# Comprehensive Survey: Automatic Query Expansion

Sayani Ghosal[1], Dr. Devendra Kumar Tayal[2]

[1]M.Tech-Scholar, Computer Science Department&Indira Gandhi Delhi Technical University for Women, India

[2]Professor, Computer Science Department&Indira Gandhi Delhi Technical University for Women, India

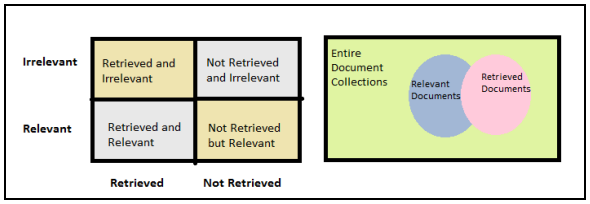
***Abstract*** *- Intoday's world massive amount of data in web is increasing exponentially. Internet users extract relevant information through few key words from loads of unstructured data; due to that reason query expansion comes into play. Automatic Query Expansion is a well-liked method and major research area which can be used to enhancethe performance of information retrieval. This paper aspires to unveil fundamentals of information retrieval, various techniques of Query Expansion, their methods and challenges. Its further explains Automated Query Expansion in detail with its inherent advantages of handling large unstructured data. This paper aims to study various researches conducted in the endeavour including comparative study of different techniques and their advantages.*

**Keywords** — *Information Retrieval (IR), Query Expansion, Automatic Query Expansion (AQE)*

## I. INTRODUCTION

Information retrieval (IR) is a procedure that generally deals with extraction of structured and unstructured data. Generally structure data means large records from database which have same syntax and Boolean expression whereas unstructured data means textual documents in response to a query. Example of structured data- All rows from a relational database table consist with same column. While for unstructured documents search engine find data with a semantic relation, example is salary of employees. Information Retrieval mainly works with unstructured data. IR is keen to find the document that pleases the given information requirements [2].

Precision and Recall are two main evaluation measures of IR. Precision term can be used for percentage of retrieved documents based on the users query and recall can be used for percentage of pertaining documents based on users query. Both terms are very much useful for evaluation measures. Measuring precision is simple where set of users can agree on relevance of each of the retrieved documents. Recall depends on the number of relevant documents for total collection. Measuring recall is more difficult. Feasibility of recall is not good for large dataset.




$$recall = \frac{|\{relevant\ document\} \cap \{retrieved\ document\}|}{|\{relevant\ document\}|}$$

$$precision = \frac{|\{relevant\ document\} \cap \{retrieved\ document\}|}{|\{retrieved\ document\}|}$$

**Fig 1: Precision and Recall**

Query Expansion is the method to increase the recall where most common form of Query Expansion is Global Query Expansion which defines by the global methods for query reformulation. Relevant results extraction from the expanded query process is called Query expansion (QE). Vocabulary problem is one of the major issues related to the Information Retrieval systems [2].

Automatic query expansion is a popular technique and major research area which can be used for improvement of information retrieval performance. Searching any valuable information from online web can be a tedious job which can give irrelevant documents with valuable documents. Volume of data in any popular web is very larger and poorly organized; due to that reason finding any useful information is very difficult job. Web user generally provides very little information to search any relevant information from the web. When users generally search with multiple topics then user search will get the result with good matches. The main purpose for AQE technique is to find the similarity and relevance of users query.

## II. INFORMATION RETRIEVAL

Information Retrieval is the method which can extract the relevant unstructured documents from a dataset. Unstructured datameans free form of natural language text. Example of unstructured data is image, video, audio but query expansion research is mainly deal with ambiguous natural language. Information Retrieval basic three mechanisms are 1) document indexing 2) searching and 3) Ranking [1].

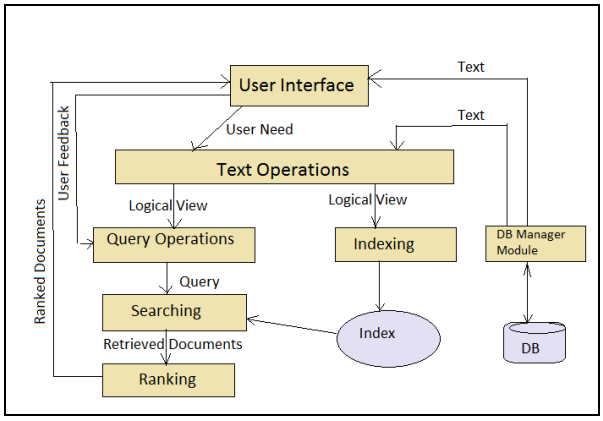



**Fig 2: Information Retrieval Method[2]**

### A. *Information Retrieval Method*

Information Retrieval performs basic steps to complete the process where retrieve content converted into logical content. Steps for Information Retrieval are described below [2].

