User-Interest ontology construction for Selecting seed URL of Focused crawler with the consideration of Semantic relations

[1]R. Nithiya, [2]S.R. Lavanya

[1]Student, [2]Assistant Professor, Master of Philosophy in computer science, Sri Ramakrishna College of Arts and Science for Women, Coimbatore, Tamilnadu-641044

[1] nithyamphil1@gmail.com

Abstract

Objectives: The main objectives of this methodology are to improve the performance of the web crawlers by retrieving the web pages based on the user queries. And also this work aims to construct the user interest ontology in terms of the user interest features by extracting the syntactic and semantic relationship exists among them.

Methods: The semantic meaning extraction for the user interest terms are proposed in this work. This is done with the help word net toolkit which aims to retrieve the concept and meaning of each and every word. Then the user interest ontology is constructed by deriving the concept relation present among the user submitted terms and the documents by calculating the relatedness and the similarity present among them with the help of tf- IDF score and the semantic relation. Finally formal concept analysis is used to represent the terms of the user interest features in the hierarchical form.

Results: The experimental tests were conducted after implementation of the proposed methodology in order to prove the effectiveness of the algorithm. This is done by comparing it with the existing work in terms of precision and recall performance measures. The experimental tests conducted were proves that the proposed methodology is improved in terms of both precision and recall measure.

Conclusion: The findings of this work demonstrate that the proposed methodology provides the user interest ontology construction and retrieve the user interested seed URL in the effective manner.

Keywords: Seed URL, User interest, ontology, semantic relation, syntactic relation

1. Introduction

Search retrieval is the most important research scenario in today's world where the numbers of network users are increasing in number. The search engines increase in numbers which aims to retrieve the query results to the users. The query results are retrieved based on the user searched topics and relevance to the files which are stored in the databases. One of the primary research issues occurs in the search engine retrieves is the accuracy of the contents that are recovered over the user questions. The search queries that are retrieved need to fill the user interest by retrieving the query results based on the user interest topics. The user interest topics are retrieved based on the user search results which point to retrieve the topics from the databases with the thoughtfulness of the user interest. The user interest calculation is done by assembling the user requests from them through the GUI interfaces.

The main limitation that may spring up in the user search, retrieval process is the privacy violation which may contribute to a reduction of the user reputation level. In order to overcome this problem, the relationship among the queries and the databases need to be retrieved in order to provide a more accurate and suitable results. However, retrieving the results with consideration of the user satisfaction level will be a more burden process in which the retrieval process will be more unmanageable. This problem needs to be broken up and the user satisfaction need to be improved well to increase the search engine reputation level.

Ontology is the ace of the methodology which aims to map the attribute in the form the relationship among the properties. The ontological representation of the user attributes and the database results can contribute to an efficient retrieval of the lookup results. Presenting the user interest in the shape of an ontology may lead to an accurate and flexible retrieval of the resolution regarding the user interests.

The primary contribution of this study is to implement an effective mechanism which aims to recover the user interest topics to the user whenever they submit the query. To reach this concept lattice form is put in which aims to extract the interrelationship between the query terms and the database contents. This process is overcome by presenting the methodology called the semantic relation extraction among the query terms and the databases. The semantic query term extraction is done with the help of the word Net toolkit which is used to retrieve the synonyms and meaning of the terms present in the query topic. This will lead to an effective retrieval of the query result with more accuracy in terms of user interest and will provide most wanted results.

The governing body of this work is made as follows: Section 1 describes and provides a detailed introduction about the search retrieval process and the user interest satisfaction. Section 2 discusses about the various researches which has been conducted previously for retrieving the user interest based results. Part 3 provides a detailed description of the proposed methodology in this study in section 4 experimental tests which has been conducted to prove the effectiveness of the proposed methodology than the previous method is given. Finally in section 5 the research of this work is concluded by using the performance metrics called the accuracy and precision.

2. Related works

In [1], web crawling methodology is discussed which intends to retrieve the suitable web pages from the large set of web pages. This is applied to regain the most similar document in terms of question submission.ion. The efficient methodology si introduced in this method to retrieve the most similar documents in terms of the query which is submitted. It is executed with the thoughtfulness of the user query retrieval which aims to recover the answers based on the survival rate. It is performed by sorting the documents with the clear separation idea exist among the several documents. So the search, retrieval can be served efficiently with the cognition of the structure representation among the various modules.

