ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Retrieving Best Web Pages Using Genetic Algorithm

J.Sugumar[*1],M.Kirankumar[2], S.Prabhakaran[3]
Department Of Computer Science and Engineering, Saveetha Nagar, Thandalam, Chennai-602105.,
Saveetha School of Engineering, Saveetha University, India
sugu.rose2010@gmail.com

Abstract

This paper is concerned with extraction of top web pages from the search engine. We use an effective algorithm for extracting the target file from the search engine. The majority of search engines like Google provide search consequence using the key terms. In this we extract the top web pages by forming keywords as set of queries with a high performance. We describe our approach for ontology extraction for an existing knowledge base of web pages and we use the word net tool for identifying the Synonymous of the given queries. As an example, the simple keyword "CAR" also have the other synonyms is "AUTOMOBILES".

Keywords: Web Extraction, top web pages list, Genetic Algorithm, Ontology Extraction.

Introduction

The Internet is the greatest way to obtain data. Even so, the majority of data on the net can be unstructured textual content with natural 'languages', and also natural words textual content can be quite difficult. So search engines goals with utilizing the original web sites to improve the particular link to retrieve inbound links which might be highly related to problem. Integrating semantic systems directly into search engines can produce highly pertinent to people. The particular semantic strategy entails removal of the most extremely essential text or even search phrases from the consumer problem also to uncover it is synonymous phrases and also look for through meaning to individuals search phrases. Including the sentence "This car or truck can be orange with shade "is exactly like the sentence "This car or truck can be orange with color". Both the paragraphs imply the same but use of phrases differs. The definition of "car" can be synonymous to the expression "automobile". They can be used interchangeably. Ontologies tend to be measured among the support beams of the semantic World Wide Web. Ontologies in many cases are equated having taxonomic hierarchies of courses, type explanations, and also the sub relative, but ontologies does not need to be restricted to these types of sorts. Ontologies are actually generally used so as to response troubles produced from the particular administration of discussed sent out expertise and also the integration of data over various programs. Even so, the process of ontology constructing continues to be a new substantial and also error-prone assignment. Therefore, several scientific studies to semi-automatically or even on auto-pilot develop ontologies via present docs are actually formulated. The particular Semantic Internet depends greatly about the proper ontologies of which structure root data for the purpose of extensive and also lightweight unit realizing. Through getting rid of pertinent ontology ideas and also their particular relationships at a expertise starting of heterogeneous textual content docs that happen to be attained while in look for tend to be highly useful with locating inbound links that may match up the particular user's fascination exactly.

System analysis

Existing System

In existing system the links the extracted from the other search engines by means of ranking algorithms. Different search engines utilize different ranking algorithms to get the final result page. Most of them are based on the number of time the link has been visited. These search engines also have a drawback of semantic analysis in processing the user query. Moreover the links retrieved by them may or may not be active currently.

Problem Statement

- Most of the search engines do not apply semantic analysis.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

- The results are based on the number of times the links have been visited.
- The search results generated are unlimited and they are time consuming to browse all links to find a best link.

Proposed System

The proposed method utilizes a set of queries instead of a single query. It is done with the help of Word Net ontology. The ontology is used for extracting the similar words for the user query. When the query set is created, then it is subjected to search in Google search engine and the top results are selected.

Advantages

- The search engine proposed results in retrieval of relevant results for user queries.
- It does semantic analysis by using Word Net tool and ontologies are extracted.
- The results are limited and highly appropriate.

System design

Architecture Diagram

It gives the basic architecture of the developing project.

