ISSN: 0974-8571 Vol.13 No. 2 December, 2021

International Journal of Computational Intelligence in Control

Smart Sindhi Documents Retrieval System Based on Pattern Discovery Approach for Students Search Services

Mashooque Ali Mahar[1*], Javed Ahmed Mahar[1], Mir Sajjad Hussain Talpur[2], Mumtaz Ali Mahar[1], Naveed Khan[2]

[1]Department of Computer Science, Shah Abdul Latif University, Khairpur, 66020, Sindh, Pakistan.
[2]Information Technology Centre, Sindh Agriculture University, TandoJam, 70050, Sindh, Pakistan.
Corresponding author: Mashooque Ali Mahar (mashooq.mahar@salu.edu.pk).

ABSTRACT: Information retrieval (IR) can be defined as performing actions on database storing, searching, and retrieving based on a user's requirements. The increase in globalization makes it important to find information in local languages such as Sindhi. Users' requirements regarding services provided in local languages are nowadays common. Hence, a lot of work should be done in the Sindhi language. Due to rich morphological structures, the retrieval process of Sindhi documents is not as simple as in English. A pattern discovery technique was used to develop morphological patterns. This paper presents a state-of-the-art-based approach for the retrieval of Sindhi documents. To develop the Sindhi IR system, we used intelligent techniques for query formulation, document identification and indexing, pattern matching, and document ranking. The soft copies of 31 books were collected from various sources and prepared into a database of 7283 documents for experimentation. A set of 50 queries was inputted to the developed IR system by the selected students using three created scenarios, i.e., single word, double word, and sentence. The execution time, along with elapsed time as well for each query and retrieved documents, was calculated. The performance of each scenario was evaluated individually in terms of precision and recall. It is evident after the analysis of the received results that the developed Sindhi IR system based on our proposed mechanism performs better.

INDEX TERMS: Documents indexing, documents ranking, information retrieval, morphological patterns, pattern discovery, Sindhi language.

1. INTRODUCTION

Natural Language Processing (NLP) is an important research area and a lot of research has been done in this field so far. The purpose of NLP is to deal with the development of computer systems that are able to perform meaningful tasks pertaining to natural languages [1]. The scope of NLP is increasing day by day as its importance in software applications based on NLP models and algorithms is felt. The IR is a branch of NLP known as the science of searching, and it is specifically used in document searching. It is also termed as the satisfaction of the user's request for IR. In this system, a user puts a query for the retrieval of a particular block of information. The query is then converted into the system understandable form, which further processes and searches for the required information from the database of documentation [2]. IR generally processes a user's query and searches for the requested set of documents or knowledge. Sindhi is known as one of the oldest Indo-Aryan languages [3]. South Asian countries like India and Pakistan are the ones where the Perso-Arabic script-based Sindhi language is mostly used. The Devanagari script is used for writing the Sindhi language, particularly in some areas of India. While the people of Pakistan, particularly those in Sindh province, use Arabic script with the assistance of the Persian language to accommodate implosive, retroflex, and nasal sounds [4]. IR as a research field is currently at its pinnacle [5].'s requirements regarding services provided in local languages like Sindhi are generally accepted. Based on the morphological structure of languages, IR provides many good results while dealing with other languages, but when it comes with the Arabic-script based languages, it has to deal with a rich morphological structure which ultimately provides low results. Lots of words are similar in Arabic, Urdu, Persian, and the Sindhi language in terms of orthography. But the meaning of the same words with reference to the context has changed from language to language. Many researchers have been working so far in order to develop the IR system of Pakistani languages. This task of developing an IR system provides access to users for retrieving data written in Sindhi. This advancement in technology helps a lot in almost every field of life, such as e-governance, agriculture, rural health, education, national resource planning, disaster management, and others. The development, the testing, and the training of the IR system involve a huge collection of documents. In this regard, we approach the numerous sources for the collection of Sindhi documents, specifically Sindhi books in soft form. We used material freely available on the internet to collect Sindhi language books, and we visited local publishers to request soft copies of their published books.

