Efficient Data Mining Model for Question Retrieval and Question Analytics Using Semantic Web Framework in Smart E-learning Environment

https://doi.org/10.3991/ijet.v17i01.25909

Subhabrata Sengupta[1(✉)], Anish Banerjee[1], Satyajit Chakrabarti[2]
[1] Information Technology, Institute of Engineering and management, Kolkata, India
[2] Computer Science Engineering, Institute of Engineering and management, Kolkata, India
subhabrata.sengupta@iemcal.com

Abstract—In the field of Information recovery, the fundamental target is to discover important just as most applicable data concerning a few questions. However, the essential issue regarding recuperation has reliably been, that the request an area is enormous so much that it has gotten very difficult to recuperate applicable information capably. In any case, with the latest progressions in profound learning and AI models, calculations, applications brilliant and computerized data recovery component matched with text examination to decide different characterizing boundaries alongside intricacy and weight-age assurance of inquiries. By focusing, the cutoff points and hardships, like CPU cost, efficiency, automation and congruity, we have assigned our information recuperation structure particularly towards Academic Institutional Domain to consider the interest of various association related inquiries. The aim is to make an efficient data mining and analytical model that can automate an efficient question retrieval and analysis for complexity and weight-age determination.

Keywords—data mining, deep learning, smart and automated information retrieval mechanism, weight-age determination

1 Overview

The World Wide Web fills in as an enormous, broadly circulated, worldwide data administration place for news, promotions, customer data, monetary administration, instruction, government, online business and numerous other data administrations. With the unstable development of data accessible on the World Wide Web, different wise web administrations have been created to help client's entrance significant data from the Web. Web search has its root in information retrieval (IR) [1]. Customary IR accepts that the essential data unit is an archive, and an enormous assortment of records is accessible to shape the content information base. Recovering data just methods finding a bunch of reports that is applicable to the client inquiry. A positioning of the arrangement of reports is typically additionally performed by their importance scores to the question [2]. The most ordinarily utilized inquiry design is a rundown of catchphrases,

which are likewise called "Terms". IR is not quite the same as information recovery in data sets utilizing SQL questions on the grounds that the information in data sets is profoundly organized and put away in social tables, while data in content is unstructured [4].

Web personalization is a technique, an advertising instrument, and a craftsmanship. Personalization requires undeniably or unequivocally collecting guest data and utilizing that information in your substance transport configuration to control what data you present to your clients and how you present it [6]. Web mining means to find helpful data or information from the Web hyperlink structure, page substance, and utilization information. Despite the fact that Web mining utilizes numerous information mining strategies, as referenced above it isn't absolutely a use of customary information mining because of the heterogeneity and semi-organized or unstructured nature of the Web information.

The system aims to achieve a smart automated question retrieval mechanism with the implementation of text processing and analysis techniques. Using the different algorithms for key phrase and keyword determination from texts, topic classification and similarity comparison for creating a relevant question bank through determining based on a set of questions fed. After it the acquired questions are classified based on its defining parameters like textual complexity, lexical complexity, difficulty and accuracy rate.

Our aim is to couple the AI and deep learning algorithms for customized learning and personalized information retrieval along with smart learning frameworks like semantic web technologies to influence the E-Learning system.

2 Background study

Novel highlights dependent on reference network data is constructed and utilized related to customary highlights for key phrase extraction to acquire noteworthy enhancements in execution over solid baselines. [7]

Kea, a calculation for naturally separating key phrases from text. Kea distinguishes applicant key phrases utilizing lexical strategies, ascertains include values for every up-and-comer, and utilizations an AI calculation to anticipate which competitors are acceptable key phrases. The AI conspire first forms an expectation model utilizing preparing archives with known key phrases, and afterward utilizes the model to discover key phrases in new documents. [8]

Report is treated as a bunch of expressions, which the learning calculation should figure out how to group as certain or negative instances of key phrases. GenEx calculation explicitly for naturally separating key phrases from text. The test results uphold the case that a hand-crafted calculation (GenEx), fusing specific procedural area information, can create preferred key phrases over a broadly useful calculation (C4.5). Abstract human assessment of the key phrases produced by GenEx proposes that about 80% of the key phrases are adequate to human perusers. This degree of execution ought to be good for a wide assortment of applications. [9]

Text summarization is arisen as a significant examination zone in late past. In such manner, survey of existing work on content rundown measure is valuable for doing promote research. [10] For expand questions, to comprehend the pith of the inquiry text synopsis is truly advantageous.