1) User interface is demarcated by user requirements and it will be in the form of textual query.
2) Textual operation can convert the textual query and same happened with content indexing.
3) Pre-processed query is converted into system level demonstration by query operation
4) The query is implemented on top of the documents which can able to retrieve the set of appropriate documents. Index structure is a fast query processing method which is a document form.
5) Retrieved documents are ranked according to the user requirements.
6) User will be benefited by using ranked documents which is subset of main document.

### B. *Query Expansion Definition [5]*

Query expansion recalculate user's actual query to improve the information retrieval efficiency.

Let a User query consist of n terms,

$$Q = [\, t_1, t_2, t_3, \dots \dots . t_i, t_{i+1}, \dots t_n \,]$$

Adding together of new terms

$T' = t'^1, t'^2, t'^3, \dots \dots t'^m$From the data source D and Removal of stop words

$$T'' = [t_{i+1}, t_{i+2}, \dots t_n \,]$$

The reformulated query can be expressed as:

$$Q_{exp} = \left(Q - T''\right) \cup T'$$
$$= [\, t_1, t_2, t_3, \dots \dots . t_i, t'^1, t'^2, t'^3, \dots \dots . t'^m \,]$$

In the above definition, the key aspect of Query Expansion is:

$T'$Set of significant terms which added with original user's query for retrieving applicable documents computationcan occursusing set$T''$and data sources D [2, 5].

### C. *Query Expansion Techniques*

***1) Manual Query Expansion:***In this case user can manually reformulate the query

***2) Automatic Query Expansion:***System can automatically reformulate the query without any user interference. Both the technique is work out set and choice of data sources is included into system's intelligence.

***3) Interactive Query Expansion:*** Query reformulation can be done as a result of joint cooperation between the system and user. Initially system returns search results for an automatically reformulated query and the users chooses significant results between them. Based on user's preference, the system further reformulates query and retrieves the results. The procedure continues until the user is fulfilled with the actual search results [4].

***4) Log-based:***The method for query expansion is based on query logs. The main idea of this method is to build the correlations between user query terms and main document terms through taking out the query logs.

**5) *Co-occurrence:***Different approach is used in co-occurrence data to improve indexing and query building. The term co-occurrence data can be used in document retrieval systems for detection of indexing terms. [16]

## III. Automatic Query Expansion (AQE)

AQE is an improved method which is beneficial for Natural Language processing Research. Below point is described for understanding of basic level knowledge of Automatic Query Expansion.

### A. *AQE Phases*

According to the survey of different research paper the automatic query expansion technique consists of four steps [3].

***1) Data Processing-***

Data processing is the main important tasks for any Natural language processing systems. Data processing started with the extraction of data and then construct the data structure for easily accessible by the users. Transforming the raw data source is used for expanding the user query. Pre-processing of a data source is autonomous for user query but the

processing is specific for data source. In pre-processing expansion method is also considered. In many query expansion techniques, the top ranked items are retrieved to the response of user query. Original user query is generally retrieved from the collection of several documents. For pre-processing stage it is important to index the collection of documents and need to run the user query based on the collection index. [3].

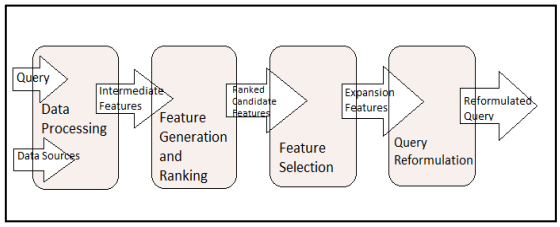



**Fig 3: Automatic Query Expansion Phases [3]**

#### *2) Features Generation and its Ranking*

In features generation and ranking stage the system will generate and ranks the expansion features. The most important query expansion method is to choose the minimal proportion of candidate expansion features. Candidate expansion features then added to the user query [28]. The input is actual user query and converts the data source for this stage. The output for this stage is expansion features with associated scores. The user query is pre-processed to remove the common words and important expanded terms. According to the execution of candidate generation techniques is also classified. Ranking estimation is measured by relationship between the expansion features generated and the actual user query [3].