We explore a number of techniques and methods used to accomplish the task of information retrieval. Among the various approaches, we selected the Pattern Discover (PD) data mining method to implement in the Sindhi IR system

because the PD method is very efficient and suitable. This approach not only provides the best Sindhi patterns, but it is also rich in other structural patterns that can be identified from the huge dataset which includes many sequences, subsequences, and substructures [6]. The PD approach works with the user's, groups, and organization's defined patterns among the number of available patterns like open, close, and max patterns [7] [8]. The Google search engine provides a number of language interaction opportunities, including the Sindhi language. Searching through Google in Sindhi is also available, but the accuracy percentage is not at a satisfactory level. The Sindhi document retrieval process is quite different from the rest of the recognized languages, like English, due to its rich morphological structure. In Sindhi, every letter has its own diacritic sign on it, which is its phonological property. The Sindhi text used in books, newspapers, and magazines is written without diacritics. The frequent use of compound and homographic words in writing creates ambiguity for novice readers and children. This uncertainty is created due to the similar shapes of letters and the location of dots. Many Sindhi words contain prefixes and suffixes. Hence, for computational processing of Sindhi text, there is a need for a stemmer. All the reasons mentioned above enunciate the need for the development of an IR system. The information percentage is increasing day by day, which ultimately increases users' requirements for efficient, authentic, and user-friendly search and retrieval systems [9].

2. LITERATURE REVIEW

Literature reviewed against the proposed method of Sindhi documents retrieval system for students is divided into multiple directions such as information retrieval, indexing of documents, topics, modeling, topic searching, data mining, and pattern discovery. A scheme was presented in [10] for IR that was completely based on text information. The proposed approach is based on ensemble and data-driven techniques, which eventually highlight the combinations of various default sets. Esposito [11] presented another text-based data retrieval mechanism that used a question-answer scenario to recover the data. The machine used natural language-based questions, which were used as an input for the data retrieval process. A challenging task of plagiarism detection is done by Roostae [12]. The proposed system consists of three phases where information extraction is performed in the first phase from the user's query. After extracting information, the system reduces the record calculations in the first phase. The detailed analysis of the candidates is done in the second phase, after which the required information is retrieved. For better results, the first and second phases were combined, and this task is considered the third phase of the system. The indexing of documents is another complex task but mandatory for IR systems. A lot of IR systems are focused on document indexing so that the job of data retrieval can be done rapidly and accurately. However, the term "inverted files" is widely used in document indexing. Qureshi [13] proposed an indexing approach for a natural language named Urdu. In the proposed research work, the author recommended that the creation of indexes does not require ending words and that indexed files should be in ordered format. Alotaibi [14] presented a study that was conducted independently on IR stemming. The core purpose of the research work is the reduction of morphological properties of words in cognitively inspired languages. The IR systems are mainly designed for the purpose of understanding user queries. Hence, in this regard, Fernandez [15] presented a model of word embedding. The model extracted the relevant information from its job description. While the BM25 classic model performed the tasks of information sorting, document indexing and document ranking as well. This research is aimed at retrieving Sindhi documents. To achieve this type of task, many of the text mining applications use topic modelling approaches, including IR-based applications. Exploitation of topic modelling is an important concern in text mining applications. Hence, the topic evaluation method is used to access the efficiency and quality of topics in the model. Xu [16] proposed a model for the evaluation of topics in the model. The author used concept matching in ontology. To solve the problem of searching for similar concept words, the authors used the word embedding technique. State-of-the-art based models such as term-based, phrase-based, and topic-based were used for the evaluation of the proposed model. A bug localization method is proposed by Yang [17]. The proposed method determines the matching topics according to the generated bug report. After that, the extraction of similar bug reports along with equivalent pledge information similar to those topics is performed by the system. The model training is performed through a convolutional neural network using a long-term and short-term memory algorithm based on these extracted features. Moreover, there are a lot of search engine platforms available for user interaction, which involves numerous online activities. Since the search for a particular topic or knowledge is a common informational query that involves user intention, In order to improve the capacity of supervised models regarding knowledge prediction, Yu [18] presented a set of features known as resource-centric features. The proposed features increase the user's knowledge of the state of web search sessions.

