Hadoop Based Effective Metadata Indexing for the Improvement of Information Retrieval From E-Learning Assets

Shardul M Upadhyaya, Dr. Shyamal Tanna
,
[1] *Student in Masters of Computer engineering, Department of Computer Engineering L.J. Institute of Engineering and Technology, Ahmedabad, India*
[2] *Assistant professor, Department of Computer Engineering L.J. Institute of Engineering and Technology, Ahmedabad, India*

ABSTRACT

In current information technology era with the growing use of various online educational resources such e-Learning platforms, learning portals, learning object repositories etc. it is necessity to define some approach which works on indexing for this educational resources which can facilitate in the task of knowledge discovery by performing efficient and accurate searching and retrieving operation on this learning resources. Traditional approach use on TF-IDF(term frequency-inverse document frequency) based automatic indexing which works on bag of word model and result in high dimensionality search space To solve this problem we proposed ontology based indexing approach which use OF-IDF (ontology frequency-inverse document frequency) and overcome the shortcoming of traditional indexing approach which is based on TF-IDF. We also develop recommendation engine which recommend the indexing article based on computer programming domain related ontology.

Keyword: - *Keyword Extraction, Information Retrieval, Automatic Indexation, Semantic Web, Ontology Learning Recommendation System, E-Learning.*

1. INTRODUCTION

The measure of information accessible over Web is substantial. It is hard to experience complete text material; along these lines there is a prerequisite of good information extraction and outline strategies which can give substance of a given document in exact way. As keywords are littlest unit of information, they can give a reduced representation of a document's substance. Numerous current methodologies for keyword assignment depend on human indexers for manual assignment of Keywords. Manual keyword extraction utilizes altered scientific classification, is tedious and a troublesome errand; Examination is hence required to concentrate on strategies that can automatically extract keywords from documents. This can go about as a guide to produce rundown highlights for documents that would somehow or another be difficult to reach. The objective of automatic extraction is to improve information revelation and association without the downsides related with manual assignments. In this paper with the help of automatic keyword extraction, and ontology mining, new approach is developed which overcome the traditional TF-IDF based approach by reducing dimensionality by introducing new OF-IDF [7] base approach recommendation engine is built as part of system to get relevant article based on learning domain.

The next few sections of this paper are organized as follows: section 2 presents some Literature Review related to research works that deal with the problem of automatic indexation based on traditional TF-IDF based method. Section 3 presents our proposed ontology based approach for extracting a set of descriptive keywords helpful in the automatic indexing and recommendation of educational resources. Experiments and results are described and discussed in Section 4. Finally, section 5 is dedicated to conclusion and future works.

2. A LITERATURE REVIEW ON VARIOUS METHODS USED FOR AUTOMATIC KEYWORD EXTRACTION IN CURRENT SCENARIO

We saw various approach which enhance the capacity of keyword extraction and indexing and facilitate the searching from very large unstructured web data. In Keywords Extraction for Automatic Indexing of e- Learning Resources Hendez, M. et al. propose an approach to help the indexing operation. This approach consists in automatically extracting a set of relevant terms describing the educational content of a resource it is based on the TFIDF algorithm, the usage of a domain lexicon and exploits the structure of educational documents and ranking It was applied to facilitate the searching and retrieving from online educational resources By Performing automatic indexing [1]. In Keyword Extraction for Mining Meaningful Learning- Contents on the Web Using Wikipedia Toyota, T. Res. et al. present a method for extracting appropriate keywords to identify meaningful learning contents on the Web using Wikipedia. They use PF-IBF method to calculate degrees of association between the articles and the keywords and improve the accuracy of learning-related keyword extraction. They proposed a method that would allow dynamic adjustment to weighted relevance to learning in response to the school year or level of the learner, but accuracy was low for some learning items, indicating that there is still room for improvement. So it useful in application which suggest article based learning area of the student's school year [2]. In Keyword Extraction of Web Pages Based on Domain Thesaurus Guowan H et.al propose a keyword extraction method based web domain thesaurus the method extract keywords based on statistics and choose the selected frequency location and a combination of factors associated with a particular field of the thesaurus to access the weight of the keywords. Overall process overcome the problem of Traditional keyword extraction method only considers word frequency, but does not consider certain areas of the low-frequency vocabulary which contain amount of information[5]. In A Novel Statistical and Linguistic Features Based Technique for Keyword Extraction Gupta, A. et al. proposed the statistical and language model based keyword extraction which overcomes the shortcomings existing solutions, require either training models or domain specific information for automatic keyword extraction. This Approach works on an individual document without any previous parameter adjustment and takes full advantage of all the features of the document to extract the keyword. The extracted keywords can than assist in domain specific indexing. This system consists of two major subcomponents: Keyword Extractor Module and Domain Extractor Module. Keyword Extraction module takes input from document collection. These input HTML pages are analysed for extracting significant keywords from the HTML pages. The candidate keywords are supplied to Domain Extraction module for matching with ontological constructs to determine the domain of the corresponding webpage. The web page along with specified domain and significant keywords is then stored in domain specific webpage repository. The simplicity and efficiency of the proposed algorithm makes it applicable for a wide range of documents and collections and may be useful in any application where there is a need of representing the content of a web page within small set of keywords [3]. In Feature Extraction for Co-Occurrence-Based Cosine Similarity Score of Text Documents Ammar Ismael Kadhim,et.al proposed approach to improve and to reduce the high dimensionality of feature spaces where cosine similarity score lacks in performance to achieve this propose TF-IDF term weighting is used to extract features and Two different weighting methods (TF-IDF and TF-IDF Global)[4].