#### *3) Features Selection*

In Features selection stage the top elements are selected for Automatic query expansion. The top element selection done by individual basis and it will not consider the mutual dependencies between the expansion features. Evaluation measure of many experimental results for the automatic query expansion technique has done based on individual assumption [25]. Generally limited number of features is selected for query expansion. Resulting query processed more quickly through query expansion technique because of retrieval effectiveness. Small set of quality terms is not essentially less successful than adding up all candidate expansion terms. Most experimental research shows that quantity of expansion features is of low. Features score can be interpreted based on probabilities and selection can be done based on greater probability term compare to threshold value [3].

#### *4) Query Reformulation*

The last stage for automatic query processing is query reformulation, where system needs to describe the expanded query which will be submitted to the IR system. Generally system needs to assign a weight to every feature and then describes the expanded user query. Many popular queries reweighting technique is used in Automatic Query expansion. Expansion terms are extracted from pseudo-relevant documents forcomputing their score for AQE query reformulation.

### *B. AQE Data Source*

AQE data source is described below [5]:

#### *1) Hand-built Knowledge resource*

For hand-built knowledge resources the knowledge extraction from hand-built data sources like dictionaries, Ontologies, thesaurus and LOD cloud. Thesaurus-based query expansion can be defined as automatic or hand-built. One of the well-known hand-built thesauruses is WordNet [21]. WordNet is increasing the actual query with semantically similar terms. It was observed that retrieval efficiency was enhanced expressively for unstructured query, while only minimal improvement was visible for structured queries. As per the various researches, some other articles also use the WordNet to increase the actual query which use synonyms of the query and assigns weight. ConceptNet is also useful for mutual sense knowledge of data and it defined as a relational semantic network [6].

#### *2) Documents for retrieval process*

Additional term for initial query started playing a very important role in query expansion. The two types of clustering are used in documents retrieval system- 1) clustering of terms and 2) clustering of documents. As per the different researchers Corpus based technique is used similarity thesaurus toincrease the actual query. Similarity thesaurus is defined as the collection of documents which mainly based on specific domain knowledge.Every term for thesaurus is defined as weighted document vector. The disadvantage of corpus specific query expansion technique is to establish relationship between words which are connected with other communities [5, 6].

#### *3) External text collection and resources*

External Text collection is used to extract the most shared and operative data sources for query expansion. Query expansion approaches show better performance in comparison to all three data sources. Somedata sources are under External Text collection category which requiredvarious procedures for text

collection. Generally extract data from anchor tag. Stop word removal and word stemming is part of data extraction. Query logs are also data source for query expansion, where the query using association between query terms and data. Most of the search engines in research papers are using Query Expansion which based on the query logs [6].

### C. AQE Methods

Automatic Query Expansion methods mainly divided into five techniques [3]

#### 1) Linguistic Approach

Linguistic analysis focuses on morphological, lexical, syntactic and semantic relationships between words in the query. Linguistic analysis is based on different databases like dictionaries, thesauri or WordNet.

#### 2) Corpus-Specific Approach

Corpus specific technique utilizes for large structured set of texts (corpus). For this technique it analyses the full database content for finding the features that are similar. It will used to find correlations between terms at the document level or paragraphs or sentences.

#### 3) Query-Specific Approach

Query specific technique mainly utilizes the local context which can be provided by the user's query. It can use the top ranked documents which retrieved by the user's query to generate the features of candidate. The query specific technique can be more useful than corpus specific techniques which can able to retrieve the candidate features that appear frequently in document collections that are irrelevant to the user's query

#### 4) Search Log Specific Approach

Search log analysis uses query logs for query associations. Search log generally contain the past user queries and URLs for the pages. The technique may convert implied relevance feedback other thanactual retrieval feedback data query. Search log technique can able to extract the candidate features from past and current user queries which also related to the current query.