In [2], web ontology construction methodology is discussed in the detailed style which points to retrieve the contents in terms of the petition. Web ontology construction is used to represent the network terms in the form of the ontology in which the relation terms can be extracted by using the effective search retrieval process. The best ontology can lead to a successful result generation in terms of the most iterative environment. The web ontology needs to be updated in order to make sure the retrieval of content id given in the updated manner. The novel approach is discussed in this work for efficient updating of the ontology.

In [3], has discussed a fresh path to retrieve the web pages in accordance to the knowledge which are retrieved later.. The knowledge learning procedure depends the previous transaction and credits which has been done before. To do so topic specific web crawler has been inserted in this study which intends to recover the contents based on the user interest scores which has been measured earlier. Crawling has been performed to recover the contents based on the user interest topics in terms of the user interest scores.res. This is done by calculating the page retrieval score for each and every topics which are retrieved later in order to evaluate and retrieve the most specific contents.

In [4], discussed a novel approach for getting the ontology with the thoughtfulness of the user context ontological profile is constructed in this work by using the knowledge of the user interest which is gathered from the users. The knowledge's that are gathered from the users are gathered and have to be gained before capturing the user knowledge. The constructed user interest ontology is specified in terms of the user concept relation extraction model where the concepts of users will be extracted from them and then the evaluation of the concepts will be worked into the ontology. This ontological representation may used to define the user interest representation in terms of concepts of domain attributes and the association rules.

In [5], discussed a way of constructing ontology with the help of horn clauses which means to retrieve the contents and pulls up the relationship present among them by using the horn clause representation ontological representation is used to extract and derive the relationship among the concepts and terms that are extracted from the various concepts which has been derived. The formal concept analysis methodology is used to derive the relationship present among terms which are extracted and used to construct the most valuable ontology. So the users can extract the substance from the constructed ontology in the flexible manner.

In [6], integration of information from the more sites is discussed in the detailed style which points to derive the meaningful data. done to evaluate the concept of the terms in order to derive the more meaningful

information. It is served by merging more than two ontologies in terms of the relationship present among the diverse concepts and field attributes. To do so, the, merging of the ontologies is done. This merging of ontologies is used to retrieve the most meaningful information in order to make the user satisfaction level by regaining the most meaningful information from the data. The merged ontologies can be utilized to regain the most meaningful data which can be utilized to retrieve the meaningful term extraction.

In [7], the introduction of a novel approach for the merging ontologies is introduced which aims to recover the more meaningful info. It is done by comparing the feature sets of domain attributes which are collected from the most regions. It is performed to assess the concepts of the terms that are assembled from the various websites which purport to predict the answers of the termini. The knowledge extracted from the multiple sites can be combined to extract the various information. The concept relation extraction is done with the thoughtfulness of the triplets by using which the concepts of the terms can be derived.. The triplets consist of the attributes of the most valuable metrics like identity and concepts of the terms which are extracted from the various entities. It is performed to assess the significance of the various terms and extract the knowledge.

3. Ontological based seed URL Extraction

The web crawler is the concept of indexing the web pages which are recovered over the search inquiry. The retrieved web pages will be indexed with the numeric values by using the ontological representation. This agency of the ontological profile construction is applied to deduce the contents based on the user interest. Ontological based user profile construction is implemented which aims to store the user interest by with the help of user log profile and retrieve the contents in the future scenario based on the user request terms. The ontological profile construction leads to an efficient retrieval of the user query contents in terms of user satisfaction level. It is done by achieving the ontological profile construction and also contributes to the efficient retrieval of the user query. At the time of user interest ontology construction this worked only considers the syntactic meaning of the field attributes. In improver to that this work aims to retrieve the contents and construct the user ontology with the thoughtfulness of the semantic meaning also.. The semantic meaning is used to retrieve the synonym of the domain attributes which are submitted by the user users. With the help of it one can learn and retrieve the accurate and more improved retrieval of contents. The seed URL extraction is performed as follows:

1. Construct the user interest ontology

2. Update the Ontology

3. Computer the semantic relations present among topics

4. Finding the user interest topics

The above steps are adopted in the proposed work to recover the user interest ontology which aims to recover the user interest based topics. These steps are discussed detailed in the following subjects.