2.1 LIMITATION AND CHALLENGE OF PREVIOUS APPROACH: -

By reviewing of all previous methods we conclude that TF-IDF depends on the bag-of-words (Bow) model, subsequently it doesn't measure any semantics, co-occurrences in various documents so it results in high dimensionality of feature space for automatic keywords extraction and indexing operation. Because of this reason TF-IDF is just valuable as a lexical level feature-Cannot catch semantics according to domain.

3. PROPOSED APPROACH FOR AUTOMATIC KEYWORDS EXTRACTION AND INDEXING

We propose domain specific ontology to extract relevant Keyword automatic and use semantic Approach called OF-IDF(ontology frequency-inverse document frequency) to deal with reduce high dimensionality and enhance efficiency to indexing operation of learning related documents. We also build recommendation engine which recommend the relevant learning document on the bases of this indexing operation.

This approach is based on two principal steps:

(1) Learning resources pre-processing;

(2) Keywords extraction ranking and Recommendation.

Before beginning of execution collection of e-learning documents in various format collected and used as input dataset in algorithm

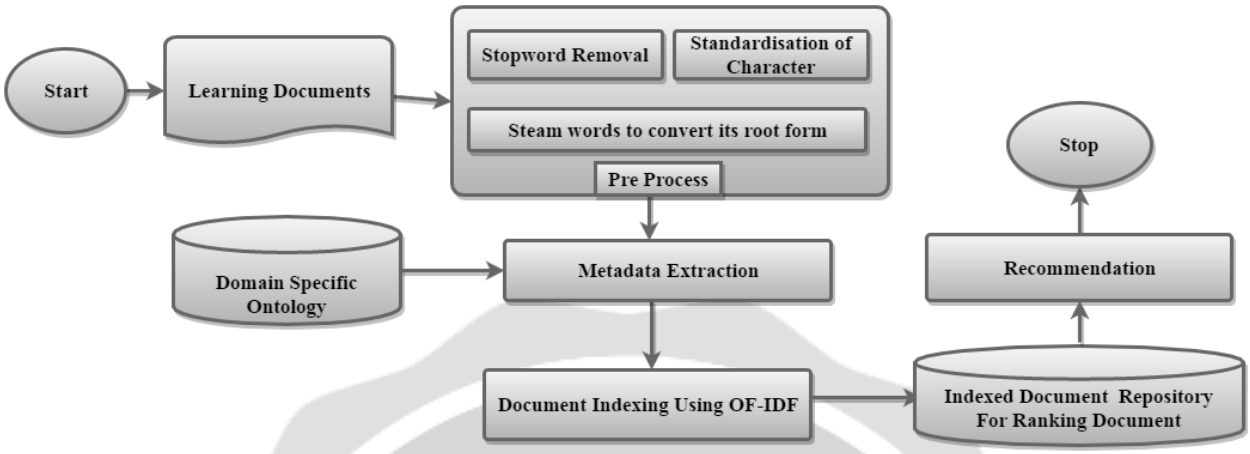

FIGURE -1:-PRPOSED ALGORITHM FOR AUTOMATIC KEYWORD EXTRACTION AND INDEXING

- **Pre-Processing Module:-**

1) Textual content extraction:-

Textual content is extracted from given learning resources by removing extra overhead information for example stop words, advertisement, images etc.

2) Standardization of character case:-

All the words in the corpus of learning resources, furthermore those in the lexicon are changed to lowercase structure.

3) Lemmatization:-

Each word is converted into its root form or canonical form by using steaming algorithm so it become easy to extract candidate term or key term form learning resources

4) Metadata extraction:-

Meta data and relevant keywords from learning repository are extracted from title, Meta tag, abstract by using pre crated learning domain specific ontology.