Programmed keyword extraction is a significant examination heading in content mining, characteristic language handling and data recovery. keyword extraction empowers us to speak to message records in a consolidated manner. An extensive investigation of looking at base learning algorithms (Naïve Bayes, uphold vector machines, strategic relapse and Random Forest) with five generally used ensemble techniques (AdaBoost, Bagging, Dagging, Random Subspace and Majority Voting) is led. [11]

On the off chance that the likelihood conveyance of co-event between term an and the continuous terms is one-sided to a specific subset of regular terms, at that point term an is probably going to be a catchphrase. The level of inclination of a circulation is estimated by the $\chi 2$-measure. Our calculation shows equivalent execution to tfidf without utilizing a corpus. [12]

Graph based strategy is quite possibly the most proficient solo approaches to extricate watchword from a solitary web text. In any case, seldom did the past diagram-based techniques think about the sentence significance. In this paper, we propose a diagram-based watchword extractor WS-Rank which brings sentences into chart where sentences are unmistakably treated by their significance. [13]

The system evaluation is assessed in an unexpected way, including correlation with human annotated keywords utilizing F-measure and a weighted score comparative with the oracle system execution, just as a novel elective human evaluation. [14]

In a supervised system for extricating watchwords from meeting records, a classification that is fundamentally not the same as composed content or other discourse spaces, for example, broadcast news. Notwithstanding the customary recurrence or position-based hints, we research an assortment of novel highlights, including semantically persuaded term explicitness highlights, dynamic sentence-related highlights, prosodic prominence scores, just as a gathering of highlights got from synopsis sentences. [15]

The issue of programmed catchphrase extraction in the gathering area is handled by a sort fundamentally not the same as composed content. For the administered structure, we proposed a rich arrangement of highlights past the average TFIDF measures, for example, sentence remarkable quality weight, lexical highlights, synopsis sentences, and speaker information. [16]

The extractive outline methods as a rule spin around finding generally important and continuous catchphrases and afterward extricate the sentences dependent on those watchwords. Manual extraction or clarification of applicable catchphrases are a dismal methodology flooding with blunders including heaps of manual effort and time. [17]

The appropriation of EDM by advanced education as a logical and dynamic apparatus is offering new occasions to misuse the undiscovered information created by different student information systems (SIS) and learning management systems (LMS). This paper portrays a half breed approach which utilizes EDM and relapse examination to dissect live video streaming (LVS) understudies' internet learning practices and their presentation in their courses. Understudies' collaboration and login repeat, similarly as

the quantity of talk messages and questions that they submit to their teachers, were inspected, close by understudies' last grades. [18]

3 Issue articulation and proposed methodology

Data recovery is the examination of helping customers with finding information that organizes their information needs. Actually, IR examines the securing, association, stockpiling, recovery, and circulation of data. Key phrases and important key-points are extracted from a set of questions, matching with which our mining model searches and stores for relevant questions. Then these questions are filtered and clustered according to their complexity and difficulty. Text analysis algorithms are used to calculate the lexical complexity, key words, difficulty and are classified using clustering algorithms with a set of sample questions to determine weight-age, difficulty etc.

In the illustration, Figure 1, the full scaled architecture of the proposed model is shown. The collected set of questions on a particular topic are processed using many textual analysis algorithms and techniques for extracting similar patterns, key words and phrases that define that set of questions of a particular topic. Based on these the mining model, searches for questions on that topic comparing and analysing these phrases, patterns make a relevant set of questions. The recovery module utilizes the report file to recover those archives that contain some question terms (such records are probably going to be applicable to the inquiry), register importance scores for them, and afterward rank the recovered reports as per the scores. The positioned reports are then introduced to the client. The archive assortment is likewise called the content information base, which is ordered by the indexer for proficient recovery.

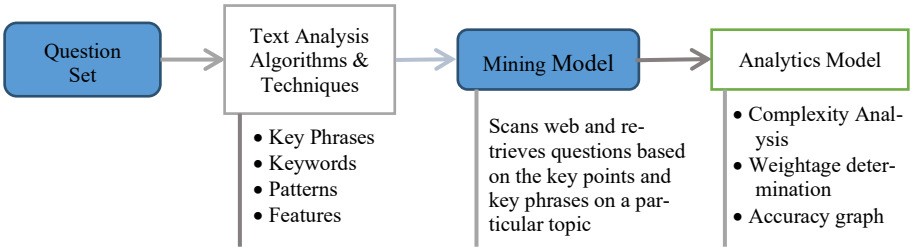

Fig. 1. Proposed mining and analytics model

Upon retrieval, the questions are fed to an analytics model for proper analysis. The complexity in terms of difficulty, lexical complexity, topic relevance etc. are all measured through this model and on passing through several pre-trained classification model it is classified rather grouped with the most relevant set of questions based on nearest neighbour algorithm. All these analysed units cater to the weight-age and parameter determination the questions individually.

