International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 05, May-2016

Extracting and Updating of Learner Interest using WordNET in an E-Learning System

T. Sheeba,
Research Scholar,
Department of Computer Science & Engineering
Karpagam University,
Coimbatore-641 021

Dr. Reshmy Krishnan,
HOD,
Department of Computing,
Muscat College,
Muscat, Sultanate of Oman.

***Abstract*— Keyword extraction is one of the most important themes in E-Learning environments. In this paper a model which could improve the extraction of keywords from the frequently visited documents using WordNet are employed. First of all, learner's words are selected using TFIDF algorithm. Then the related words are semantically represented using WordNet and then updated to the learner profile using semantic similarity using WordNet. The experiment shows the successful extraction and updating of learner interest to the learner profile.**

Keywords*—*E-Learning; Keyword extraction; WordNet

I. Introduction

Compared with the traditional face-to-face style teaching and learning, E-Learning is indeed a revolutionary way to provide education in life long term. E-Learning is said to be an innovative way of learning suited to meet today's learner's learning requirements, particularly as the industrial economy evolves into a knowledge based economy. The major challenge in E-Learning system is that the system is not able to satisfy the interest of online learner and the learner is also unable to extract correct information on the web based on their interest. In order to satisfy the diverse range of requirements of online learner and to design an adaptive learning system, there is necessity to extract the learner interest that would reflect the true learner needs. Several methods have been proposed for extracting the keywords from the documents. Keyword extraction is a process in which a limited number of words are selected to induct the whole document purpose. This process should be done in a systematic manner and by at least or no human interferences. Various methods have been used in the extraction of keywords: statistical, linguistic, machine learning and hybrid methods. Statistical methods are simple and don't need training data. [1] used NGram method which is a statistical method for automatic document indexing. Other statistical methods such as term frequency, TFIDF and word co-occurrence [2] can also be used. The main advantage of using word co-occurrence is its simplicity and high performance compared to tfidf. The main benefits of statistical methods are their ease of use and also their good results. Linguistic methods pay attention to linguistic features such as part of speech (name, verb, adjective), syntax and semantic. [3] examined different methods in keyword extraction such as term frequency, inverse document frequency and relative position of first use of keywords and part of speech label using linguistic features. The results show that using such techniques will significantly improve automatic keyword extraction process. [4] used machine learning techniques to improve keyword extraction process. It uses simple methods such as naïve Bayes which is applicable significantly in not changing the quality. Other methods of machine learning are available which are more complex and have higher computational cost. [5] has applied natural language processing techniques for keyword extraction from radio news. Encyclopedia and journal papers were used as resources to determine the keywords relations. [6] used statistical methods to find the keywords. Irrelevant words are removed using a dictionary of mathematics. The result shows improvement in keyword extraction, but the most important weakness of this method is the need for determination documents subjects before processing them and the dictionary in that field should also be provided in advance. The main drawback of the above said existing methods of keyword extraction either need training data or specific knowledge in that field. WordNet based keyword extraction in E-Learning: Keyword extraction and concept finding in learning documents is one of the most important subjects in E-Learning environments. [7] presented a novel model to improve keyword extraction in learning objects. For this reason, text mining methods accompanied by ontology of WordNet dictionary are employed. First, keywords are selected using TFIDF algorithm. Then unrelated concepts are deselected by proposed algorithm using WordNet dictionary. The remaining concepts having highest similarity are considered as the most important keywords in the provided learning object. WordNet based Learner Interest extraction in E-Learning: In this work [8], interested terms of the learner are extracted by analyzing the web log using the method Vector Space Model (VSM) which is used to extract feature from document. In VSM, each document is identified by n-dimensional feature vector in which each dimension corresponds to a distinct term. The term frequency used to reveal the importance of term within a particular document. The drawback of this method is that is does not represent semantics to the terms i.e. to take into account the meanings of log terms and the semantic relations between them. A fuzzy clustering method and statistical K-means clustering method is used for the prediction and clustering of learners based on the e-learners interests. This paper recommends to use ontology based user profiles to maintain sophisticated representations of personal interest profiles. In this work [9], user profile is created by collecting information from different search space such as user's blog, personal/organizational web page, and any other cites. To extract features from documents, methods such as WordNet and Lexico-Syntactic pattern for hyponyms were used. Further improvement is applied to the constructed profile by taking

