ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.4

MOOCLink: Building and Utilizing Linked Data from Massive Open and Online Courses

Pranali Subhash More[1], Swapnali Suresh Sakpal[2], Raj Anil Shinde[3], Aishwarya Ashok Kulaye[4]
BE Student[1, 2, 3, 4]
Department of Information Technology
Terna Engineering College, Mumbai, India

Abstract:
Due to the widespread of internet, online education is gaining popularity. In the field of education, Massive Open Online Courses (MOOC) are used in delivering learning content to any person who wants to take the course with no constraint on attendance. The courses by different providers may differ in session timing, price, difficulty level etc. Hence a user has to visit every MOOC provider's site and go through the course details. To make this task user-friendly, a Information aggregator is used which can aggregate online courses from multiple course providers. Before aggregating this courses from different MOOC's , data pre-processing is performed. And to combat the limitation of stemming, we use lemmatization. In Information Retrieval , one of the important task is retrieving the relevant information. However an important issue for retrieval effectiveness is the mismatch problem wherein the indexers and users do not often use the same words. Query is often too short and may not contain relevant terms. This issues are handled by query expansion. User query terms like synonyms using a dictionary or WordNet .

Keywords: semantic application; linked data; education; ontology engineering; information aggregator.

I. INTRODUCTION

Internet is an important technology of the information age. It serves as a large reservoir of data from which one can retrieve required information. However, information available on Internet is not stable. At any time this information may be altered, moved or deleted which leads to a problem of finding relevant information on internet. One of the problem web is facing today is information overload. There are a large number of information sources over the web which provide similar or related information for a particular topic. It is the users job to go to each of these sources and get the required information. To effectively use this data from multiple sources it needs to be aggregated at one place.

Benefits of MOOC are as follows:

1. It helps students to find a right course.
2. Courses are offered for free.
3. Courses are available to large and diverse audience across the globe.
4. It provides easy access to global resources and promotes sharing of ideas and knowledge.
5. It enhances active learning.

In the domain of education, there are a large number of MOOC providers such as Coursera,Udacity,Udemy etc. MOOCs are Massive Open Online Courses. They act as a medium for collaborative sharing of knowledge and unlimited participation via web. Each of these course providers may be offering similar courses at the same time. Therefore, if a user wants to take up a particular course, he has multiple choices. The courses by different providers may differ in session timing, price, difficulty level etc. Hence a user has to visit every MOOC providers site and go through the course details. To make this task userfriendly, a Information aggregator is used which can aggregate online courses from multiple course by extracting data from MOOCs providers.

Advantages of Information Aggregator:
1. It enables the users to efficiently search courses by different course providers at one place.
2. It saves time as the user doesn't have to go around different websites and compare the on-line courses provided by different MOOC providers.

However, retrieving relevant information is one of the important tasks. Information retrieval is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections of data .Traditional keyword based search in information retrieval may not fetch all the relevant results, whereas semantic search is based on meaning of terms. Whenever user enters the query in search box information retrieval process begins. Web search is an application of information retrieval techniques to access large amount of data anywhere. One of the important issue for retrieval effectiveness is the mismatch problem wherein the indexers and users do not often use the same words. This is termed as the vocabulary problem. To deal with this problem, many methods have been proposed that includes interactive query refinement, word sense disambiguation, relevance feedback and search results clustering. One of the most successful techniques to overcome this problem is query expansion. Query expansion is currently considered as an extremely promising technique to improve the retrieval effectiveness. For instance, consider user query 'car' then the expanded query is 'car cars automobile automobiles auto' etc.

A. Applications of query expansion

1: Question Answering (QA)- QA faced a problem of mismatch between question and answer vocabularies. Query expansion using lexical ontologies such as WordNet can be used to resolve this problem.