If we can apply this collective model in the Unifying Logic Layer of Semantic Web Architecture, then we can set new query rules over information retrieval.

3.1 Key phrase extraction

Key phrases are unique set of words that help to define the essence and motive of the text. These key phrases help to summarize the text, in the form of list of relevant concepts portrayed in the article.

3.2 Key phrase extraction includes

Lemmatization of text. Bringing the root word out from the words as it makes it is unnecessary to use each and every word in the text, so they are brought down to their root word.

Selection of potential phrases. Texts contain many irrelevant words, stop words etc. These must be filtered out so that important words can be selected for keywords and key phrases formation and consecutive words bearing contextual similarity must be grouped together.

Scoring phrases. Once the list of possible phrases is extracted, their importance and influence is measured and scored.

RAKE, Rapid Automatic Keyword Extraction calculation, is a region self-sufficient expression extraction computation which endeavours to choose key articulations in an assortment of text by stalling the repeat of word appearance and its occasion with various words in the substance. Instead of returning the key phrases in order of score, it returns the score and the extracted key phrases. RAKE_NLTK uses natural language processing toolkit for few calculations.

Genism, carried out in Genism Python and Cython, Genism is an open-source library for regular language preparing, utilizing present day measurable AI. These algorithms are used over texts to extract key phrases with their influence scores.

These key-phrases will be categorized as per the relevance. The key phrases, categories are organized in an xml based data repository maintaining the ontological structure of the information pieces. During data recovery measure, the client takes care of a particular inquiry as for the expected space. A parse tree will be created from the given question by the NLP parser. The parse tree yields an articulation created by means of DFS (Depth First Search) crossing of semantic tree relating to the inquiry. The DFS expression represents POS (Parts of Speech) tagged phrases organized in a tree as per their dependencies with each other. The key phrase extraction module extracts key phrases by parsing the POS tagged expression in an order depicting the degree of relevance with the query. Subsequently, the removed key expressions are given to the key expression classification module to arrange each expression to a cloister classification. Each pair of key expression: classification is utilized by the proposed recovery framework to recover the website links requested concerning their importance to the question. The recovery module counsels the xml based web connect vault to coordinate with the key expression class pair with the put away key expressions. As mentioned earlier, each stored key phrase is also associated with relevant web links, which aids the retrieval module to list the relevant web links.

Fig. 2. Key-phrase extractor

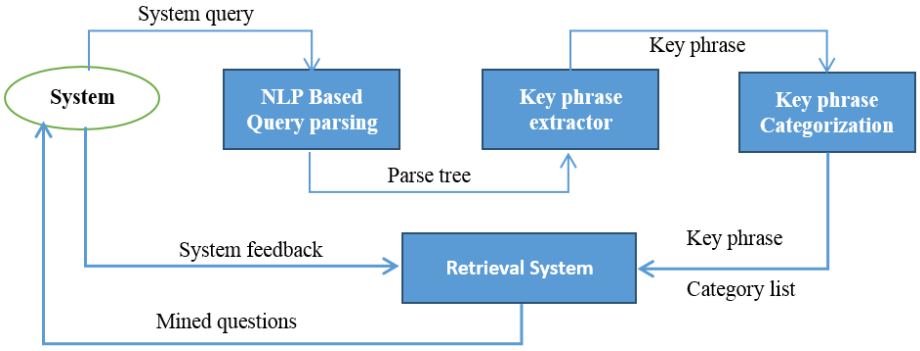

Fig. 3. Proposed mining model for question retrieval

3.3 Question analytics and weight-age determination

After creating the collection of the questions from the retrieval system, these questions were passed through NLTK algorithms for analyse these questions. The questions were processed to determine which category they belong to using clustering algorithms (e.g. KNN). Then through keyword extraction their subjective complexity is determined.

Lexical complexity is an indication of readability of the text. This readability can be affected by both the use of complex sentence formation, use of complex word combinations and also in the technical field, with the use of technical terms it can be affected. Many algorithms have tried to formulate the lexical complexity on various parameters for different use cases ranging from general case, technical or even education and healthcare.