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

collaborative user methods in which group of users with similar interest is determined by taking similarity score among them. An ontology matching approach is also applied in order to learn the profile with other similar users. In [10], a new method to develop and maintain a user profile is created in a "music" domain by analyzing user's web access behavior. The user's current interests and new interests are extracted from the analyzed user web logs. A method of identifying new terms that can be of highest relevance to user interests is done with the help of an importance measure. This measure is combined with ontology based semantic measure which compare items browsed by a user on the web with the items from a user's profile. The proposed relevancy measure is applied for a process of updating a user profile in the music domain. In [11] & [12], each user profile is built using the learning objects published by the user based on the user's interests and preferences. This user profile is built using ontology and an algorithm to construct this profile automatically is also proposed. This work first constructs a fuzzy ontology by calculating the relatedness degree for each pair of two distinct terms. Then appropriate tests are done to select the association which will be incorporated into the ontology. The entire approach has been successfully integrated into the management tools for learning objects in AGORA platform. The result shows that this fuzzy ontology has demonstrated as a good representation of the user's interests and preference. This research has suggested improving the user profile quality, using a pruning process to avoid concepts which have no significance.

As a conclusion, the proposed work is to develop an adaptive learning in E-learning system in which learner interest is extracted using the ontology-based representation of WordNet which is an improved method of describing the semantics of the word.

The rest of this paper is organized as follows: Section 2 presents the methodology. Section 3 presents the conclusion and finally Section 5 for references.

II. METHOLOGY

Learner interest can be obtained from the documents visited by the learner by performing web log analysis. The steps used to retrieve learner interest from the web log files are shown in Fig 1. Initially weblog files are first preprocessed to select the frequently visited documents of the learner. Next step is to represent the documents selected by the learner in an effective way. Ontology-based representation using WordNet is used to identify semantic concepts related to document terms. The main advantage of selecting WordNet for the representation is that: It is one of the richest thesauruses which contains around 100,000 terms, organized into taxonomic hierarchies and it comprises of more than 150000 synsets. It is compatible with a large number of dictionaries and other semantic sources e.g. DBpedia. Finally the extracted terms represent the true learner interest. The extracted terms are updated in the learner profile using semantic similarity method using WordNet.

The proposed approach was experimented on online learners taking courses in computing department. They were involved in on course created in Moodle. This course has rich course topics that can be represented using different component forms provided by Moodle LMS such as texts, ppts, hyperlinks, quizzes, assignments. Also, interaction between learners is possible by using forums and chats so learner can post topics and reply to others.

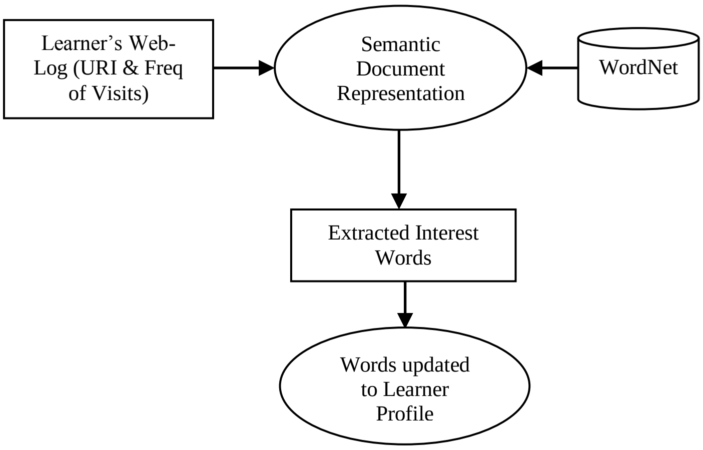
Learner's Web-Log (URI & Freq of Visits)
Semantic Document Representation
WordNet
Extracted Interest Words
Words updated to Learner Profile

Fig. 1. Extracting Learner Interest

A. Data Collection

First of all, the web log files visited by the learner are collected and then analyzed in order to get the frequently visited documents of learner.

B. Document Preprocessing

1) Removal of unwanted words:

The main goal of this step is to prevent unwanted words from being selected as keywords from the frequently visited documents. This step would help to remove some irrelevant words from the list of keywords. The preprocessing is done in order to remove stop words, extra spaces, number etc. Stop words are extremely common words which would appear to be of little value in helping documents matching a keyword. Some of the stop words which are excluded from the documents are "a", "about", "are", "as", "because", "before", "by", "from", "him", "I", "our", "own", "yours" etc. Next special characters are excluded from the selected documents which are control characters and not a letter, number, symbol, or punctuation mark. The special characters considered are "!", "$", "*", ":", ";", "+" etc. In addition to that numbers, extra spaces also excluded from the documents.