- **Ranking and Recommendation Module:-**

5) Automatic extraction of index terms:-

After preprocessing step keywords are extracted using ontology frequency and inverse document frequency OFxIDF it performance search in semantic space and generate learning repository with characteristics such as reduced dimensionality and semantic relations between automatically keywords.

6) Ranking The Document:-

In this steps indexing and ranking of learning document related repository. It's calculated by combining OFxIDF with key score and all documents are arranged in descending order of this ratio for future operation. key score is relevance measure it describe the importance of keyword by calculating ratio number of times keywords appears in important part of learning documents such as meta-tag, title tag, unordered-list, ordered list, anchor tag.

7) Recommendation:-

Recommend engine recommend the learning related article on the bases of user query it use ranking mechanism which recommend articles on the bases of the descending order ratio of key score combined with OFxIDF value of keywords.

4. EXPERIMENTS AND RESULTS

In this section, we will first portray the tools we utilized for executing the proposed approach of keyword extraction and for measuring its performance. At that point we will also present and discuss about the results of our experiments.

4.1 Learning Resources Corpus

To test the proposed approach we have gathered a set of 33 files with in form of web pages to be input in our system. This corpus is a collection of web pages describing tutorials of xml courses and programing with c language learning courses available on various e-learning platforms. This collection of corpus covers the various topics like functions, loops, data types, decision making statements in c it also contains collection of various interesting topics of xml like dtd, schema etc. We also build domain specific ontology which specify more than 60 parameters which and describes semantic relationship of various domains of programming in c and xml related learning resources.

4.2 Tools and Technology Used In Experiment

For creation of ontology Protégé is used which is a free, open-source platform that provides a growing user community with a suite of tools to construct domain models and knowledge-based applications with ontologies. Pre-Processing and Ranking Logic Implemented in Hadoop-2.3.0 distributed framework environment by using eclipse kepler version. Recommendation engine build in java language which is implemented by Net Beans IDE 8.0.1.In Java environment jdk version JDk-7u-75 for windows is used. Whole experiments tested on single node system contains Intel core i-7 second generation processor with 8 gb ddr-3 ram on Microsoft windows 10 platform.

4.3 Experiment Analysis

In this segment we introduce the results of our experiments for evaluating the proposed approach. For this, we used the precision recall and F-measures. These three metrics are commonly used to evaluate the performance of information retrieval tools and natural language processing. We are going to compare efficiency of our approach by using these three metrics. We are going to compare Traditional TF-IDF and our OF-IDF based automatic indexing approach both approaches are implemented in Hadoop-2.30 distributed framework.

Recall:-

The recall is used to calculate the ratio between the number of correct annotations found and the total number of correct annotations.

Recall = Number of correct words retained / Total Number of correct words [1]

Precision:-

Precision is the ratio between the number of correct annotations found and the number of annotations extracted by the system. It measures the quality of annotation and indicates to what extent the system is giving correct annotations.

Precision = Number of correct words retained / Number of words automatically extracted [1]

F –measure:-

The F -measure is used to evaluate the performance results based on a weighted combination of the two previous measures. It is considered as a harmonic average for recall and precision.

F-score = recall x precision / (recall + precision) / 2[10]

As the proposed approach of extracting keywords provides a set of ranked index terms by relevance the following table contains the results obtained after two experiments:

Measure	Traditional TF-IDF	Proposed OF-IDF
Recall	0.44510	0.46362
Precision	0.72222	0.75
F-measure	0.54983	0.76205

Table -1: comparison table of traditional approach vs. proposed approach

According to the data in Table 1, the accuracy rate, recall rate, F-Score values can be plotted as a line chart as shown in Figure 2, 3, and 4. The chart shows that as the keyword extraction quantity increasing, accuracy decreases, and the recall rate improves. Comprehensively measuring accuracy and recall rate, the F-Score value in this method is larger than that in the TF-IDF method. Experimental results show that this method achieve better affection in the keyword extraction on e-learning related web page.