Table 1. Readability complexity determination algorithms

Formula / Graph	Year created	Use Case
Dale Chall Readability Score	1995	General / Education
Coleman-Liau	1975	Education
Flesch-Kincaid	1975	General Use case
Smog Index	1969	Healthcare
ARI (Automated Readability Index)	1967	Technical
Linsear write Formula	1966	Technical
Gunning Fog	1952	Business and product related
Flesch Reading Ease	1948	General Use case

Automated Readability Index or ARI is very straight forward to calculate. It considers characters, words and sentences, leaving far from being obviously true measurements like "complex words". It is the most usable formulation in case of technical domain.

$$\text{ARI} = 4.71 * (\text{characters} / \text{words}) + 0.5 * (\text{words} / \text{sentences}) - 21.43 \quad (1)$$

The Dale-Chall readability score measures a text against a word knowledge of a random fourth-grader. According to its scale, the more unfamiliar words used, the higher the reading level will be. In our case we can actually tweak the original formula where instead of going for complex words we can go for technical keywords and can calculate the score using the formula.

$$DCRI = 0.1579 * (technical\ key - words/total\ words * 100) + 0.0496 * (total\ words/sentences) \quad (2)$$

Depending on the domain of implementation, we can use a combination of both these indexes with different influences to determine readability index.

Weightage of the question treated as text is determined as a combination of key phrase and keyword density, readability index and inverse of accuracy on attempt.

$$Accuracy\ index = correct\ attempt/total\ attempts \quad (3)$$

$$Weightage = (no\ of\ keywords/total\ words) * (1/(accuracy\ index)) * readability\ index \quad (4)$$

*Accuracy index is a dynamic quantity changing from time to time so weight-age is also a dynamic quantity.

4 Semantic web architecture overview

The principle worries of our own is the layered engineering of Semantic Web orchestrated in chain of command. Each layer is workable to the upper layer and goes about as customer to its hidden layer imitating an upward pyramid, roundabout and a transcending design. As this work process is grouping driven and plan of steps, it addresses a reflection of higher request - change in a solitary layer will be influencing the usefulness of different layers. The most eminent occasion is the International Standards association (ISO) and the Open Systems Interconnected (OSI) [19]. The most up to date semantic web engineering as planned by Tim Berners-Lee is clarified as:

Unicode layer and URI gives an improved on intends to distinguish assets like a site, a picture, a record, or an individual: fundamentally any unit with a character. Unicode, utilized for PC character portrayal, is a widespread standard encoding framework: it addresses all language dissimilar to other encoding frameworks. Followed by the XML Layer depicts the archive content. XML composition then again gives lexical help for confirmed XML archives. Then comes the RDF & Data interchange Layer: URI is utilized to target web assets and uses chart models to depict connections among different assets. The outline of RDF is an improved on displaying language that has

presented classes of assets, properties and their interrelationships [20, 21]. The following SPAQL - query, Ontology (OWL), RIF & RDF-S Layer: The foundation of Semantic Web, Ontology, gives semantic executable by the machine and for better man to application correspondence - a shareable space. It's primary target is to give semantic to produce web meaning, which will assist machines with unravelling the significance and help in data sharing [22]. The XML design is indicated by the RIF (Rule Interchange Format) for rules in similarity with RDF and OWL at expressive force at a halfway level as per the RIF Working Group [23]. Unifying Logic Layer passes on as a foundation layer for uniting the two recently referenced layers into a whole, to coordinate requests and attract rules over data tended to in the RDF close by ontologies related with it and schemata. Different works in this area have centred at joining rules with question implanted office, with a blend of rules and language structure. Then comes the Proof Layer is for validation of particular statements. The following Trust Layer depends upon the information source which can deny bothersome applications or customers induction to these sources. It's everything except a piece of trust and conviction between data sources and units of clients. The UI and Application Layer acts as an interface the users interacts with so it must satisfy them along with the applications. The Vertical Layers like Crypto is utilized for encryption purposes and for advanced mark. Existing in firth to 6th layer is utilized to set up trap of trust. XML marks, by applying it to the asset content, it very well may be distinguished. [24]