The Internet IR is a challenging task, especially when there is a huge flow of data. This scenario makes text feature extraction quite important, which plays a key role in increasing the efficiency of IR in web mining and text mining applications. A customer's actions on the web site are categorized into two distinct sets: one is searching and the other is browsing. To avoid bringing a huge volume of inappropriate data and to fulfil the user's requirement for particular topic searching, the user needs to deliver a web scheme while demanding a search for knowledge on the internet [19]. Providing services to users is a vital task. This object is achieved in a tailored manner by implementing PD. Because the data dissemination activity is brief, the user's personalization experience is enhanced. In this regard, the proposed method should be capable of dealing with additional non-related data and provide users with a friendly experience by extracting and retrieving related information to fulfil user requirements [20]. The user's directional and navigational performance helps them attain web services efficiently. Using PD methodology, this task of benefiting

from a user's experience and information was completed [21] [22].This technique performs the tailoring function for a website and helps to search for specific requested content in a document from a huge data set. Another study was done on the semantic structure of web data by Kumar [23]. The author preferred the combination of feature extraction as well as feature selection methodology in order to improve the data mapping and data retrieval processes. Standard features were used, which increased the effectiveness of the task of text mapping.

The Web is a common platform for the collection of huge amounts of data where text data is available in a structured manner. The task of PD requires different data models to be implemented with web data, apparently as per user requirements. To make any web site responsive to user requirements, the web data can be evaluated using web mining services for user navigation patter and PD. To satisfy the user's need for particular information search, Zhong [7] proposed a method of pattern deploying and pattern evolving. The aim of this new method is to improve and increase the effectiveness of discovered patterns in both terms of usability and upgradation. A detailed analysis was made by Aggarwal [24] on numerous techniques of PD on the ground of data mining. Researchers conducted the analysis by keeping the domain factor in mind, as different domains have respective advantages and limitations in PD. Based on certain parameters, the authors made a comparative analysis of various PD approaches. Similarly, Dipli [25] proposed a data mining approach for the purpose of pattern mining. The authors adopted a term-based approach for text mining tasks and presented advanced techniques for pattern deploying and evolving.

Eldin [26] presented an enhanced approach to opinion retrieval to accomplish the task of user's requirements recognition. The proposed method is based on opinion mining, which comprises explicit features. Initial requirements in the presented methodology expand with the help of Arabic opinion-based heuristics and linguistic patterns. The accuracy of the system was measured based on several factors, which included the importance of opinion, weight of features, and sentiment polarity. The importance of IR systems besides data retrieval activities also influences usable in other computational fields. To provide authentic information to an individual in a well-mannered opinion, Hayashi [27] developed a system to overcome the gap between professionals and nonprofessionals. To efficiently implement the knowledge base task, the author acquired reliable and consistent information. Scholars majorly focused on certain blood diseases and collected the respective data in text format. Extraction of relevant content is performed on the collected text-based dataset.

3. PHASES OF SINDHI IR SYSTEM

This research work is mainly focused on the development of IR system for Sindhi language. Some basic elements of the research are discussable which are query formulation, identification and recognition of Sindhi text-based documents, documents indexing, pattern mapping technique and documents ranking. Moreover, among these mentioned phases, some are further divided in to the sub phases. Figure 1 is the complete depiction of proposed research methodology.

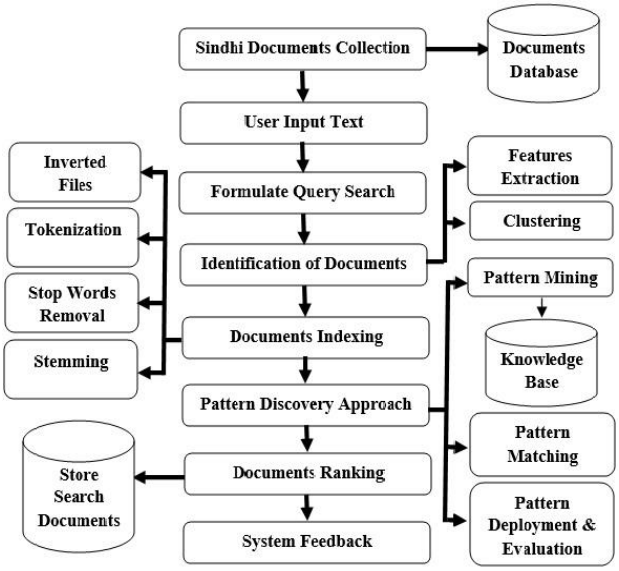

FIGURE 1. Proposed research methodology.

To implement the proposed approach of PD, we need a set of documents written in Sindhi language. A database of Sindhi documents is created where all the scanned and downloaded Sindhi documents and Sindhi books were stored separately. Hence, this phase of documents collection is considered as most significant phase after the task of finding research gap. It is therefore mentioned as the first phase of Sindhi IR system. The induction of main six phases used to accomplish the task of Sindhi IR system is presented below.