A. Construct the user Interest Ontology

The user interest ontology is made to offer the relationship present among the domain attributes which are stated by the users and equally well provide the meaningful retrieval of contents. The user interest ontology is constructed with the count of the semantic and syntactic meaning presents, among the conditions which are excerpted. This semantic meaning is applied to offer the clear representation of the relationship exists between the conditions which are stated by the users. In this work word net is used to recover the meaning of the contents which are stated by the users. With the assistance of the word Net toolkit one can retrieve the synonyms of each and every word that are stated by the users. The word net tool retrieved the concept of every word that is stated by the users by using the knowledge bases where the meanings of the values are stored.

In improver to that fuzzy formal concept analysis methodology is used which aims to recover the contents based on the user interest features. This Fuzzy formal concept analysis is applied to retrieve the contents in terms of the basic feature vector that are recovered over the user questions. By using this FCA one can retrieve the real world concepts by applying the concept lattice. The concept lattice is nothing only the representation of the seed URLs in the contour of the formal context. So the user interest of hunting can be accurately foretold.

B. Update the Ontology

After creation of ontologies, the created ontology will be updated with the consideration of the concept maps which are extracted from the various modules. The ontology merging concept is practiced with the thoughtfulness of the knowledge infrastructures and the user interest feature vectors. This user interest feature vector is applied to regain the most essential concept relation present among the different characteristics. Finally the attribute selection is managed with the thoughtfulness of the relationship present among the different ontologies in terms of the field attributes and the relationship present among them. It is achieved by putting back the domain attributed with some of the vector attributes as like to follow: to follow: If the present attribute consists of some identical entities with the attributes x which was created, then allow one attribute to share other attribute which was shared.

(a) If the present attribute leads to an management of the all of the other attributes then the present node will be represented as the father node and the current object will be predicted as the child node.

(b) If present attribute don't satisfy the above conditions, then the merge the present attribute with the other attributes like a shot.

Later Updation of the user interest ontology, the concept lattice will be derived which aims to evoke the meaning present among the corresponding concept and the terms are fixed. The guests present in the ontology tree will be mapped as the concept hierarchy as follows:

(1) If the LUB concept of the optimized concept lattice is Ø, we initialize the source node of the ontology tree into Ø; otherwise we map the LUB concept into the user-interest ontology tree as the source node;

(2) If the node C in the optimized concept lattice has many synonymous concepts C1, . . ., Cn. C1 is treated as the super-concept and the other concepts are inserted in a left-to-right order as sub-concept of C1;

(3) If the node C has sub-concepts, we enter them into the branch of the node C;

(4) If C1, . . ., Cj have a common sub-concept (not an empty concept), in order to avoid repetition, we simply hold on one concept in one leg of one client;

(5) Repeat (2)–(4) until C1, . . ., Cj have a common empty concept.

C. Computer the semantic relations present among topics

In this research, the ontology is constructed with the consideration of both syntactic and semantic representation of nodes. This is by using the methodology called the term based VSM and tf-IDF score values. And also semantic meanings are taken out with the assistance of the word net toolkit. The semantic relation between the terms that are submitted by the users and the context information's by using the following formulae:

$$\mathrm{Rel}\left(t, c_i \middle| d_j\right) = \frac{1}{|T|-1} \sum \frac{1}{|CS\,1|} \sum \mathrm{SIM}(c_i, c_k) \qquad (1)$$

Where

T - term set of the j^{th} document d_j ,

t_l - term in d_j except for t and

cs_l - candidate concept set related to term t_l.

SIM (ci, ck) is the semantic relatedness between two concepts, which is computed as follows:

$$\mathrm{SIM}(c_i, c_k) = 1 - \frac{\log\ (\max\ (|A|,|B|)) - \log\ (|A \cap B|)}{\log\ (|W|) - \log\ (\min\ (|A|,|B|))} \qquad (2)$$

where

A and B - sets of all articles that link to concepts ci and ck respectively, and

W - set of all articles in Wikipedia

The page relatedness values will be separated in the ascending order and then the higher value of real (t, ci|DJ) will consider as conceptualising chi is more semantically related to term it. to term it. Based on Rel(t, ci|dj) and term's weight w(tk, dj), the concept's weight is defined as their weighted sum as follows

$$W(c_i, d_j) = \sum w(t_k, d_j) * Rel(t_k, c_i|d_j)$$

Grounded on this weight score, user ontology will be updated and as well as the context will be retrieved more accurately in the efficient manner as per user demand.