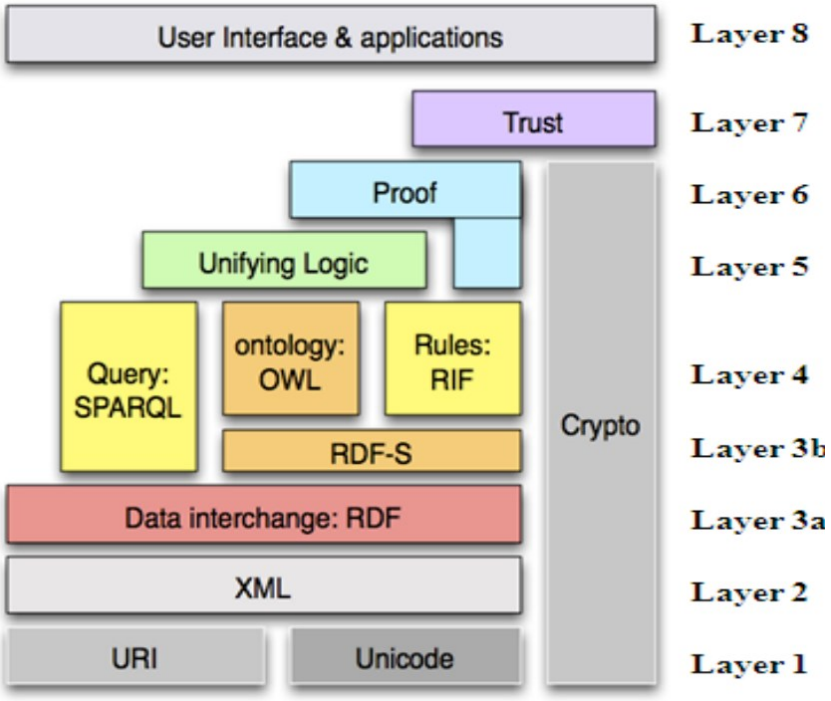

Fig. 1. Diagrammatic representation semantic-web block architecture

5 Experimental setup and flow

- **Step 1:** A set of collected questions are processed to extract the keywords and key phrases. Each question goes through textual pre-processing, like removal of stop words, lemmatization etc. Then keywords and key phrases are extracted from them using NLTK algorithms.
- **Step 2:** Using these keywords and key phrases as searching tokens along with topic name, the mining model searches through web pages for relevant questions. The class of question it belongs to is detected using KNN algorithm trained on the acquired set of questions. The topic relevance is measured using average cosine similarity to the questions belonging to the class.
- **Step 3**: These questions are then processed using NLTK algorithms to extract keywords and their key phrases. Then their lexical complexity (readability index) is calculated and their accuracy is set to 1. Then their initial weightage is calculated using the formula:

$$Weightage = (no\ of\ keywords/total\ words) * (1/(accuracy\ index)) * readability\ index \quad (5)$$

Readability index is calculated based on the sector of application; ARI algorithm for technical domain, DCRI for general/medical domain.

- **Step 4:** With time the accuracy indexes changes and thus, the weightage changes. This helps to monitor the acceptability, accuracy and dynamic drill set for examinee.

```
anish@BurpMaste-98:~/paper2021$ python3 sample.py
Driving Text - According to Java Operator precedence, which operator is considered to be with highest precedence?
Stop words -  {'don', 'to', 'my', 'been', 'such', 'these', 'other', 'more', 'so', 're', 'under', "you've", 'didn',
"she's", 'mightn', 'our', 'now', 'and', 'very', 'doesn', "haven't", 'not', 'out', "should've", 'had', 'a', 'no', "w
eren't", 't', "wasn't", 'won', 'haven', 'are', 'shan', 'ain', "that'll", 'off', 'this', 'be', 'for', 'aren', "needn
't", 'we', 'hadn', 'yours', 'who', "shouldn't", 'above', "it's", 'all', 'ourselves', 'whom', 'was', 'any', 'ours',
'isn', 'yourself', 'your', 'if', 'couldn', 'should', 'him', 'that', 'himself', 'then', 'into', 'between', 'how', 'd
', "isn't", 'their', 'where', 'themselves', 'll', 'from', 'over', 'only', 'through', "shan't", "couldn't", 'ma', "w
ouldn't", 'own', 'yourselves', 'against', 'me', 'his', "won't", 'during', 'of', 'once', 'itself', 'at', 'as', 'nor'
, 'can', 'just', 'is', 'both', 'o', 'do', 'up', 've', 'being', 'same', 'herself', 'while', 'there', "doesn't", 'fur
ther', 'hers', 'its', "didn't", "don't", 'having', 'needn', 'her', 'an', 'has', 'below', 'have', 'when', 'they', 'd
oes', 'it', 'those', 'why', 'doing', 'about', 'with', "mustn't", 'shouldn', 'most', 's', 'y', "hasn't", 'in', 'but'
, "mightn't", 'm', 'each', 'hasn', 'after', 'down', 'which', 'mustn', 'i', "you'll", 'what', 'or', 'were', 'until',
 'wouldn', 'again', 'few', "you'd", 'myself', 'she', 'did', 'some', 'here', 'theirs', 'by', 'weren', 'before', 'the
', 'will', "hadn't", 'you', 'am', 'too', "aren't", 'than', 'he', 'because', 'on', 'wasn', "you're", 'them'}
Word tokens-  ['According', 'to', 'Java', 'Operator', 'precedence', ',', 'which', 'operator', 'is', 'considered', '
to', 'be', 'with', 'highest', 'precedence', '?']
Filtered Sentence -  ['According', 'Java', 'Operator', 'precedence', ',', 'operator', 'considered', 'highest', 'pre
cedence', '?']
anish@BurpMaste-98:~/paper2021$
```

