,D4,D6,D7	8
6.	.1	.9	D5,D1,D3,D2,D4,D6,D7	18
7.	.9	.1	D2,D1,D5,D3,D4,D6,D7	2

The results obtained in table 1 for ranking are then compared to the human rating which provides the actual rank and the variance of each rank is computed to obtain the final ranking of the documents. The rank which is obtaining the minimum variance from the actual rank is scored as the final rank for the set of documents with respect to the query given by the user.

4. PERFORMANCE ANALYSIS

Performance of our approach for finding semantic similarity between the web documents definitely depends on how keywords and associated concept relations are extracted from the document, then on the process of spreading used to create and would depend from domain to domain as well as formulation of concept relations. We have compared the performance of our semantic similarity scheme using genetic algorithm for different set of pages related to a domain like education. From these pages we determined the actual similarity of the pages using the description and conceptual level for finding the similarity relevance. The score for the conceptual and descriptive level scores were computed at different number of iterations. The results obtained from the novel approach and have been presented in Table 2. The results in the table 2 shows the conceptual and descriptive score for the set of documents, and also the score modified at 5th and 9th iteration which is obtained using the genetic algorithm which helps in setting the value of the two defined weights w1 and w2. Finally the score obtained for the set of documents gives the semantic relevance of the document with respect to the user query which provides the ranking of the set of documents.

Table 2: Results of semantic similarity using Genetic Algorithm approach.

SNO	Set of Documents	Description Score	Conceptual Score	Score at 5th iteration	Score at 9th iteration	Final score for set of documents
1.	D1, D2	1.2,1.1	.90,.4	.7,.5	.78, .72	.76,.64
2.	D2, D3	1.1,1.25	.4,.6	.5, .31	.72, .40	.64,.37
3.	D3, D4	1.25, 1.35	.6,.75	.31, .53	.4, .65	.37,.62
4.	D5, D6	1.25,.8	.95,.4	.73, .6	.81, .76	.79,.72
5.	D6, D7	.8,.45	.4,.3	.6, .25	.76, .29	.72,.25

Note: D1:en.wikipedia.org/wiki/education.

D2:www.teach_kids_attitude_1st.com/definition of education.html.

D3:www.motivation_tools.com/youth/what_is_education.html.

D4:education.svtution.org/2011/06/what_is_education.htm.

D5:Dictionary.reference.com/brouse/education.

D6:psychology.about.com/od/educationalpsychology/educational_psychology.htm.

D7:press.chicago.edu/ucp/books/Chicago/w.html.

For deep analysis of the performance of our approach with lexical matching, we further looked to the pages retrieved from Google search engine having similar content but not represented with same words. The large number of PDF files giving similar content was taken and the summarization process applied to the documents and then the keywords extracted from the documents summarization and also along with their weight age. The terms which were extracted with similarity greater than the threshold value were taken into account for the process of finding semantic similarity using genetic algorithm. Then the similarity between the concepts and the relationship that exists between the concepts was then calculated using the approach given in this paper. We found that for maximum number of documents the approach produces good similarity measures. In each case the similarity computation of our method is much better than traditional similarity approach showing the superiority.

5. CONCLUSION

The semantic similarity concept using Genetic Algorithm between the semantic web documents and the user query improves the ranking by retrieving relevant web pages in the result-set produced by the search engine.

The ranking approach proposed in the paper takes the concept of Genetic Algorithm for finding the better result-set for the user query. The Genetic Algorithm in our approach helps in finding the optimal solution by considering the exhaustive approach. Our future efforts would be to design more meaningful and exhaustive ranking strategy by using the semantic analysis of web pages and by deeply statistical analysis relevance of documents, so that the semantic search engine can evaluate more precisely relevance and also the similarity between the web page and the user query. We will also try to make our approach scalable for the semantic web.

6. REFERENCES

[1] Berners-Lee T., Hendler J., and O. Lassila, “The Semantic Web,” Scientific Am., 2001.

[2] Brin S. and L. Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine”, Proc. Of 7th Int’l Conf. on World Wide Web(WWW ’98), pp. 107-117, 1998.

[3] Ding L., Kolari P., Ding Z., and S. Avancha, “Using Ontologies in the Semantic Web: A Survey”, Ontologies, integrated series of information systems, vol 14, pp. 79-113, Springer, 2007.

[4] E. Greengrass, “Information Retrieval: A survey”. DOD Technical

Report TR-R52-008-001, November 2000.

[5] Grossman D., and O. Frieder. “Information retrieval algorithms and heuristics”. Second ed. . Springer. 2004.

[6] Hajian B., and Tony W., ”A method of measuring semantic similarity using a multi-tree model proceedings IJCAI 2011 - 9th Workshop on intelligent techniques for web personalization & recommender systems (ITWP'11) Barcelona, Spain, 16 JULY 2011.

[7] Lamberti F., Sanna A., and C. Demartini, “A relation-based Page Rank algorithm for semantic web search engines”, IEEE Trans Knowledge and Data Eng., vol. 21, no. 1, Jan 2009.

[8] Oleshchuk V., and Asle P., “Ontology Based Semantic Similarity Comparison of Documents”, Proc. of IEEE 14th workshop on database and expert systems applications,2003.

[9] Othman R., Deeris S., Illias R., Alashwal R.and Rohayanti H.,”Incorporating semantic similarity measure in genetic algorithm: an approach for searching the gene ontology terms”, International Journal of computational intelligence, vol 3, no. 3, 2006.

[10] Page L., S. Brin, R. Motwani, and T. Winograd, “The Page Rank Citation Ranking: Bringing Order to the Web”, Stanford Digital Library Technologies Project, 1998.

[11] Protiti M., “Semantic web: The future of WWW”, Proc. Of 5th Int’l Conf. CALIBER, Punjab University, Chandigarh, 08-10, 2007.

[12] Shamsfard M., Namehtzadeh A., and S. Motiee “ORank: An Ontology based System for Ranking Documents”, Int’l Journal of Computer Science, vol 1, no 3, ISSN 1306-4428, 2006.

[13] Takale S., and Sushma N., ”Measuring Semantic Similarity between Words Using Web Documents”, (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 1, No.4 October, 2010.

[14] Thiagarajan R., Manjunath G., and Markus S.,”Computing semantic similarity using ontologies “ ISWC 08, the International Semantic Web Conference (ISWC), Karlsruhe, Germany, 2008.

[15] Wei W., Barnaghi P. and Andrezj B.,” Search with meanings: An overview of semantic search systems”, School of Computer Science, University of Nottingham Malaysia.

[16] Yuxin M,” A semantic-based genetic algorithm for sub-ontology evolution”, Information Technology Journal, 9(4), ISSN1812-5638, pp 609-620, 2010.