#### 5) Web Data based Approach

Web data technique is used for anchor texts on web pages to generate the candidate features. Anchor text is visible, clickable text of a hyperlink. Anchor texts are similar for real user queries. Anchor texts can be describe by the contents of the linked documents. This technique can be utilized for the complex net of Wikipedia documents and hyperlinks

**TABLE I**

**AQE METHODS [3]**

| Techniques | Sub Techniques | Advantages |
|---|---|---|
| Linguistic Approach | 1)Stemming<br>2)Ontology Browsing<br>3)Syntactic Parsing | 1) Influence by the global language property |
| Corpus-Specific Approach | 1) Concept term<br>2)Term clustering | 1) Analyse the inside of data<br>2)Subject point handle in a better way |
| Query Specific Approach | 1)Distribution difference<br>2)Model based<br>3)Document Summarization | 1) Easy to specify<br>2) query can provide the local context<br>3) Recurrent for assortment but unsuitable for query |
| Search log Specific Approach | 1)Relative Query<br>2) Exploiting query Relationship | 1) System can encode implied relevance Feedback |
| Web Data Based Approach | 1)Anchor Text<br>2) Wikipedia based | 1) Performed well in long text<br>2) Represented by Anchor Text |

### D. AQE Ranking Methods

The comprehensive ranking methodology of AQE is described below. [3, 8]

***1) Text segmentation:*** In text segmentation system can recognized the single term which mainly occurred in textual collection but it canignore the punctuation, case etc.

***2) Word stemming:*** The system used very large morphological lexicon. It can contain Verbs in many forms like past participle, past tense, future infinitive, third person singular etc.It also contain Nouns and adjectives where Nouns in the form of singular and plural genitive singular, plural and adjectives in comparative, base and superlative form.

***3) Stop word*:**Stop word is used for stop list. As an example of stop word contained CACM dataset and common function words can be deleted. System removed the terms which seemed in less than 3 and more than 100,000.

***4) Document weighting-***System assign weights to the query terms for each document. **TF-IDF** scheme is used for document weighting. (**TF**: Term Frequency in the query and **IDF**: Inverse Document Frequency) [26]

***5) Weighting of unexpanded query-*** System also used the function log TF-IDF for weighting term in the unexpanded query. (**TF**: Term Frequency in the query and **IDF**: Inverse Document Frequency)

***6) Document ranking with unexpanded query-***System computes document ranking between the document vectors and the unexpanded query vectors. It also includes cosine normalization [3].

***7) Expansion term ranking-*** For Expansion term ranking the terms is used with higher estimated probability for the first retrieved documents compare to the entire collection.

***8) Weighting of Expanded Query-***for Weighting of Expanded query systems used vector space notation and normalized score. The normalization can be performed by dividing every score by the maximum score. Smoothing function is used for large fraction with very low scores. [3].

***9) Document ranking with expanded query-*** Document ranking also computed by using inner product between the expanded query vector and document vectors [3].

### *E. Application Of AQE*

Automatic Query Expansion are used in many research areas [5]

***1) Information Filtering:*** Process for monitoring the stream of documents and selecting the relevant documents.

***2) Multimedia Information Retrieval:***It performs the text based search over the media metadata (html/xml description).

***3) Question Answering:*** Main aim is to provide certain type of natural language questions instead of full documents.

***4) Cross-Language Information Retrieval:*** In this cross language retrieval the user query and retrieved can be in different language.

***5) Text categorization:***searching for hidden web content that will not be indexed by standard search engine.

**TABLE III**

**AQE APPLICATIONS [5]**

| **Area** | **Analysis of applications in recent Research** | | |
|---|---|---|---|
| | **Application** | **Data Source** | **Related Publication** |
| Information Filtering | 1) e-mail<br>2) e- commerce<br>3) Multimedia system<br>4) Search Engine | 1)Anchor Text<br>2)Twitter<br>3)Wikipedia | T Chantzios, 2019 [16]<br>M Chahal, 2016 [25] |
| Multimedia Information Retrieval | 1) Semantic information Search<br>2) Audio, image and Video Retrieve | 1) Anchor Text<br>2) Annotation<br>3) Captions<br>4) Query log | V Rocha, 2016 [17]<br>PSchäuble, 2012 [9]<br>D Patil, 2015 [22] |
| Question Answering | 1) User Query answer<br>2) Response quickly | 1) Social Network<br>2) WordNet<br>3) CQA Archive | Y Chen, 2015 [18]<br>P Garg, 2013 [24] |
| Cross-Language Information Retrieval | 1) Different language communication through user | 1) User log<br>2) Co-occurrence term<br>3) Word Embedding | V K Sharma, 2016 [19] |
| Text categorization | 1) Text Classification<br>2) IOT<br>3) Biomedical | 1) top tweets<br>2) CLEF and TREC | L Trieu, 2017 [20]<br>A Saritha, 2014 [23]<br>J Ababneh, 2014 [26] |