2: Cross-language information retrieval -Retrieving documents written in a language which is other than the language of the user query. However, there are many limitations due to

untranslatable terms, translation ambiguity between the source and target languages and insufficient coverage. To combat this errors query expansion is used to get better results.
3: Information Filtering- Information filtering (IF) is the process of monitoring a stream of documents and selecting those documents that are relevant to the user. Better queries can be learned using query expansion based on similar users or links.
4: Text categorization: Using query expansion with text classification attempts to capture more relevant documents or information.

B. Scope
1: Textual data about the courses from MOOC websites can be extracted using APIs and web crawlers.
2: Categorization of data collected: Manually categorizing courses into different categories is time consuming and error prone. Therefore, this process can be automated by using machine learning algorithms such as Nave Bayes classification.
3: Semantic search by query expansion: Users query may not always contain relevant words to fetch correct courses. Hence users query needs to be enriched with semantically related terms like synonyms.
4: Query Classification: Task of assigning the query to one of the predefined categories based on content of the query.
5: Search Module: Responsible for firing the query to the categorized courses dataset and collecting the results which will be displayed on the user interface.

II. PROBLEM DEFINATION

Information retrieval is the process through which a system can respond to a user's query for text-based information on a specific topic. IR is one of the most important problems in the domain of natural language processing (NLP). Today, keyword search continues to be the most common technique used to search information over web. The limitations in keyword search generate unsatisfactory results when this method is used to search relevant information.

The limitations are as follows:
a. Keyword search possesses the weaknesses of finding too few information or finding too many information in search results.
b. Keyword searches may also overlook key information due to human error.
To overcome these problems we will use semantic search based on meaning of terms. To implement this kind of robust semantic search, users query needs to be expanded to include more meaningful terms which will be done using query expansion. One of the first steps in the information retrieval is stemming. A stemming algorithm aims at obtaining the stem of a word that is its morphological root. In stemming, two main errors have been identified: Stemmer clears the terminations of the terms. Sometimes when the stemmer deletes only the terminations that are almost certainly a suffix that must be cleared, there is a risk of keeping some more complex suffixes or ambiguous terminations in the remaining stem that should also be stripped to obtain the morphological root of the word. This error is called under-stemming. Stemmer makes an over-stemming error when it strips more terminations than it should, 11 Efficient Search over Information Aggregator clearing parts of the word that belong to the morphological root. When this error happens, there's a loss of linguistics data as a part of the morphological root is deleted . This problem can be handled by using lemmetization technique which makes use of vocabulary and morphological analysis of words. It aims at removing inflectional endings only and returns the dictionary form of a word which is known as the lemma.

A. Proposed system
This section presents overview of the proposed system and its components:

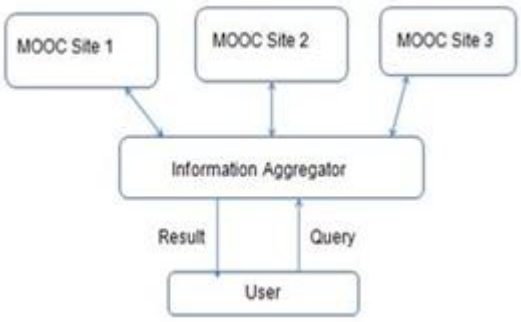


Figure.1. System level architecture

System Level Architecture:
Fig.1 consists of two main components: Information Aggregator and Efficient Search Module Using Query Expansion. Information aggregator collects information from different MOOC websites and utilizes the data to discover and compare online courseware. Efficient Search Module using Query Expansion to include more meaningful terms and fetch the results. The MOOCs considered are: Udacity , Coursera and Udemy. B. Detailed Architecture Detailed Architecture Components consists of the following modules:

1.Data Extraction Module:
It is responsible for extracting the data from MOOC websites and delivering the extracted data to a MySQL database.
It consists of 3 sub modules: Coursera
Data Extraction: Data will be fetched from Coursera using Coursera API.
Udacity Data Extraction: Data will be fetched from Udacity using Udacity API.
Udemy Data Extraction: Data will be extracted from Udemy website using web crawler.

