ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 3, March 2017

Exploratory Search for Retrieving Unaware Fields for Users Using Ontology Clustering

R.Manjula[1], M.Chellammai [2] , S.Preethi[3], M.Savitha[4]

Associate Professor, Dept. of IT, Panimalar Engineering College, Chennai, Tamil Nadu, India[1]

B. Tech. Student, Dept. of IT, Panimalar Engineering College, Chennai, Tamil Nadu, India[2,3,4]

ABSTRACT: As information retrieval using web grows at a very fast pace, there has been great interest in techniques that help efficiently locate deep-web interfaces. An exploratory search may be driven by a user's curiosity or desire for specific information. When users investigate unaware fields, they may want to learn more about a particular subject area to gain their knowledge rather than solve a specific problem. A matching query style has significant limitations. Search results are satisfactory only when users give the right search query. "Smart Crawler" ranks websites to prioritize highly relevant ones for a given topic. To achieve more accurate results for an exploratory search, we find the synonym, hyponym and hypernym of the root word after pre-processing the given query by user. We have added a feature "Bookmark" to save the websites both locally and globally so that it's browser independent.

KEYWORDS- Successor stemming algorithm, Customized stopping keywords, Vector space model ,Clustering algorithm, Ontology clustering, Term Frequency and Inverse Document Frequency

I.INTRODUCTION

To retrieve information about the query which is given by the user in depth manner by finding synonym, hypernym and hyponym of the query given by the user that is pre-processed. We use customized stop list to pre-process the user's query for removal of stop words. Further, successor count root words are found by "successor stemming algorithm" for eliminating the suffix of the given query. There are four techniques used in this stemming process to stem the words. They are cut-off, peak and plateau, complete word method and entropy. From the obtained root word, we find hypernym, hyponym and synonym using "Natural Language Processing". Then for those root words, C and inverse document frequency are calculated and based on that calculations, we have stored website links. These links are retrieved when the user selects the word to know in detail. Images are also stored for perceptual view. Additionally a feature "bookmark" that saves the websites both locally and globally which is diagrammatically represented in Fig: Architecture Diagram of Ontology clustering. So that they are browser independent and can also be used in all types of browsers without importing it.

II.RELATED WORKS

[1] It introduces an indexing network model and five related normal forms to advance the field. As a basic model, indexing network organizes and manages various information service resources through analyzing the relationships among them.

[2] It propose a phrase-based document similarity to compute the pair wise similarities of documents based on the Suffix Tree Document (STD) model.

[3] It is using clustering-by-directions algorithm. The algorithm introduces a novel approach to interactive query expansion. It is designed to support users of search engines in forming Web search queries.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 3, March 2017

[4] It proposes scalable method for measuring the relevancies between categories through topic models, which takes consideration of both content and user interaction based category similarities is developed here.

[5] This paper proposed a new mechanism called Tf-Idf based Apriori for clustering the web documents. It then rank the documents in each cluster using Tf-Idf and similarity factor of documents based on the user query.

[6] A novel probabilistic retrieval model is presented. It forms a basis to interpret the TF-IDF term weights as making relevance decisions.

[7] This paper presents a summary of such graph-based methods that are found in recent technical publications plus an analysis of their component functions and their maturity calculated from information found in the referenced papers.

[8] This paper proposes a new keyword search approach which basically utilizes the statistics of underlying XML data to decide the promising result types and then quickly retrieves the corresponding results with the help of selected promising result types.

[9] This paper discusses a framework of adding background knowledge from Linked Open Data to a given data mining problem in a fully automatic, unsupervised manner. It introduces the FeGeLOD framework and its latest implementation, the RapidMinerLinked Open Data extension.

[10] This paper reports on the formal language developed for encoding information and present approaches to solve the inference problems related to finding information, to determining if information is usable by a robot, and to grounding it on the robot platform.

[11] This paper presents a method for feature selection and region selection in the visual BoW model. This allows for an intermediate visualization of the features and regions that are important for visual learning.

[12] The aim of the paper is to determine the sentiment polarity expressed in a text. In online customer reviews, the sentiment polarities of words are usually dependent on the corresponding aspects. a novel cross-lingual topic model framework which can be easily combined with the state-of-theart aspect/sentiment models.