**TABLE III**

**RECENT RESEARCH FOR AUTOMATIC QUERY EXPANSION**

| Publication | Automatic Query Expansion Survey | | | |
|---|---|---|---|---|
| | **Datasets** | **Approach** | **Techniques used** | **Evaluation Parameter** |
| DK Sharma, 2019[10] | 1) CISI<br>2) CACM<br>3) TREC - 3 | Hybrid Evaluation Algorithm | 1) Cuckoo Search<br>2) Accelerated particle swarm optimization<br>3) Fuzzy Logic | 1)average recall<br>2)average precision<br>3)Mean-Average Precision<br>4) F-measure |
| DK Sharma, 2019 [11] | 1) CACM<br>2) CISI | Soft Computing | 1) particle swarm optimization<br>2) Fuzzy Logic | 1)Mean-Average Precision<br>2) F-measure |
| Ahlem Bouziri, 2018[12] | 1) TREC-Robust<br>2) CLEF | Term Correlation | 1) Association Rules<br>2) Pairwise learning to Rank<br>3)AR Ranking | 1) MAP |
| Yogesh Gupta , 2017[13] | 1)CACM<br>2) CISI<br>3) TREC-3 | Term Weighting | 1) particle swarm optimization<br>2) Filtering Method<br>3) Fuzzy Logic | 1) recall<br>2) precision<br>3)Mean-Average Precision<br>4) F-measure |
| Dwaipayan Roy, 2016[14] | 1) TREC ad hoc<br>2)Web data | Word Embedding | 1) k- nearest neighbour<br>2) word2vec | 1) MAP |
| Jagendra Singh, 2015 [15] | 1) TREC-3<br>2) FIRE ad hoc | Pseudo Relevance feedback | 1) Term Co-occurrence<br>2) Semantic information of Terms | 1)Precision<br>2)Recall<br>3) MAP |

### *F. Advantages of AQE*

Please find the below advantages for AQE techniques [1, 3]

#### *1) Increase the recall*

More chance for a relevant document that does not contain the actual query terms which need to be retrieved.

#### *2) Strict recall improvement*

When query terms are added (AND) composed with any Web search engines then strict recall will be improved.Expanded user query can submitted by using the Boolean operators (AND, OR).

#### *3) Combined recall/precision measure*

Information Retrieval systems measured by both It will be giving better retrieval effectiveness for AQE results.

### *G. Disadvantages of AQE*

Please find the below limitations for AQE techniques [3]

#### 1) *Parameter Settings:*

AQE techniques depends on several parameter-

a) number of top-ranked documents which select from the actual query
b) number of expansion features
c) different variables within term-ranking
d) weighting functions

Parameters rely on the quality of the actual query and the size of data source. Fixed values can be used for the same but it may not work well for all types of query.

#### *2) Efficiency:*

Good Automatic Query expansion technique is computationally costly due to the execution of the query. Significant performance time cut occurs due to extreme decrease in the quality results. AQE will deliver real time results for large number of users which will require the balance the performance time with quality results.

#### 3) *Usability:*

The AQE technique of Information Retrieval system is usually hidden from users. The AQE technique can add synonyms with the original query

and user will receive thehigh rankeddocuments that does not contain the original query term. In anchor text, the irrelevant documents can include the original query term. To solve this above problem user need to follow the IQE technique like, user need to expand the query and user needs to revise the query.

## IV. CONCLUSIONS

Automated Query Expansion performs best in specialized domain. Recall efficiency observed in areas of multimedia, cross language search etc. Recent research reports suggest that combined recall and precision efficiency is higher in AQE. Nowadays a great number of techniques are obtainable (Linguistic, query specific, corpus specific, search log and web data) that accomplish the different necessities such as computational efficiency, query type, external data accessibility and uniqueness of ranking system. Plenty of researches conducted based on Automatic Query Expansion have been done on standard dataset. However Automatic Query Expansion is computationally costly and time consuming it will required stability between performance time and quality of result.

## REFERENCES


[1] Singh, J., Sharan, A., & Siddiqi, S. (2013). A literature survey on automatic query expansion for effective retrieval task. *International Journal of Advanced Computer Research*, *3*(12), 170.

[2] Greengrass, E. (2000). Information retrieval: A survey

[3] Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)*, *44*(1), 1..

[4] Chum, O., Philbin, J., Sivic, J., Isard, M., & Zisserman, A. (2007, October). Total recall: Automatic query expansion with a generative feature model for object retrieval. In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on (pp. 1-8). IEEE.

[5] Azad, H. K., & Deepak, A. (2019). Query expansion techniques for information retrieval: a survey. Information Processing & Management, 56(5), 1698-1735.