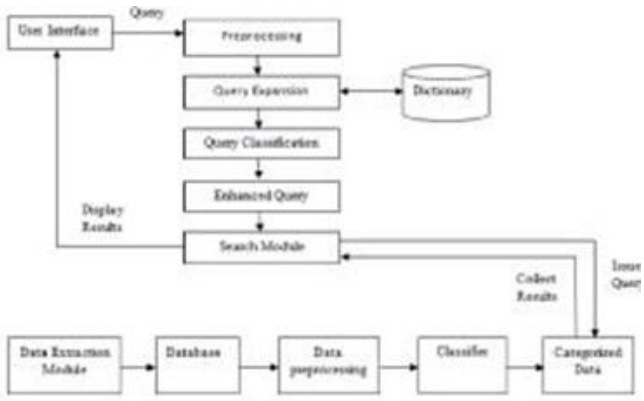


Figure.2. Detailed Architecture

Data Preprocessing:
The data needs to be cleansed before the classifier is applied.

Stop words:
Stopwords are removed from data to reduce noise. Stopwords are the common words that carry no information (eg. prepositions, pronouns etc).

Tokenization:
It is the method of breaking a stream of text into words, phrases, symbols, or different meaty components referred to as tokens. Lemmatization:
It uses vocabulary and morphological analysis of word and tries to remove inflectional endings.

Advantages of lemmatization over stemming:

a. It returns words to their dictionary form.

b. It analyzes if query words are used as verbs or noun.

c. It also helps to match synonyms.

2. Classifier:
Using a machine learning algorithm such as Naive Bayes classifier the courses data from MOOCs will be categorized uniformly into predefined categories. When we have large amount of data available, Classification is an essential task for efficient information management and retrieval. Text Classification is the problem of assigning a text to one or more categories based on the content of the text. Classification can be of following types depending on the number of classes present .

a. Binary Classification: Classify the input object into one of two classes..

b. Multiclass Classification: Classify the input object into one of multiple classes.

Based on the number of classes that can be assigned to an instance, classification can be divided as follows:

a. Single-label classification: Only one class label can be assigned to input object.

b. Multi-label classification: More than one class labels can be assigned to input object.

For text/web page classification, Naive Bayes classification algorithm is widely used. This is a supervised learning algorithm. In Naive Bayes, algorithm has been used for web page classification. This is a probabilistic classifier because it uses probability to find out the target class which maximizes the posterior probability that a particular hypothesis holds true for given data and in case of web page classification, this hypothesis refers to whether a web page belongs to a category or not.
There are two variations of Naive Bayes that are generally used for text classification: Multinomial Naive Bayes and Multivariate Bernoulli model. Assume C is the category set. Naïve Bayes classifier assigns a document D to class label y such that, given a document D, the posterior probability of class Ci i.e. P(CijD) is maximum. Thus the decision rule can be given as following:

y= argmax P(CijD) where Ci is the target category.
s. When the user enters the query, it is expanded using a set of words given in the dictionary. The simplified formula is: y=argmaxP(Ci) _P(tkjCi) where P(Ci) is the prior probability ofclassCi, P(tkjCi) is the conditional probability of term tkoccurring in documents of class Ci, (t1; t2; tnd) are the tokens in document D and nd is number of such tokens in D. The parameters P(Ci) and P(tkjCi) are estimated using the training data. The difference between two variations is that Multinomial Naive Bayes considers multiple occurrences of terms where as Bernoulli model ignores multiple occurrences of terms and thus makes many mistakes while classifying large documents[8]. Hence Multinomial Naive Bayes is preferred for text classification task.

3.Naive Bayesian Training:
In the training phase, P(Ci) and P(WijCi) are calculated for all combinations of class Ciandterm Wi based on the training set. P(Ci) denote the probability of class Cj, P(WijCi) denotes the conditional probability of term Wi given class Ci.