4. Finding the user interest topics

Then in the end, with the aid of the user interest ontology, results will be retrieved based on the user query.y. It is done with the consideration of the semantic relationship and the syntactic meaning present among the query terms. The proposed methodology which means to recover the user interest topics are presented as follows:

ALGORITHM

(1) Combine the user-interest ontology, improve the base set of HITS, call an algorithm discovery user topic area. In the algorithm, the hub and authority web pages are adopted to describe the user topic area.

(2) Here use HITS algorithm to generate the hub and authority web page set for a user query, the hub web page set is considered as F of BDG, the authority web page set is believed as a C of BDG.

(3) Extract CBDG from BDG, Form Hset and Aset of CBDG.

(4) Let URLs in Hset be the seed URLs of the focused Web crawler.

(5) Make use of Hset and Aset on web graph G to discovery other URLs for user topic area.

(6) Delete all URLs of Hset and Aset from the hub web page set and the authority web page set, update the two bands.

(7) Repeat step (3), until the number of the seed URLs of the focused Web crawler is enough.

So finally, efficient seed url extraction mechanism has been implemented which aims to retrieve the contents in terms of the user interest anthology features. Whenever the user submits the query the corresponding documents will be retrieved based on the user interested topics.

EXPERIMENTAL RESULTS

After execution of the proposed methodology, the operation of the suggested approach is derived by comparing it with the existing approach in terms of the performance metrics called the precision and recall measure. These standards are utilized to foretell how much correctly the user documents are recovered with the gratification of the user interest features. The comparing is done as follows and explained in the detailed manner.

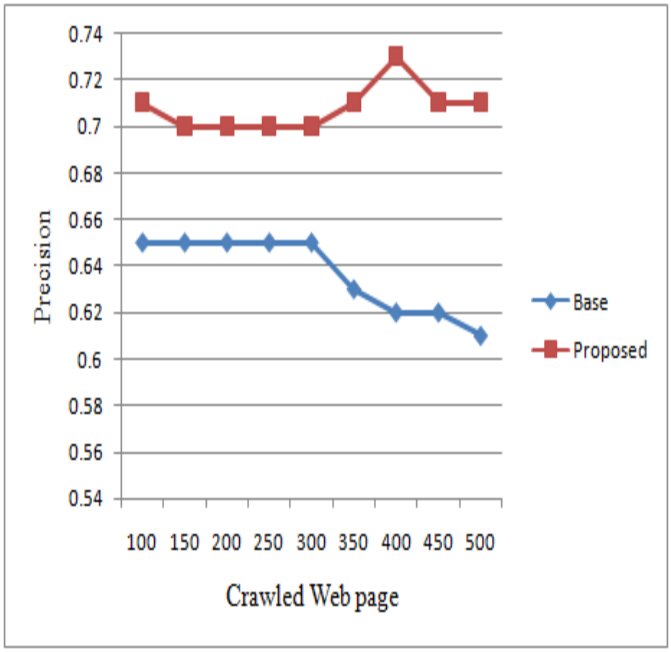
A. Precision

Precision value is used to indicate that whether the retrieved result is relevant to the corresponding document or not. The precision is computed as follows:

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}$$

The comparison graph is depicted as sticks with in figure 1. In that graph, the precision value is compared against the various numbers of crawling web pages. In the x axis numbers of crawled web pages are consumed and in y axis precision value is used up.