III.SCOPE OF RESEARCH

The aim of the project is to develop a search engine which can be used for exploratory search for retrieving unaware fields for different kind of users using ontology clustering and to provide a browser independent book marking system. The search engine existing now will provide good results only if the keywords are given correctly in search query. So we are going to develop a search engine which allows user to search even if they doesn't know exact keywords using ontology clustering and multi term search and to provide a bookmarking system independent of browsers. The purpose of project is to provide an exploratory search for users to know about different fields using Hypernym, Hyponym, Synonym of the keywords given in search query and to allow them to bookmark globally and locally by storing bookmarks in database.

IV.PROPOSED METHODOLOGY AND DISCUSSION

We proposed an efficient and flexible search scheme that supports both multi-keyword ranked search and synonym hyponym and hypernym based search. To implement multi-keyword search and result ranking, Vector Space Model (VSM) is used to build document index, where each document is expressed as a vector and each dimension value is the "Term Frequency (TF) weight" of its corresponding keyword. A new vector is also generated in the query phase. The

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 3, March 2017

document has the same dimension with document index and each dimension value is the "Inverse Document Frequency (IDF) weight."

To compute similarity of one document in a large dimension to search a query we use "Cosine measure" that calculates the angle between them. Tree base index structure is used to improve search efficiently and balanced by Binary tree. The searchable index tree is constructed with the document index vectors, So the related documents can be found by traversing the tree.

V.ARCHITECTURE DIAGRAM

To retrieve information about the query which is given by the user in depth manner by finding synonym, hypernym and hyponym of the query given by the user that is pre-processed. We use customized stop list to pre-process the user's query for removal of stop words. Further, successor count root words are found by "successor stemming algorithm" for eliminating the suffix of the given query. From the obtained root word, we find hypernym, hyponym and synonym using "Natural Language Processing". Then for those root words, term frequency and inverse document frequency are calculated and based on that calculations, we have stored website links. These links are retrieved when the user selects the word to know in detail. Images are also stored for perceptual view. Additionally a feature "bookmark" that saves the websites both locally and globally.

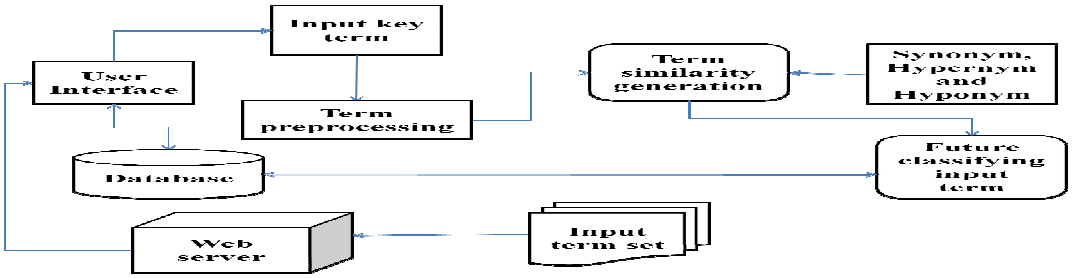


Fig: Architecture Diagram of Ontology clustering

ALOGORITHM

K-MEANS CLUSTERING ALGORITHM

For finding similarity between keywords we are using k-means clustering algorithm as it is simple and more efficient for searching purpose.

Let $X = \{x_1,x_2,x_3,\ldots\ldots,x_n\}$ be the set of data points and $V = \{v_1,v_2,\ldots\ldots,v_c\}$ be the set of cluster centers.

1) Randomly select '*c*' cluster centers.

2) Compute the distance between each seed point and centers of clusters which we selected randomly

3) Assign the seed point to the cluster center whose distance from the cluster center is less than that of all the cluster centers..

4) Recalculate the new cluster center like finding mean for seed points

5) Recalculate the distance between each seed point and new obtained centers

6) If no seed point was reassigned then stop, otherwise repeat from step 3.

This algorithm aims at minimizing squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (||x_i - v_j||)^2$$

where,

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.

'c_i' is the number of seed points in i^{th} cluster.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 3, March 2017

'*c*' is the number of cluster centers (mean)

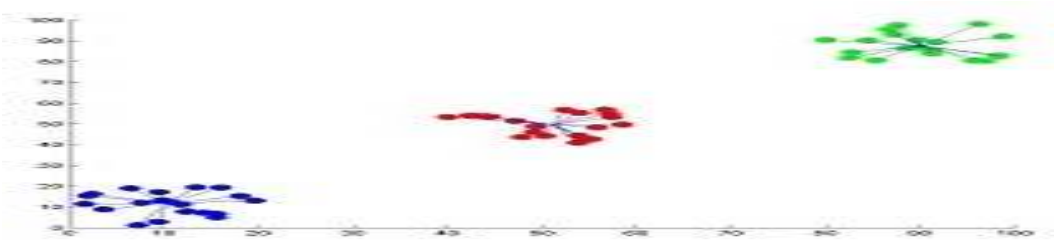


Fig: an example for forming clusters using similarity

Advantage

Relatively efficient: O(tknd), where n is # objects, k is # clusters, d is # dimension of each object, and t is # iterations. Normally, k, t, d << n. Fast and robust too.