Fig. 2. Text pre-processing - Tokenization, stop words removal

```
anish@BurpMaste-98:~/paper2021$ python3 sample.py
driving text -  Which are the two subclasses under Exception class?
lemmatized text -  ['Which', 'two', 'subclass', 'Exception', 'class', '?']
anish@BurpMaste-98:~/paper2021$
```

Fig. 3. Lemmatization of texts to find root word

```
What is the difference between object oriented programming language and object based programming language?

['object oriented programming language',
 'object based programming language',
 'difference']
```

Fig. 4. RAKE-NLTK for key phrase extraction

```
Enter The Question: How to find duplicate number on Integer array in Java


Post Evaluation Weightage determined:  6



Data Structure and Algorith : 1
Programming Languages : 1
Code/Error Analysis : 0
```

Fig. 5. Class and weightage determination

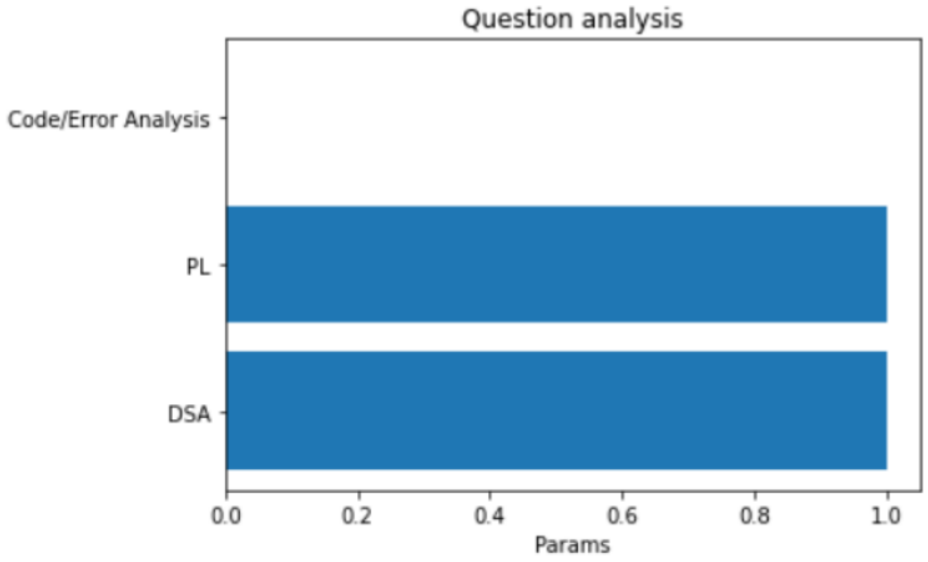

Fig. 6. Key phrase-based domain determination

6 Architecture & Influence of AI services with web based semantic concept

To plan a keen performing E-Learning framework to offer simple admittance to looked through courses, modules and so on Different significant things are to be created and observed and taken into concern like data and theory-based information which is

the focal point of the engineering. They are enormous for going most likely as a vault where ontologies metadata deriving rules, instructive assets, course materials and client profiles are dealt with. The metadata can be arranged in a similar report or can be protected outside in an intricate information unit. Benefits of outside stockpiling or remotely put away parts incorporate the simplicity of clearing the different metadata-depiction put away in the data set and along these lines helps in productive memory the executives. Additionally, as far as the re-convenience of similar materials, there are conflicting methodologies by clients and creators which sees the chance of various depictions as indicated by various ideas.