2) Calculate Tf-idf (term frequency-inverse document frequency):

Next step is to calculate the Tf-idf, which is the term most often used in information retrieval and text mining to measure the importance of a word is to a document. The importance may proportionally increase as the number of times the same word appears in the document. The main advantage of using this method is that the words which are distributed over most of the documents in the same manner would not be considered and the words which are used a lot in specific subjects would also not be selected.

Computation of Tf-idf involves three main steps: first step is to construct a document term matrix that describes the frequency of words that occur in a collection of documents as shown in Table I.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

Second step is to compute normalized term frequency (TF). This measure is used to calculate how frequently a term occurs in a document using the formula given in (1).

TF(w) = (Number of times word w appears in a document) / (Total number of words in the document) - (1)

Next step is to calculate Inverse Document Frequency (IDF), which measures how important a term is. Term frequency calculation considers all terms as equally important which may include some frequent terms like "is", "of", and "that" etc but have little importance. So, it is necessary to weigh down the frequent terms using inverse document frequency (IDF) which is computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears as given in Eq (2).

IDF(w) = log_e(Total number of documents / Number of documents with word w in it) -(2)

Final step is to get the weight of words in the selected document. The output of this step is the weight of words in selected document that can be used to build learner interest profile which is calculated by multiplying TF(w) * IDF(w).

database which is available online, and provides a large repository of English lexical items. It was designed to establish the connections between four types of Parts of Speech (POS) - noun, verb, adjective, and adverb. The smallest unit in a WordNet is synset, which represents a specific meaning of a word. It includes the word, its explanation, and its synonyms. The specific meaning of one word under one type of POS is called a sense. Each sense of a word is in a different synset. For one word and one type of POS, if there is more than one sense, WordNet organizes them in the order of the most frequently used to the least frequently used (Semcor). Synsets are connected to one another through explicit semantic relations. Some of these relations (hypernym, hyponym for nouns, and hypernym and troponym for verbs) constitute is-a-kind-of (holonymy) and is-a-part-of (meronymy for nouns) hierarchies. The various types of semantic relations used are holonymy vs. melonymy (is part of), hyponymy vs. hypernymy (is a), etc. It can be used as an ontology by using the word sense nodes as concepts or entity types or classes, based on the formalism.

	network	topology	wireless	interface	cable	devices	logical	installed	perform	physical	equipment	access	security	standards	lans
Document 1	17	24	0	2	5	5	10	1	2	12	2	5	0	0	0
Document 2	7	0	15	0	4	0	0	0	0	1	2	2	2	2	3

TABLE I. DOCUMENT MATRIX FOR SAMPLE DOCUMENTS

	network	**topology**	**wireless**	**interface**	**cable**	**devices**	**logical**	**installed**	**perform**	**physical**	**equipment**	**access**	**security**	**standards**	lans
Document 1	0.04	0.06	0	0.01	0.01	0.01	0.03	0	0.01	0.03	0.01	0.01	0	0	0
Document 2	0.01	0	0.03	0	0.01	0	0	0	0	0	0.01	0.01	0.01	0.01	0.01

TABLE II. SHOWS A SAMPLE OF THE WEIGHTED WORDS IN THE DOCUMENTS

3) Semantic Representation Using WordNet

In this step, a list of three words having more weightage is selected for each document. According to our experiments, this list contains most of the keywords selected for each documents. Semantic representation using WordNet is used to improve the performance of traditional VSM (Vector Space Model) model by giving semantic representation to the words selected from documents. It supports the semantic-conceptual relation that links concepts that may be expressed by more than one word. It returns the ordered list of synsets based on a term and the first synset is taken as the identified concept for a term. These concepts from the documents reflect the true learner interest which in turn is used to retrieve the relevant learning contents based on the learner interest. This concept will solve the major challenge of E-Learning system. The representation of learner profile based on WordNet will enhance acquiring from learner profile and is profoundly necessary in an E-Learning system.

This paper used semantic representation using WordNet for the representation of document. This method improves the semantic relation relations between words thereby enhancing the representation of documents. WordNet is a lexical

The ordering is supposed to reflect how common it is that the term is related to the concept in standard English language.