VECTOR SPACE MODEL

For ranking websites we are using vector space model as it is efficient and involves cosine similarity function to find similarity between keywords and links.

1.Find term frequency for keywords using Robinson tf (okapitf):

$$\text{okapi\_tf} = \frac{tf}{tf + .5 + 1.5\frac{doclen}{avgdoclen}}$$

2. Find inverse document frequency for website links using below formula

IDF(t)=log(N/Nt) where, N= #of docs

Nt= #of docs containing term t

3. Compute cosine similarity for tf and idf like below,

$D_1 = (0.5T_1 + 0.8T_2 + 0.3T_3)$ $\quad Q = (1.5T_1 + 1T_2 + 0T_3)$

$$Sim(D_1, Q) = \frac{(0.5 \times 1.5) + (0.8 \times 1)}{\sqrt{(0.5^2 + 0.8^2 + 0.3^2)(1.5^2 + 1^2)}}$$

$$= \frac{1.55}{\sqrt{.98 \times 3.25}}$$

$$= .868$$

4. Compute tf-idf base similarity using below formula,

$$\frac{\sum_t (TF_{query}(t) \cdot IDF_{query}(t)) \cdot (TF_{doc}(t) \cdot IDF_{doc}(t))}{||doc|| \cdot ||query||}$$

SUCCESS VARIETY STEMMING ALGORITHM

For stemming less weight-age words we are using successor variety stemming algorithm as it is efficient and find correct root words using different strategies stated below:

1. Using the **cutoff method**, some threshold value is selected for successor varieties and a limit is identified whenever the threshold value is reached. The problem with this technique is how to select the threshold value. If it is too small, incorrect splits will be made; if too large, correct splits will be missed.
2. With the **peak and plateau method**, a segment split is made after a letter whose successor variety exceeds that of the letter immediately preceding it and the letter immediately following it. This method removes the need for the threshold value to be selected.
3. In the **complete word method**, a split is made after a segment if the segment is a complete word in the corpus.
4. The **entropy** method takes advantage of the distribution of successor variety characters. The method works as follows. Let |Dαi| be the number of words in a text body beginning with the i length sequence of letters. Let

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 3, March 2017

|Dαij| be the number of words in Dαi with the successor j. The probability that a member of Dαi has the successor j is given by $H_{\alpha i} = \sum_{p=1}^{26} - \frac{|D_{\alpha ij}|}{|D_{\alpha i}|} \cdot \log_2 \frac{|D_{\alpha ij}|}{|D_{\alpha i}|}$

These four strategies are combined together for efficiency and it is the best stemming algorithm by performance wise.

STOP WORD LIST

Customised Stop Word List

Combination of

Snowball stop word list -Terrier stop word list and minimal stop word list together and make a customized list for efficient removal of stop words.

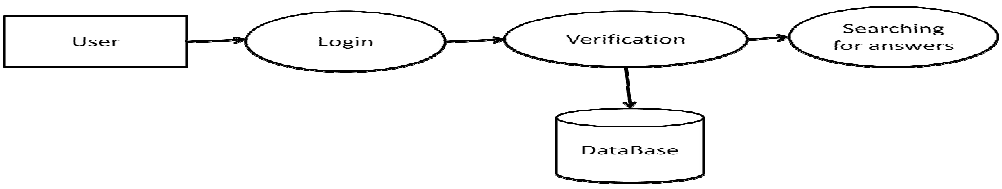
i)*Snowball stop word list* – this stop word list is associated with the Snowball Stemmer

ii)*Terrier stop word list* – this is a simple and précised stop word list associated with the Terrier package.

iii)*Minimal stop word list* – this is a stop word list that made consisting of determiners, coordinating conjunctions and prepositions and some question words.

USER INTERFACE

New user register to retrieve information in our project. Registering contains User name, password, mail-ID and phone number. Mail-ID is used for login as it's an unique field and if the user forgets his password a mail is sent to the user for change of password in a secured manner. The user can enter the web-page after the verification of the user. After login, web user can enter the search space page. In this environment, the user searches the content from the web server. This Search Space is the interface for the user and web server. The input is given as a query by user for gaining knowledge by searching relevant information.