Yet, the main part will be the UI GUI with which the client will communicate. Clients will associate with it, explore through it and will get their ideal inquiries. This is the Access Interface. Semantic Search Engines furnish the API with various strategies for directing questions and information bases. The Interference Engine is a canny blend of real factors that helps with handling and answer requests and is at risk for the completion of new genuine elements through the brilliant blend of real factors as of now have in their knowledge base. Services incorporate the assortment of predefined services that are given to the clients like appearance subtleties, obviously, course map and so on in the services segment, effective robotized tweaked searching arrangements can be incorporated utilizing the execution of AI where the catchphrases extraction can make the searching outcomes more proficient. The NLP calculations pre-measure the searching watchwords to discover and plan results all the more effectively. RAKE calculation is utilized to remove the catchphrases with expanding significance past a limit post pre-processing (disposal of stop words, lemmatisation, tokenization and so on). The catchphrases are then planned with the themes and course units and from the chose bunch cosine work helps in the exemplary outcomes with the most proficiency.

7 Conclusion

In the proposed system we have tried to make an end-to-end question retrieval and analytics model. The model is trained on a set of questions which is analysed through NLTK algorithms for key-point and key-phrase extraction. These key-points and key-phrases are ordered in accordance to relevance and importance based on which the mining model retrieves the questions from the web pages. These questions are the categorized and grouped into classes using classification algorithms. Then these questions are passed through the analytics model for weight-age determination. This helps in a dynamic examination system where the questions are shifted or changed according to accuracy rate of the candidate.

8 Future scope

Question analytics system is currently only working on determining the weight-age based on the various complexity parameters, but in future more in-depth analytics can be conducted which can explore more qualities of the question, like categorize it as

subjective, numerical, philosophical etc. These will also help in auto evaluation of the examination.

9 Reference

[1] M. Riad. et.al. PSSE: An Architecture for a Personalized Semantic Search Engine. International Journal on Advances in Information Sciences and Service Sciences Volume 2, Number 1, March 2010, pp.102 – 112.

[2] Myungjin Lee. et.al. Semantic Association-Based Search and Visualization Method on the Semantic Web Portal. International Journal of Computer Networks & Communications (IJCNC), Vol.2, No.1, January 2010, pp.140 – 152

[3] Ulf Brefeld. et.al. Document Assignment in Multi-site Search Engines. WSDM'11, ACM 978-1-4503-0493-1/11/02, February 2011, pp.575 – 584. https://doi.org/10.1145/1935826.1935907

[4] Fabrizio Lamberti. A Relation-Based Page Rank Algorithm for Semantic Web Search Engines. IEEE Transactions on Knowledge and Data Engineering, Volume 21, Number 1, January 2009, pp. 123 – 136. https://doi.org/10.1109/tkde.2008.113

[5] Nguyen Tuan Dang, and Do Thi Thanh Tuyen. Natural Language Question Answering Model Applied To Document Retrieval System. World Academy of Science, Engineering and Technology 51 2009, pp.36 – 39. https://doi.org/10.1109/iccsit.2009.5234481

[6] A.Jebaraj Ratnakumar. An Implementation of Web Personalization Using Web Mining Techniques. Journal of Theoretical and Applied Information Technology, 2005, pp. 68 – 73

[7] Caragea, Cornelia & Bulgarov, Florin & Godea, Andreea & Gollapalli, Sujatha. (2014). Citation-Enhanced Keyphrase Extraction from Research Papers: A Supervised Approach. 1435-1446. 10.3115/v1/D14-1150. https://doi.org/10.3115/v1/d14-1150

[8] Witten, Ian & Paynter, Gordon & Frank, Eibe & Gutwin, Carl & Nevill-Manning, Craig. (1999). KEA: Practical Automatic Keyphrase Extraction. ACM DL. 254-255. 10.1145/313238.313437. https://doi.org/10.1145/313238.313437

[9] Turney, P.D. Learning Algorithms for Keyphrase Extraction. Information Retrieval 2, 303–336 (2000). https://doi.org/10.1023/A:1009976227802

[10] Bharti, Drsantosh & Babu, Korra. (2017). Automatic Keyword Extraction for Text Summarization: A Survey. http://arxiv.org/abs/1704.03242

[11] Onan, Aytuğ & Korukoğlu, Serdar & Bulut, Hasan. (2016). Ensemble of keyword extraction methods and classifiers in text classification. Expert Systems with Applications. 57. https://doi.org/10.1016/j.eswa.2016.03.045

[12] Matsuo, Yutaka & Ishizuka, Mitsuru. (2003). Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information. International Journal on Artificial Intelligence Tools. 13. https://doi.org/10.1142/s0218213004001466

[13] Yang F., Zhu YS., Ma YJ. (2016) WS-Rank: Bringing Sentences into Graph for Keyword Extraction. In: Li F., Shim K., Zheng K., Liu G. (eds) Web Technologies and Applications. APWeb 2016. Lecture Notes in Computer Science, vol 9932. Springer, Cham. https://doi.org/10.1007/978-3-319-45817-5_49