Figure 1. Precision Comparison



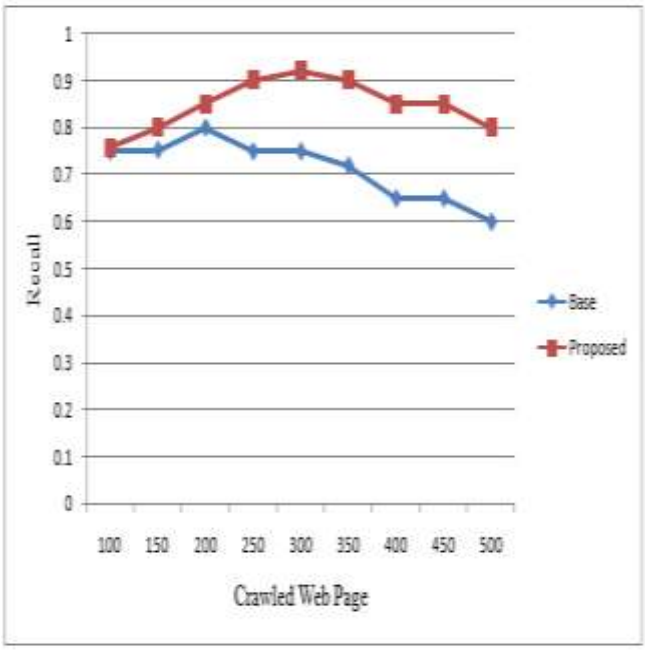
B. Recall

Recollection is the performance standard which is applied to assess whether the retrieved result is relevant to the document or not.t or not. That is recall is used to define whether the retrieved result is successfully retrieved or not. The computation of the recall value is executed as follows:

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

The comparison graph is drawn as depicted in figure 2. In that graph, the recall value is compared against the various numbers of crawled web pages. In the x axis numbers of crawling web pages are required and in the y axis resale value is needed.

Figure 2. Recall Comparison



4. Conclusion

Personalized seed URL selection is the process of retrieving the most similar document and the contents related to the user query. In the existing system entirely the terms are considered for relation extraction. This will degrade the overall accuracy rate and performance of the organization. The time complexity rate of the system is more eminent; it will precede the lower performance in this system. In decree to sweep over this issue and improve the functioning of the organization, our proposed system the feature vectors are presented with both text and concept. The syntactic level represents each web page as a term vector. The semantic level represents a web page with concepts related to terms in syntactic level.. Search complexity and memory complexity rate of the system is abbreviated in this arrangement compared to the living scheme. The suggested scheme is well effective than the existing system by using this syntactic information as well as the semantic relation of users. Experimentation result of the system is shown that our suggested scheme is well reduce the time complexity rate and memory complexity as well as enhance the carrying into action of the organization.

5. Reference

1. Ahmed Patel, Nikita Schmidt. Application of structured document parsing to focused web crawling. *Computer Standards & Interface*. 2011; 33, 325–331.
2. Asad Masood Khattak, Khalid Latif,Sungyoung Lee. Change management in evolving web ontologies. *Knowledge-Based Systems.* 2013; 37, 1–18.
3. Niran Angkawattanawit, Arnon Rungsawang. Learnable crawling: An efficient approach to topic-specific web resource discovery. *First International Conference on Computational Intelligence, Communication Systems and Networks, CICSYN* 2009, Indore, India, 23-25 July, 2009; 01/2009.
4. Geir Solskinnsbakk, Jon Atle Gulla. Combining ontological profiles with context in information retrieval, *Data & Knowledge Engineering*. 2010; 69, 251–260.
5. Hele-Mai Haav. A Semi-automatic method to ontology design by using FCA. *Concept Lattices and their Applications – CLA*. 2004.
6. Rung-Ching Chen, Cho-Tscan Bau, Chun-Ju Yeh. Merging domain ontologies based on the WordNet system and Fuzzy formal concept analysis techniques. *Applied Soft Computing.* 2011; 11, 1908–1923.
7. Gerd Stumme, Alexander Maedche. FCA-MERGE: Bottom-Up merging of ontologies. *Proceedings of the 17th International Joint Conference on Artificial Intelligence.* 2001; 1, 225-230.

The Publication fee is defrayed by Indian Society for Education and Environment (iSee). www.iseeadyar.org

Citation:

R. Nithiya, S.R. Lavanya. User-Interest Ontology Construction for Selecting Seed URL of Focused Crawler with the Consideration of Semantic Relations. *Indian Journal of Innovations and Developments.* 2014; 3 (3), 61-67