A. DOCUMENTS DESCRIPTION

Based on the adopted approach in [7], we split all the collected documents in to the paragraphs where each document *d*

provide a set of paragraphs PD(*d*). Assume *D* as a training set consisting of a positive set of documents D^+ and negative set of documents D^- as well. Let $T = \{t_1, t_2, \dots, t_m\}$ be a set of keywords which can be retrieved from the set of positive documents. This research work is aims at retrieving topic based documents. Hence, in this regards we uses topic modeling technique similar to [16]. This technique is based on a collection of algorithms work collaboratively to accomplish the task of identification specific topic from a set of documents. Numbers of approaches have been introduced so far for the task of hidden topic generation. Latent Dirichlet Allocation (LDA) is one of them widely used for this job. The basic concept of LDA is that it considered a single document as a multinomial distribution topic where each topic is multinomial distribution over words. Let $D = \{d_1, d_2, \dots, d_M\}$ be a collection of *M* documents where each document is represented by topic distribution$\theta_d = \{V_{d,1}, V_{d,2}, \dots, V_{d,v}\}, \sum_{j=1}^{v} V_{d,j} = 1, V_{d,j} = P(z_j|d)$, while 'v' represents the number of topics. The probability distribution over words represents each topic. Hence, for the j^{th} topic, we have $\Phi_j = \{\emptyset_{j1}, \emptyset_{j2}, \dots, \emptyset_{jm}\}$, m is the number of words per topic, $\emptyset_{ji} = P(w_i|z_j)$. $P(w_i|d) = \sum_{j=1}^{v} P(w_i|z_j) * P(z_j|d)$is used to calculate the word probability w_i in the document *d*. Each topic z is represented as a set of words which is denoted by$T(z) = \{w_1, w_2, \dots, w_m\}$. The term probability distribution is used for the task of words to topic conversion where high probability words can be termed as topical words.

B. QUERY FORMULATION

Sindhi IR system we developed is capable to take Sindhi text-based input from the user, process the input data and retrieve the respective prerequisite documents. In the IR system, user put the request in the form of a query, the inputted query then processed by the system and converted into the system understandable form for further processing. Subsequently, query formulation is the upcoming phase and is considered as more complex and more important especially when there is a request of information retrieving of a particular content from a huge dataset of documents. This request of document finding is passed to the IR system.

The approach we used to develop an IR system for Sindhi language is same as described in [8]. Also it defines mathematically issue with the development of IR system. Assume we have group of *m* objects $\wedge = \{\wedge_1, \wedge_2, \dots, \wedge_m\}$ and the group of *n* terms $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_n\}$. Every object $\wedge_i$ is a subset of terms in $\mathcal{T}$ ($\wedge_i \subset \mathcal{T}, \forall i \in [1 \dots m]$). Given the group of queries $\mathcal{Q} = \{\mathcal{Q}_1, \mathcal{Q}_2, \dots, \mathcal{Q}_l\}$, where every query $\mathcal{Q}_i$ is composed by the group of terms, that is, $\mathcal{Q}_i \subset \mathcal{T}$, the purpose of IR system is to find, for every query $\mathcal{Q}_i \in \mathcal{Q}$, the most relevant subset of objects $\wedge'$, such that $\wedge' \subset \wedge$.

C. DOCUMENTS IDENTIFICATION

Identification of documents is the next phase after query formulation. This phase is consisting of feature extraction techniques and document clustering approach. Word mapping task can be done using the method word frequency. In this method the frequency of the word measured as an extracted feature. The text feature vector is maintained with the counted word frequency. Resemblance of words can be identified using distance calculation through dot product and cosine similarity. A trained classifier is used for the text-based classification with the help of text feature vector. While for the feature extraction of Sindhi words, we use N-gram based word frequency model [28]. The calculated word frequencies also help to measure the retrieved documents.

The task of documents collection having similarities from a created database of huge documents, the best and preferable technique is document clustering technique. Due to the authenticity in text based data classification by reducing the dimensionality of feature space, most of the researchers preferred clustering techniques. Implementing K-means clustering algorithm provide quite acceptable results to researchers [23]. The clustering algorithm collecting mapping-words and convert them to a single feature. This approach of clustering is suitable for small type of datasets. It provides accurate results in text classification where there is an involvement of smaller dataset. Furthermore, we divided the document dataset in to small clusters where each cluster represents changing center. The process begins with the assigning of initial query value. K-mean clustering algorithm used to calculate the distance between input and center value. The output of k-mean clustering algorithm will then be used as an input value for the nearest center. The clustering process performed on the formulated vector space based on text features to detect natural clusters. To compute the clustered objects around the centroids $\mu i \forall i = 1 \dots k$following objectives should be minimized.