[1] Liu, Fei & Liu, Feifan & Liu, Yang. (2011). A Supervised Framework for Keyword Extraction From Meeting Transcripts. Audio, Speech, and Language Processing, IEEE Transactions on. 19. 538 - 548. https://doi.org/10.1109/TASL.2010.2052119

[14] Liu, F. Liu and Y. Liu, "A Supervised Framework for Keyword Extraction From Meeting Transcripts," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 3, pp. 538-548, March 2011. https://doi.org/10.1109/TASL.2010.2052119

[15] Lee-Feng Chien. 1997. PAT-tree-based keyword extraction for Chinese information retrieval. SIGIR Forum 31, SI (July 1997), 50–58. https://doi.org/10.1145/278459.258534
[16] Fei Liu, Feifan Liu and Yang Liu, "Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion," 2008 IEEE Spoken Language Technology Workshop, Goa, 2008, pp. 181-184. https://doi.org/10.1109/slt.2008.4777870
[17] Bharti, Drsantosh. (2017). Automatic Keyword Extraction for Text Summarization in Multi-document e-Newspapers Articles. European Journal of Advances in Engineering and Technology. 4. 410-427.
[18] Garlan, D. et. al, "An Introduction to Software Architecture. In: Advances in Software Engineering and Knowledge Engineering", World Scientific Publishing Company, NY, 1993.
[19] Buraga, S. and G. Ciobanu., 2002. A RDF- based model for expressing spatio-temporal relation between web sites. In The 3rd International Conference on Information Systems Engineering. IEEE Computer Society. pp: 355. IEEE Computer Society Washington, DC, USA. https://doi.org/10.1109/wise.2002.1181671
[20] Matthews, B, Semantic Web Technologies. JISC Technology and Standards Watch, Joint Information Systems Committee, 2005.
[21] Boley, H. et. al., "RIF Basic Logic Dialect. W3C Working Draft", 2008.
[22] Mark Bartel et al. "XML Signature Syntax and Processing (Second Edition)", 2008.
[23] Russell Cloran et. al, "XML Digital Signature and RDF", 2005.

10 Authors

Mr. Subhabrata Sengupta, obtained his Bachelor's in Technology degree in the field of Information Technology from Bengal Institute of Technology, India. Then he obtained his Master's degree in the same field from Jadavpur University, India, and is currently pursuing his PhD in Computer Science majoring in Machine Learning and Information Retrieval from University of Engineering and Management, Kolkata, India. Currently, he is a professor at the Faculty of Information Technology, Institute of Engineering & Management, Kolkata, India. His specializations are in Machine Learning, Information Retrieval Systems, Smart and adaptive education systems. His current research interests are smart information retrieval systems, student and academic performance analysis, educational data mining and predictive modelling for smart and adaptive education systems with the applications of Machine Learning.

Mr. Anish Banerjee, is pursuing his Bacherols in Technology (2017-2021) in Information Technology from Institute of Engineering and Management, Kolkata, India. He worked as a machine Learning Engineer and a Team Lead at Analysed, Hyderabad and as a Machine Learning Developer and consultant at Durbin Technologies, Kolkata, India and is currently placed in Tata Consultancy Services' Digital unit. His research interests are in Machine Learning, Deep Learning, Data Analysis and Statistical Modelling of important data. Has has international experience of working in these said fields for companies like MACO, Quinch and FedEX.

Dr. Satyajit Chakrabarti, obtained his Bachelor's in Technology degree in the field of Computer Science Engineering from Institute of Engineering and Technology, Kolkata, India (1998-2002). Then he obtained his Master's in Computer Science from The University of British Columbia, Vancouver (2002-2004) and his PhD in Electronics

and Communication Engineering researching in Nanotechnology from National Institute of Technology, Agartala, India (2012-2016). Currently, he is a professor at the Faculty of Computer Science, Institute of Engineering & Management, Kolkata, India. He was the Teaching and Research Assistant in University of British Columbia, worked as a Developer Analyst and then as a Project Manager in TELUS. He has served as the President of US Chamber of Industry & Innovation, American Innovation Venture Capital Corporation, American STP Corporation and as the CEO of SAT VENTURES LIMITED. He has always put forward his enthusiasm and support towards productive and innovative ideas and has given his utmost help and co-operation in constructive approaches in the domain of science, research, innovation that builds a better tomorrow.

Article submitted 2021-08-01. Resubmitted 2021-10-03. Final acceptance 2021-10-15. Final version published as submitted by the authors.