DATA PRE-PROCESSING

For data pre-processing, two main methods are used namely,

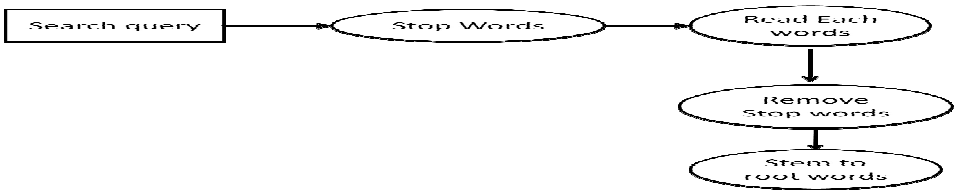
CUSTOMIZED STOP LIST:

For the given query by user, stop word is filtered. Stop words means removal of less weighted words like 'who', 'what', 'how', 'a', 'is', 'the', etc which is done by pre-processing. It is controlled and implemented using customized stop list.

SUCCESSOR STEMMING ALGORITHM:

After stopping process is done, another process of stemming is implemented. Stemming of word is the removal of suffix to get the root word. Stemmers employ a lookup table which contains relations between root forms and normal forms. To stem a word, the table is queried to find a matching word.

If a matching word is found, the associated root form is returned. Eg: A stemming algorithm reduces the words "running" and "runner" to the root word, "run".



ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 3, March 2017

ONTOLOGY CLUSTERING

Words ending in 'nyms' are often used to describe different classes of words, and the relationships between words.

Synonym: A word which have similar meaning.

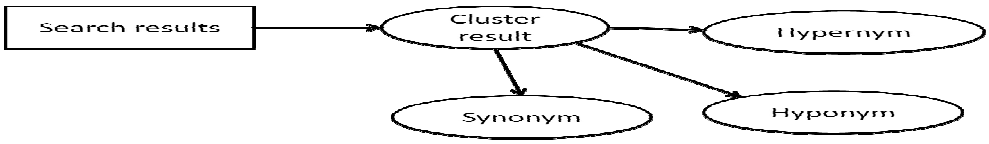
For example, one synonym of sad is "gloomy" however, this word carries quite a negative connotation. Depending on the circumstance you can use it, but if you just want to say that someone is "down," then another synonym such as "blue" or "unhappy" would be more applicable.

Hypernym: A word that has a more general meaning than another.

For example, the word "color" is a hypernym for red, green, blue, brown, etc.,

Hyponym: A word that has specific meaning.

For example, it is the specific meaning of hyponym like red, green, blue, brown, etc.,



MULTI TERM SEARCH

The term "multiple word search" is applied to written words that can have more than one use or definition, and the user can choose the word from different meanings. "Homographs**"** are words that are written the same, but have different meanings and pronunciations. Example : *lead* (the metal) and *lead* (the verb for going ahead). Without the sound difference, it may need to be clarified.

CLUSTER THE MOST RELEVANT CONTENT

Clustering sentences by keywords. SIMBA produces a generic summary. Thus, the keywords that represent the global topic within the collection of texts are identified. The candidate keywords list contains common and proper names. It is built considering the lemma of the words, to ensure that the words in the collection are unique. There-after, the list is ordered considering the score of each word. in the clusters that represent the first set of keywords. These sentences are considered more significant than the others, since they address the main topics conveyed by the collection of texts.

BOOKMARK

Local and global book marking can be done for the page user using and it is very helpful as it is not browser dependent and book marks are stored in database.

This can overcome the drawback in real browsers since in real browsers can only book mark in one particular browser and if it is needed in another browser it has to import the book marks. Instead of that we can save book marks in database or user's folders.

V.EXPERIMENTAL RESULTS

The performance graph is calculated and displayed based on the analysis of the previous survey that we have referred and the performance of the system that we have proposed with additional features. The processing of search, information retrieval and bookmarking is taken as the component that is measured. And it is represented in a bar graph below rated to the highest measure of 10.