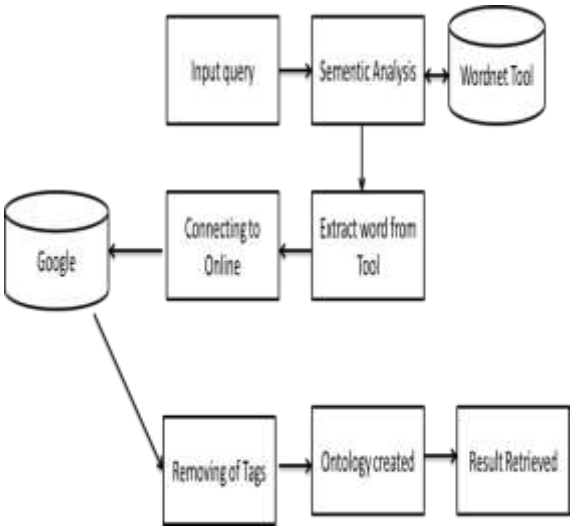

Figure 3.1

System implementation

Module Explanation:

1) Input search query

In this module the user will be entering the search query according to their interest.

2) Synonymous identification

In this module, we identify the exact synonymous of the user entered search query. To resolve synonymous problem we have utilized word net tool for finding the exact meaning for the particular word. The Word Net is usually a large lexical dictionary which contains Nouns, Verbs, and adjectives. Many of the word contain different meanings for the same word for example the word Apple can be differenced as a fruit or apple product.

3) Terminology extraction

The many previously mentioned sub phases are conducted to help extract the particular pertinent terminology related to a selected sector. We all consider terminology because the set of text or term strings which share a single quite possibly intricate propagated this means in just a community. Because of their low ambiguity and also high specificity, most of these text are also especially employed to conceptualize a knowledge base.

4) Genetic Algorithm Module

This module is used to find the relationship between the extracted terms from the links and the user query using the genetic algorithm. In a genetic algorithm, a population of candidate solutions to an optimization problem is evolved toward better solutions. Each candidate solution has a set of properties which can be mutated and altered. The evolution usually starts from a population of randomly generated individuals, and is an iterative process. The population in the each iteration is called as generation. In each generation, the fitness of every individual in the population is evaluated; the fitness is usually the value of the objective function in the optimization problem being solved. The more fit individuals are stochastically selected from the current population to form the new generation. The new generation of candidate solutions is then used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population.

4) Removal of Tags

In this parse, the removal of unwanted tags will removed from the retrieved links before it has been shown to the results. The removing of unwanted tags done because there will be many tags which does not any meanings to the result and it does not gives proper information to the user.

5) Search Results

In this parse result will be extracted after the removal of unwanted tags using the html parser once

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

the ontology has been created for the particular topic which the user has been entered in the search engine according to the Synonymous identification the appropriate links will be retrieved from the search.

Experimental result

It gives of the experimental result developing project.

Homepage:

It is a Home page for retrieving of web page.

Figure 5.1

Ontology Search:

This will be retrieving the web pages by find the word fitness.

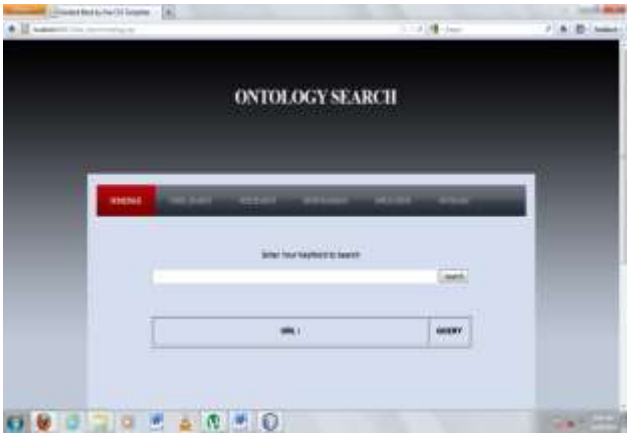

Figure 5.2

Input The Search Query:

Entering of interested search queries.