[6] Buey, M. G., Garrido, Á. L., & Ilarri, S. (2014, September). An approach for automatic query expansion based on NLP and semantics. In International Conference on Database and Expert Systems Applications (pp. 349-356). Springer, Cham

[7] Kathuria, N., Mittal, K., & Chhabra, A. (2017). A Comprehensive Survey on Query Expansion Techniques, their Issues and Challenges. International Journal of Computer Applications, 168(12)

[8] *Carpineto, C., Romano, G., & De Mori, R. (1999). Informative term selection for automatic query expansion. NIST SPECIAL PUBLICATION SP, 363-370.*

[9] Schäuble, P. (2012). Multimedia information retrieval: content-based information retrieval from large text and audio databases (Vol. 397). Springer Science & Business Media.

[10] Sharma, D. K., Pamula, R., & Chauhan, D. S. (2019). A hybrid evolutionary algorithm based automatic query expansion for enhancing document retrieval system. *Journal of Ambient Intelligence and Humanized Computing*, 1-20.

[11] Sharma, D. K., Pamula, R., & Chauhan, D. S. (2019, February). Soft Computing Techniques Based Automatic Query Expansion Approach for Improving Document Retrieval. In *2019 Amity International Conference on Artificial Intelligence (AICAI)* (pp. 972-976). IEEE.

[12] *Bouziri, A., Latiri, C., & Gaussier, E. (2017, April). Efficient Association Rules Selecting for Automatic Query Expansion. In International Conference on Computational Linguistics and Intelligent Text Processing (pp. 563-574). Springer, Cham.*

[13] Gupta, Y., & Saini, A. (2017). A novel Fuzzy-PSO term weighting automatic query expansion approach using combined semantic filtering. *Knowledge-Based Systems*, *136*, 97-120.

[14] *Roy, D., Paul, D., Mitra, M., & Garain, U. (2016). Using word embeddings for automatic query expansion. arXiv preprint arXiv:1606.07608.*

[15] Singh, J., & Sharan, A. (2015, February). Co-occurrence and semantic similarity based hybrid approach for improving automatic query expansion in information retrieval. In *International Conference on Distributed Computing and Internet Technology* (pp. 415-418). Springer, Cham.

[16] Chantzios, T., Zervakis, L., Skiadopoulos, S., & Tryfonopoulos, C. (2019, June). Ping-A customizable, open-source information filtering system for textual data. In *Proceedings of the 13th ACM International Conference on Distributed and Event-based Systems* (pp. 228-231). ACM.

[17] Rocha, V., Kon, F., Cobe, R., & Wassermann, R. (2016). A hybrid cloud-P2P architecture for multimedia information retrieval on VoD services. *Computing*, *98*(1-2), 73-92.

[18] Chen, Y., Dong, B., Shen, Y., Zhenglong, W. E. I., & Liu, X. (2015). *U.S. Patent No. 9,213,771*. Washington, DC: U.S. Patent and Trademark Office.

[19] Sharma, V. K., & Mittal, N. (2016). Cross lingual information retrieval (CLIR): review of tools, challenges and translation approaches. In *Information systems design and intelligent applications* (pp. 699-708). Springer, New Delhi.

[20] Trieu, L. Q., Tran, H. Q., & Tran, M. T. (2017, December). News classification from social media using twitter-based doc2vec model and automatic query expansion. In *Proceedings of the Eighth International Symposium on Information and Communication Technology* (pp. 460-467). ACM.

[21] Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

[22] Patil, D., & Potey, M. A. (2015). Survey of Content Based Lecture Video Retrieval. *International Journal of Computer Trends and Technology (IJCTT)*, *19*(1).

[23] Saritha, A., & NaveenKumar, N. Effective Classification of Text. *International Journal of Computer Trends 0*, *20*, 40-60.

[24] Garg, P., & Bedi, E. C. S. (2013). Automatic question generation system from Punjabi text using hybrid approach. *International Journal of Computer Trends and Technology (IJCTT)*, *21*(3), 130-133.

[25] Chahal, M. (2016). Information retrieval using Jaccard similarity coefficient. *Int. J. Comput. Trends Technol*, *36*, 140-143.

[26] Ababneh, J., Almomani, O., Hadi, W., El-Omari, N. K. T., & Al-Ibrahim, A. (2014). Vector space models to classify Arabic text. *International Journal of Computer Trends and Technology (IJCTT)*, *7*(4), 219-223.