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

Where *k* is the number of clusters i.e. *Si*, *i*= 1, 2 , . . . , *k* and $\mu_i$is the mean point or centroid of all the points $x_j \in S_i$.

D. DOCUMENTS INDEXING

After performing the task of documents identification, the next phase in the developed IR system is document indexing. In this part of methodology, each document was indexed and assigned appropriate meaning. This task of indexing reduced and converts the normal document in to an understandable term of information. The purpose of document indexing was to correlate a document with the help of descriptor which basically highlight the features group. Therefore, Sindhi words and characters indexed and mapped as per document information. From the available literature, we had found number of techniques related to the document indexing. Based on the nature of language and document, we preferred inverted file technique of document indexing through which we indexed and mapped Sindhi documents. This method involved mapping characters with different documents and indexed accordingly. Vocabulary and occurrences are two key features which were used for mapping process. Scanned and other collected Sindhi documents were converted, and

text extracted from these documents with the help of built-in function of Python language named Tesseract. The text corpus was then stored using SQLite and based on the corpus an inverted indexing table is formed.

Mahar [29] proposed a model of tokenization. Same model was used in this research work for proper indexing of collected documents. Stop Words are such type of words which have a very little contribution in query processing. Hence, eliminating such words will get more accuracy in results. So, we decide to eliminate such type of common words. These stop words are most frequently used words in NLP but has a very less importance. Removing these words will definitely increase the IR system's performance. Furthermore, formulation of a list of such words will be the time-consuming task and requires much efforts. In this research work these stop words of Sindhi language are determined by two ways such as by calculating their frequency through $freq(w) = \sum_{k=0}^{k} w_k \in c$. Another way to determine these words was to seek the help of a Sindhi linguistic expert to highlight their grammatical status. With the help of a local expert we determined the importance of such words and made a list of stop words. However, not all the words are considered as stop words because several among them have some importance and frequently used in Sindhi language. Therefore, we were processed some of the commonly used Sindhi words. The sample of stop words is given in Table 1.

TABLE 1. **Sample Sindhi stop words.**

آ	ته	رهن	يا	وارن	هن	۾
ءِ	نه	ها	بس	هر	بن	جو
جي	تي	تن	سو	تو	سان	هو

Prefix and suffix structure of Sindhi words makes it somehow complex to be processed which ultimately affects the IR system's performance. While the use of Stemmer in IR systems made search engines more efficient and more operative. The use of Stemmer in the IR system basically performed the task of reducing word's morphological variants up to the root of stem. There is a huge collection of Sindhi words having concatenated structure of prefix and suffix morphemes like بداخلاق،،لاجواب where لا and بد are prefix morphemes to the words جواب andاخلاق. While سمهڻ and ملڻوارو words of Sindhi language are the example of suffix morphemes. In these Sindhi words, ڻ,و, are suffix morphemes respectively. Shah [30] proposed an algorithm which performed the task of Sindhi words stemming. The samples of prefix and suffix morphemes are given in the Table 2 which are frequently used in Sindhi language.

E. PATTERN DISCOVERY

In the PD phase of the research methodology, a rich documented database is searched for a particular pattern of document based on the formulated query. This approach is widely used in the field of data mining where the scenario came to deal with the huge amount of data. On the basis of user's requirement, number of pattern matching, and pattern mining techniques generated in this process. Alike, association rules and sequential pattern, the outcome of PD is also based on user's requirement. The use of Frequent Pattern Mining (FPM) technique for the purpose of discovery of hidden information from a set of documents demonstrated aplenty accuracy in results [8][31] [32]. The ultimate goal of the FPM is the identification of preferred patterns which have sustenance of "no lower than a given minimum support threshold". The pattern support in the range of defined threshold can be called as infrequent pattern. While crossing the defined threshold, the pattern may be termed as frequent pattern. Mining of frequent and common set of items can be performed using Apriori algorithm. According to the pattern of this seminal algorithm it performed level-wise mining. It required frequent use of all the non-empty subsets of item sets. At the k^{th} iteration (for k≥2), it forms frequent *k*-item set candidates based on the frequent (*k*-1)-item sets and scans the database once to find the complete set of frequent *k*-item sets, L_k. while all the generated information is stored in a separate knowledge base.