4. Query preprocessing: It consists of steps such as tokenization, stop words removal and lemmatization. Lemmatization considers the morphological analysis of the words. Thusly it is important to have definite word references the calculation can glance back at to interface the frame back to its lemma.

5. Query Expansion: User query is often too short and may not contain relevant terms thus resulting in low recall. Thus, in order to make the search more robust, user query terms need to be expanded [1]. This can be done by using a dictionary of semantically related terms . It is possible to expand the queries to improve the efficiency of the retrieval .The indexing step pre-processes documents and queries to obtain keywords (relevant words, also called as terms) to be used in the query. Matching is the process of computing the similarity between documents and queries by weighting terms using the most frequently applied algorithm TF-IDF. The goal is to improve precision and/or recall. Example: User Query: 'display lyrics' so the expanded query: 'reveal lyrics, show lyrics , display poem, display words' etc.

6. Query Classification: It is a task of assigning the query to one of the predefined categories based on content of the query [5]. It decreases the number of documents to be searched and thus will improve the response time. It also uses Naive Bayes for classification of queries to get more better resulty. If the word matches with the columns in dictionary then Naive Bayes is applied over the system. Thus the query is classified in one of the given categories. But if the word does not exist then normal search is performed using like operator.

7. Search Module: It is responsible for firing the query to the categorized courses dataset and collect the results which will be displayed on the user interface. The user query will be analyzed and the results will be displayed. For example, if user enters a query 'Data Science courses' then the results will be displayed from the three MOOC providers at one place.

III. IMPLEMENTATION

This section outlines and provides relevant illustrations of
A .Web Crawler
We retrieve Coursera's course properties via their course catalog API but use screen scrapers for edX and Udacity[8]. This section details the process of writing a Scrapy crawler for edX in Python. After starting a new scrappy[12] project in the terminal. we define the Item or container that will be loaded with scraped data. This is done by creating a scrapy.Item class and defining its attributes as scrapy..

B. Planning and Cost Estimation
The time taken to complete the project is 9 months. To have effectively design and develop a cost-affordable model the Waterfall model was practiced.

C. Requirement gathering and Analysis phase This phase started at the beginning of our project, we had formed groups and modularized the project. Important points of consideration were Requirement accumulating and evaluation segment .This segment started out at the start of our venture, we had fashioned groups and modularized the venture. vital points of consideration were

1: Outline and visualize all of the objectives truely.

2: Gather requirements and examine them.

3: Remember the technical necessities wished after which accumulate technical specifications of numerous peripheral components required.

4: Examine the coding languages wanted for the assignment.

5: Define coding strategies.

6: Analyze future risks / issues.

7: Define strategies to keep away from this risks else outline trade answers to this dangers.

8: Take a look at economic feasibi lity.