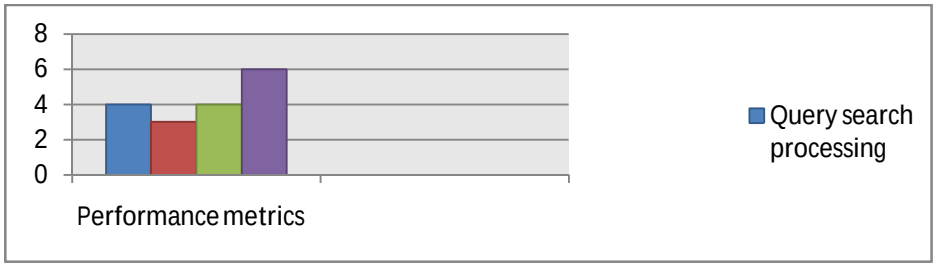
ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 3, March 2017



	Query search processing	Data Preprocessing	Page Bookmarking	Result
Performance metrics	4	3	4	6

VI.CONCLUSION

Thus, we hereby conclude by saying that our project helps to search about unaware fields in an efficient manner. It is feasible for all kinds of user. Finding synonym, hypernym and hyponym for keywords by natural language processing is a challenging task as we have to eliminate irrelevant words and find only for similar useful words. Further to make stemming process more effective we defined a new combined strategy using existing strategies for successor stemming algorithm which is somewhat complex but gives high performance in retrieving relevant links. Information retrieval through images additional to keywords can be done in order to make searching more interactive and user friendly. We can make information retrieval more secured by displaying only websites which is of high security like https protocol websites alone. And for websites which is not secured we have to get security key through Kerberos mechanism.

VII. ACKNOWLEDGEMENT

A research paper of this magnitude and nature requires kind co-operation and support from many, for successful completion. We wish to express our sincere thanks to all those who are involved in completion of this paper. Our sincere thanks to our institution PANIMALAR ENGINEERING COLLEGE and our institution head, principal, secretary, hod and the project guide who helped us to complete our paper and also to my teammates who supported and contributed to complete this paper.

REFERENCES

[1] C. J. Jiang et al., “An indexing network model for information servicesand its applications,” in Proc. 6th IEEE Int. Conf. Serv. Orient.Comput.Appl., Koloa, HI, USA, pp. 290–297, 2013
[2] H. Chim and X. Deng, “Efficient phrase-based document similarity for clustering,” IEEE Trans. Knowl. Data Eng., vol. 20, no. 9,pp. 1217–1229, Sep. 2008.yst., vol. 44, no. 1, pp. 59–71,Jan. 2014.
[3] L. Kaczmarek, “Interactive query expansion with the use of clustering-by-directions algorithm,” IEEE Trans. Ind. Electron., vol. 58,no. 8, pp. 3168–3173, Aug. 2011.
[4] Zhu, H., Chen, E., Xiong“Ranking user authority with relevant knowledge categories for expert finding” World Wide Web Volume 17, Issue 5, pp 1081–1107, September 2014
[5] R.K. Roul, O. R. Devanand, S.K. Sahay“Web document clustering and ranking using tf-idf based apriori approach “IJCA Proceedings on ICACEA, No. 2, p. 34 , 2014
[6] HoChung Wu,”Interpreting tf-idf term weights as making relevance decisions “ACM Transactions on Information Systems (TOIS) ,Volume 26 Issue 3 ,Article No.1, June 2008
[7] M. T. Mills and N. G. Bourbakis, “Graph-based methods for natural language processing and understanding—A survey and analysis,”IEEE Trans. Syst., Man, Cybern., S , 2014
[8] Li, J., Liu, C., Zhou “XML keyword search with promising result type recommendations”World Wide Web (2014), Volume 17, Issue 1, pp 127–159 ,19 January 2014
[9] HeikoPaulheim,”Exploiting Linked Open Data as Background Knowledge in Data Mining” ceur-ws.org/Vol-1082, 2013
[10] M. Tenorth, A. C. Perzylo, R. Lafrenz, and M. Beetz, “Representation and exchange of knowledge about actions, objects, and environments in the RoboEarth framework,” IEEE Trans. Autom. Sci. Eng., vol. 10,no. 3, pp. 643–651, Jul. 2012.
[11] Ji Zhao, Liantao Wang, Ricardo Cabral, and Fernando De la Torre “Feature and Region Selection for Visual Learning” ,IEEE transactions on image processing, vol. 25, NO. 3, MARCH 2016
[12] Zheng Lin, Xiaolong Jin, Xueke Xu, Yuanzhuo Wang, Xueqi Cheng, Weiping Wang, and Dan Meng, “An Unsupervised Cross-Lingual Topic Model Framework for Sentiment Classification” , IEEE/ACM transactions on audio, speech, and language processing, vol. 24, NO. 3, MARCH 2016