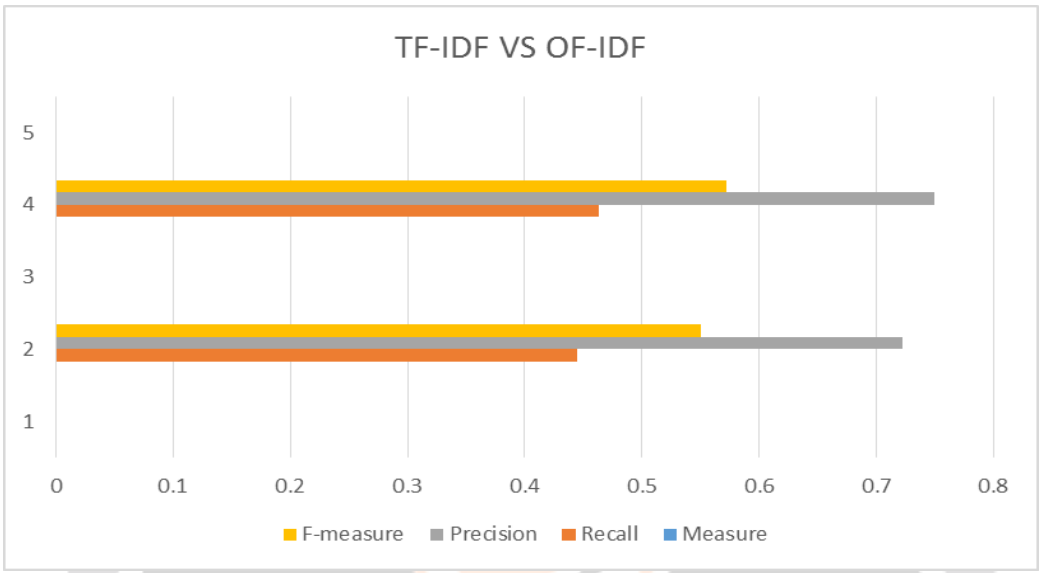

Chart -1:- Various Performance Measure Of Two Approach

From experiment analysis recall precision and f-measure we can conclude that proposed OF-IDF based approach achieve more efficiency, accuracy and less time consuming in comparison of traditional TF-IDF based indexing approach.

5. CONCLUSIONS

Keywords Extraction Plays A Very Important Role In The Information Retrieval Domain, Since The Keywords Can Represent The Asserted Main Point In A Document. Accurate And Effective Method To Find More Relevant Keywords Based On User Query Is Need Of Current Information Era. We Saw that currently widely used TF-IDF based approach suffers with high dimensionality of feature space which is not much efficient so we proposed more efficient approach which is based on OF-IDF and use domain specific ontology to overcome this short coming .This Research Can Be Further Extended By Classification Based Keyword Extraction.. We Can Also Use This With Other Languages Like Arabic, Hindi And Other Language Resources.

6. REFERENCES

[1]. Hendez, Marwa, and Hadhemi Achour. "Keywords extraction for automatic indexing of e-learning resources." In Computer Applications & Research (WSCAR), 2014 World Symposium on, pp. 1-5. IEEE, 2014.1

[2]. Toyota, Tetsuya, and Yuan Sun. "Keyword extraction for mining meaningful learning-contents on the Web using Wikipedia." In Frontiers in Education Conference (FIE), 2014 IEEE, pp. 1-4. IEEE, 2014.

[3]. Gupta, Arpan, Abhishek Dixit, and Arvind Kumar Sharma. "A novel statistical and linguistic features based technique for keyword extraction." InInformation Systems and Computer Networks (ISCON), 2014 International Conference on, pp. 55-59. IEEE, 2014.

[4]. Kadhim, Ammar Ismael, Yu Cheah, Nurul Hashimah Ahamed, and Lubab Salman. "Feature extraction for co-occurrence-based cosine similarity score of text documents." In Research and Development (SCOReD), 2014 IEEE Student Conference on, pp. 1-4. IEEE, 2014.

[5]. He, Guowan, Jie Wang, Yafeng Zhang, and Yan Peng. "Keyword extraction of web pages based on domain thesaurus." In Cloud Computing and Intelligence Systems (CCIS), 2014 IEEE 3rd International Conference on, pp. 310-314. IEEE, 2014.

[6]Achour H., Zouari M. (2013) : Multilingual Learning Objects indexing and Retrieving Based on Ontologies. In Proceedings of ICEELl' 13 - International Conference on Education & ELearning Innovations, 22-24 Ju in,2 013,S ousse - Tunisie.
[7] Ren, R., Zhang, L., Cui, L., Deng, B., & Shi, Y. (2015). Personalized Financial News Recommendation Algorithm Based on Ontology. Procedia Computer Science, 55, 843-851.
[8]InformationRetrievalandTextMining',2015.[Online].Available:http://www.tfidf.com/.[Accessed:27-Nov-2015].
[9]Latentsemanticindexing',2015.[Online].Available:https://en.wikipedia.org/wiki/Latentsemanticindexing.[Accesse d:27-Nov-2015].
[10]DataMiningMiningTextData',2015.[Online].Available:http://www.tutorialspoint.com/data mining/dm mining text data.htm.[Accessed: 26-Nov-2015].