9. Define Gantt charts and assign time span for each phase.

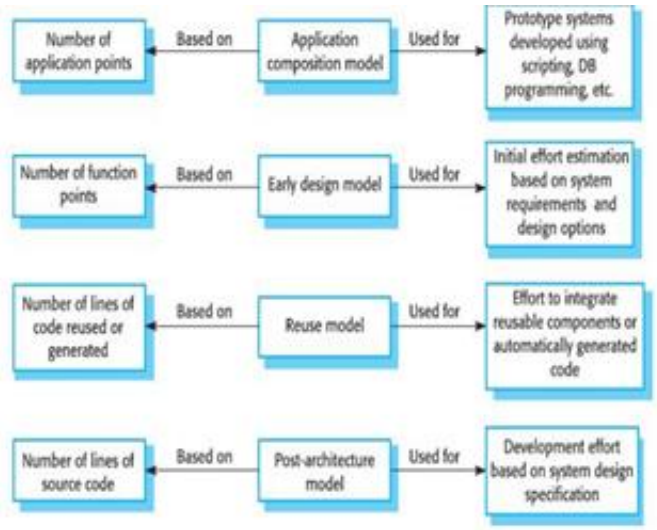


Figure.3. Cost Estimation

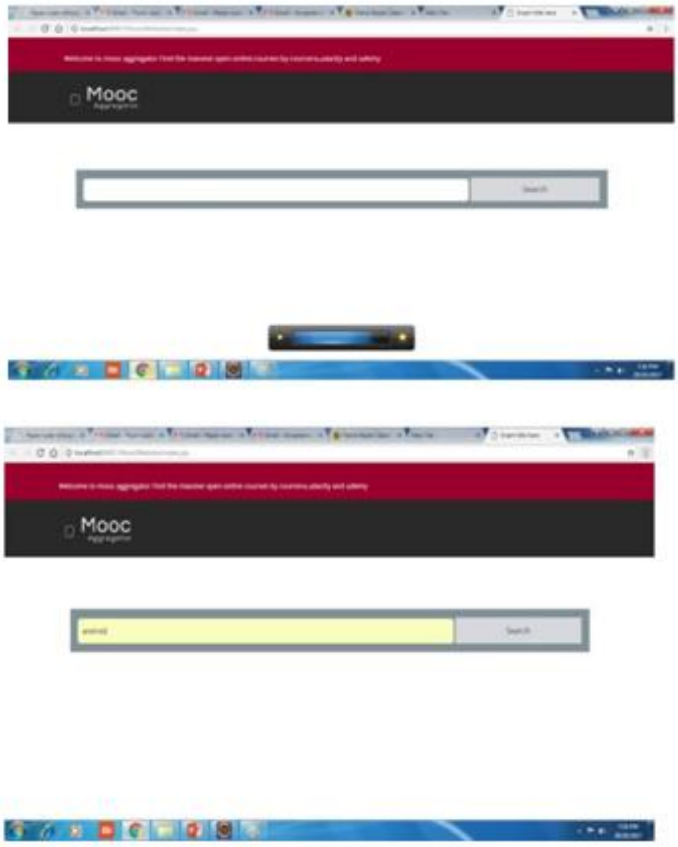
Than trying to decide specific chances of every disaster, a widespread relational score gadget of high, medium and low may be used first of all to identify the chance of the danger going on. The danger analysis also need to decide the effect of every kind of capability threat on diverse functions or departments within the employer. A threat evaluation shape, determined here (PDF format), can facilitate the process. The features or departments will range with the aid of type of corporation .The planning technique ought to become aware of and degree the likelihood of all potential dangers and the effect at the organization if that threat occurred. To try this, each department have to be analyzed one after the other. even though the principle computer gadget can be the unmarried best danger, it is not the handiest critical challenge. Even inside the most automated businesses, a few departments won't be automatic or automatic in any respect. In absolutely automated departments, critical statistics remain out of doors the device, along with criminal documents, computer data, software saved on diskettes, or assisting documentation for records access .The impact may be rated as: zero= No impact or interruption in operations, 1= important effect, interruption in operations for up to eight hours, 2= harm to gadget and/or centers, interruption in operations for eight - forty eight hours, 3= fundamental damage to the gadget and/or facilities, interruption in operations for greater than forty eight hours. All essential workplace and/or pc center capabilities must be relocated. certain assumptions may be vital to uniformly follow scores to every capability hazard. Following are ordinary assumptions which could used:1. despite the fact that effect scores could variety between 1 and three for any facility given a particular set of circumstances, ratings applied ought to replicate anticipated, likely2. even though one potential danger could result in another ability threat (e.g., a typhoon may want to spawn tornados), no domino effect should be assumed. three. If the result of the threat would not warrant movement to an exchange web page(s), the effect need to be rated no better than a "2."four .The risk assessment should be achieved by means of facility .To measure the capacity risks, a weighted factor rating system may be used. every stage of chance may be assigned factors as follows:

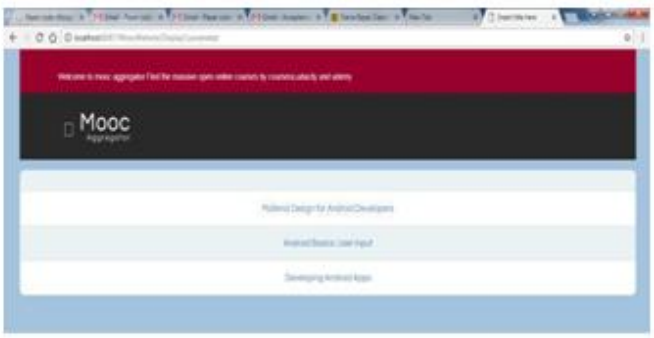
a. Although impact ratings could range between 1 and 3 for any facility given a specific set of circumstances, ratings applied should reflect anticipated, likely

b. Although one potential threat could lead to another potential threat (e.g., a hurricane could spawn tornados), no domino effect should be assumed.

c. If the result of the threat would not warrant movement to an alternate site(s), the impact should be rated no higher than 2. d. The risk assessment should be performed by facility. To measure the potential risks, a weighted point rating system can be used. Each level of probability can be assigned points as follows:

Probability	Points
High	10
Medium	5
Low	1





IV. ACKNOWLEDGEMENT

No project is ever complete without the guidance of those expert who have already traded this past before and hence become master of it and as a result, our leader. So we would like to take this opportunity to take all those individuals how have helped us in visualizing this project and thank them for their guidance and support.

V. REFERENCES

[1]. Bizer , C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. International journal on semantic web and information systems, 5(3), 1-22. Chicago

[2]. Dietze, S., Yu, H. Q., Giordano, D., Kaldoudi, E., Dovrolis , N., & Taibi , D. (2012, March). Linked Education: interlinking educational Resources and the Web of Data. In Proceedings of the 27th Annual ACM Symposium on Applied Computing (pp. 366-371). ACM.

[3]. Coursera Catalog API. (n.d.). - Cow's era Technology. Retrieved July 22, 2014, from https:lltech. coursera.org/ appplatform/catalog/

[4]. IEEE Learning Object Metadata RDF binding. (n.d.). IEEE Learning Object Metadata RDF Binding. Retrieved July 22, 2014, from http://kmr.nada.kth.se/static/ims/md-Iomrd f.html

[5]. Bohl, 0 ., Scheuhase , J., Sengler, R., & Winand, U. (2002, December).The sharable content object reference model (SCORM)-a critical review.In Computers in education, 2002. Proceedings . international conference on (pp. 950-951). IEEE. Chicago

[6]. Learning Resource Metadata Initiative. (2013, April 9). :World's Leading Search Engines Recognize LRMI as Education Metadata Standard. Retrieved July 22, 2014, from http://www.lrmi.net/worldsleading- search-engines-recognize Irmi-as- education-metadata-standard/

[7]. Coursera. (n.d.). Retrieved July 22, 2014, from http:// coursera.org/

[8]. Udacity. (n.d.). Retrieved July 22, 2014, from http:// udacity.com/

[9]. Learning Resource Metadata Initiative. (n.d.) :The Specification. Retrieved July 22, 2014, from http:// www.lrmi.net/the-specification/

[10]. RDF Schema 1.1. (n.d.). RDF Schema 1.1. Retrieved July 23, 2014, from hnp://www.w3.org/TRlrdf-schema/

[11]. Requests: HTTP for Humans. (n.d.). Requests: HTTP for Humans - Requests 2.3. 0 documentation . Retrieved July 22, 2014, from http:// docs. python-requests.org/en/latest/

[12]. Scrapy An open source web scraping framework for Python. (n.d.). Scrapy . An open source web scraping frame work for Python. Retrieved July 22, 2014, from http://scrapy.org/

[13]. Apache Jena. (n.d.). Apache Jena. Retrieved July 22, 2014, from https://jena.apache.org/

[14]. Google App Engine. (n.d.) . URL Fetch API Overview. Retrieved July 22 2014, from https:/ /developers.google.com/ app engine/docs/java/ url fetch