The purpose of this step is to identify WordNet concepts that correspond to document words. For example, in the above experiment the words such as topology, network, logical and physical are considered for the first documents and the words such as wireless, cable are considered for the second document. WordNet returns an ordered list of synsets based on a term. More common term meanings are listed before less common ones. The semanctic relations hypernyms and hyponyms are considered. Hypernyms of concepts can represent concepts with a broad meaning constituting a category into which words with more specific meanings fall and hyponym represent a a word of more specific meaning than a general or superordinate term applicable to it. Table III shows the synonyms, hypernyms and its hyponyms obtained from the WordNet 3.1 database.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

Words	Synonyms	Description	Hypernyms	Hyponyms
Topology	Network topology	the configuration of a communication network	Configuration , constellation	Bus topology, loop topology, star topology, mesh topology, physical topology and logical topology
Network	Broadcasting	a communication system consisting of a group of broadcasting stations that all transmit the same programs	Communication system, Communication equipment	-
Wireless	Radio Communication	Transmission by radio waves	Telecommunication, telecom	Radio telegraph, radio telegraphy, wireless telegraphy

TABLE III. WORDNET 3.1 DATABASE

4) *Update Learner Profile Using WordNet*

In order to update learner interest with new keywords, semantic similarity representation using Wordnet is used. Semantic Similarity a method used to measure the semantic similarity between two concepts. In this proposed approach, ontology-based semantic similarity using WORDNET is used to calculate the semantic relatedness of word sense using the information content of the concepts in WordNet . In this method, the words with a higher score are updated in the learner profile after the learner select new specific keywords. The words that are more frequent have a higher score. The main steps of this method can be described as follows: If a new keyword is found for a learner, first step is to get the synonym set from WordNet and to insert every synonym into the learner profile if it is not available. For inserting into learner profile, first similarity is calculated between all input fields and the new value using information concepts in WordNet. Then the information content similarity (sim) of two concepts c1, c2 is calculated using the formula in (3).

$$sim(a1,a2)=2log(p(c))/log(p(c1)+log(p(c2)) \qquad - (3)$$

If the information content of both the concepts are zero, then similarity score is zero due to lack of data. Ideally the information content would be zero only if that concept was the root node. If the information concept of both concepts is not zero and having a similarity score of threshold less than 0.5 will be updated in the learner profile.

New terms	semantic similarity score	Learner profile terms
network	0.3787	telecommunication
wireless	0.7319	network
wireless	0.8991	telecommunication
topology	0.0000	Data mining
computing	0.0000	computers
engineering	0.0759	software

TABLE IV. SEMANTIC SIMILARITY LEVELS FOR NEW TERMS COMPARED WITH LEARNER PROFILE TERMS

III. CONCLUSION

This paper proposes to extract learner interest from the documents frequently used by the learner and update to a learner profile in an E-Learning system. The result shows the extraction and updating of learner interest to learner profile. The paper uses WordNet for semantic representation of documents and ontology based semantic similarity using WordNet for updating the learning profile which is experimented successfully. From the conducted experiments, we conclude the successful completion of extraction and updating of learner interest to the learner profile.

IV. REFERENCES

[1] Salton, G. (1989). Automatic text processing. Chapter 9, accessed on April 2016.

[2] Matsuo, Y., & Ishizuka, M. (2004). Keyword extraction from a single document using word coocuurrence statistical information. International Journal on Artificial Intelligence Tools.

[3] Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge.

[4] Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C., & Nevill-Manning, C. G. (1999). Domain-specific keyphrase extraction. Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence.

[5] Suzuki, Y., Fukumoto, F., & Sekiguchi, Y. (1998). Keyword extraction of radio news using term weighting with an encyclopedia and newspaper articles. SIGIR.

[6] Günel, K. (2010). Extracting learning concepts from educational texts in intelligent tutoring systems automatically.

[7] Ahmad A. Kardan, Farzad Farahmandnia, Amin Omidvar, “A novel approach for keyword extraction in learning objects using text mining and WordNet”, Global Journal of Information Technology, Volume 03, Issue 1, (2013) 01-06.

[8] Khaled M. Fouad, “Adaptive E-Learning System based on Semantic Web and Fuzzy Clustering”, International Journal of Computer Science and Information Security, Vol. 8, No. 9, 2010

[9] D. Trong, N Mohammed, L. Delong , and J. Geun. (2009), A Collaborative Ontology-Based User Profiles System, N.T. Nguyen, R. Kowalczyk, and S.-M. Chen (Eds.): ICCCI 2009, LNAI 5796, pp. 540–552, Springer-Verlag Berlin Heidelberg.

[10] Marek, “Updating User Profile using Ontology-based Semantic Similarity”, FUZZ- IEEE, August 20-24, 2009.

[11] Mateus, Francisco, Victor, Alfredo, Manuel, “A Fuzzy Ontology Approach to represent User Profiles in E-Learning Environments” IEEE, 2010.

[12] Mateus Ferreirs-Satler, “Fuzzy ontologies-based user profiles applied to enhance e-learning activities”, Springler-Verlag, November 2011.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)