Figure 5.3

Retrieved Links Result:

Results of retrieved links

Figure 5.4

Console Output

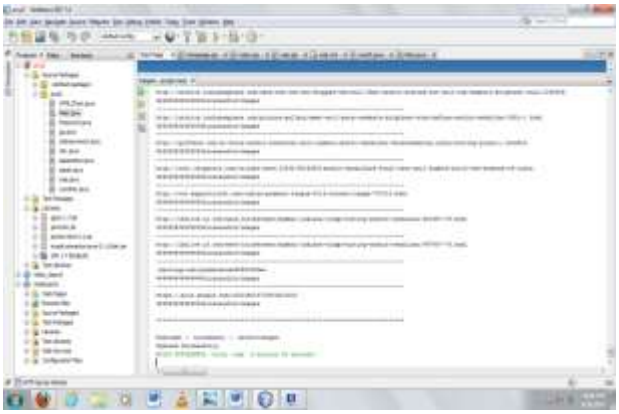

Figure 5.5

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

Database Storage:

The retrieved results are stored in the database

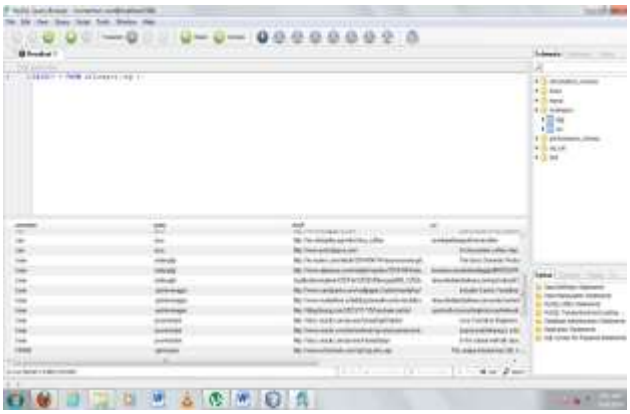

Figure 5.6

Conclusion

We have proposed semantic measure for obtaining the relevant query to search through different search engines. The overall steps of the semantic meta search engine includes process like, (i) relevant query formation using semantic similarity measure, ii) extraction of web documents based on the relevant query and iii) Retrieving of relevant search result. Here, input query and neighbors extracted from ontology is used to select the most suitable query and then, retrieving of web pages obtained from the search engine was done. The experimentation was performed with different set of queries and the performance of the results was analyzed with the help of precision. From the experimental results, we found that the proposed search engine has performed better than existing system.

References

1. T. Weninger, F. Fumarola, R. Barber, J. Han, and D. Malerba, "Unexpected results in automatic list extraction on the web," *SIGKDD Explorations*, vol. 12, no. 2, pp. 26–30, 2010.
2. "20 most influential scientists alive today," http://www.superscholar.org/ features/20-most-influential-scientists-alive-today/.
3. X. Yin, W. Tan, and C. Liu, "Facto: a fact lookup engine based on web tables," in *WWW*, 2011, pp. 507–516.
4. "Googlesets," http://labs.google.com/sets.
5. M. J. Cafarella, E. Wu, A. Halevy, Y. Zhang, and D. Z. Wang, "Webtables: Exploring the power of tables on the web," in *VLDB*, 2008.
6. B. Liu, R. L. Grossman, and Y. Zhai, "Mining data records in web pages," in *KDD*, 2003, pp. 601–606.
7. G. Miao, J. Tatemura, W.-P. Hsiung, A. Sawires, and L. E. Moser, "Extracting data records from the web using tag path clustering," in *WWW*, 2009, pp. 981–990.
8. W.Gatterbauer, P. Bohunsky, M. Herzog, B. Kr¨upl, and B. Pollak, "Towards domain-independent information extraction from web tables," in *WWW*. ACM Press, 2007, pp. 71–80.
9. F. Fumarola, T. Weninger, R. Barber, D. Malerba, and J. Han, "Extracting general lists from web documents
10. Y. Yamada, N. Craswell, T. Nakatoh, and S.Hirokawa, "Testbed for ainformation extraction from deep web," in *WWW*, 2005.