TABLE 2. **Sample prefix and suffix morphemes.**

Prefix Morphemes	نه	خوش	اڻ
	نا	سر	م
	بر	دست	پر
	بي	هم	وڏ
	خود	لا	مها
Suffix Morphemes	و	يو	ٽي
	وارو	يارو	ٽو
	واري	يان	آنتو
	ا	ن	ان
	ي	ڻ	وال

Moving forward, there is another issue with IR system is the user's query based on multiple keywords where the eventually retrieving matching document is the final goal. Vicinity and proximity of the keywords is the decisive factor for setting heuristics and ranking criteria. The task can be easy conditioned to the sorting order of the requested keywords in the document. Documents' relevancy and irrelevancy can be categorized using threshold technique. The approach we follow in this paper to accomplish the task of pattern matching is defined in [33]. Pattern matching performed through a set of queries along with the constraints that define the position of text in the document. Let $\mathcal{T}[0..n-1]$ be the text. This text has been preprocessed to answer various queries through a data structure; (1) given a pattern *P*, a position *p* and a rank *k*, output the position of the k^{th} occurrence of *P* in $\mathcal{T}[p..n-1]$, and (2) given a pattern *P*, and position *i* and *j*, count all occurrences of *P* in $\mathcal{T}[i..j]$. Each query can be considered as a two-dimensional range query, and there is existing geometric data structure with good query performance. But yet there are alternative approaches described to design the data structure. These approaches include augmentation of a binary search tree. As described in [33] the parameters we consider for the pattern matching queries are pattern *P*, position *p* and rank *k*. Among all occurrences of *P* in *T* with the matching location after text position *p*, reports the occurrence with the k^{th} smallest matching location. The difference between the research done

in [33] and in this paper is that we use suffix array data structure and store Sindhi words instead of characters while the remaining process of implementation is same as described in [33].

Understanding the discovered patterns is very important task after the successfully execution of all the previously mentioned phases. For the evaluation of term support, we need to understand discovered patterns in the light of summarization. More semantic meaning than selected terms gives high support [7]. The significance to these terms assigned on the basis of their appearances in the patterns. The performance was evaluated by applying several common measures.

F. DOCUMENTS RANKING

Ranking the documents contains results of discovered patterns and based on the given results and the relevancy of user's query, documents ranked. However, in order to uplift the authenticity and reliability of IR system, the document ranking process performed in the decreasing manner. By doing so, the most relevant document against the user's query are placed at top and less relevant document vice versa put in lower state. Hierarchical format is used for the storing of searched documents. Xu [16] determined the scoring process of documents ranking where the ranking score was used to sorting the relevant documents from all the searched documents through the probability distribution. Let $V_{D,j}$ be the average topic probability distribution of all documents in the training collection D, $\theta_D = (V_{D,1}, V_{D,2}, \dots, V_{D,v}), \sum_{j=1}^{v} V_{D,j} = 1$ and $V_{D,j}$ is measured as:

$$V_{D,j} = \frac{1}{|D|} \sum_{d \in D} P_r(z_j|d)$$

The topical words with $m_i > 1$ are selected from the documents database to represent the topic like the topical words whose probability is larger than the average probability.

$$sig(z,d) = \sum_{\substack{w_i \in d, w_i \in ET(z) \\ Pr(w_i) > avgPr(z)}} sig(w_i|z)$$

For a new incoming document *d*, the relevance score of *d* to the training collection *D* with v topics is measured using equation given below.

$$rank(d|D) = \sum_{j=1}^{v} sig(z_j, d)\ X\ Q^*(z_j)\ X\ V_{Dj}$$

Document ranking process for Sindhi documents is applied as mentioned above for getting authenticate results and is also discussed and applied in [16]. At the end, the developed IR system retrieved and provides user required document as an absolute result.

4. DOCUMENTS COLLECTION

Data collection or building a database is a very important component while developing an IR system. Similarly, we need a database of Sindhi document so that we can easily performed experiments by applying PD techniques and develop an effective Sindhi IR system. Developing a database step required collection of many Sindhi documents. All the required documents were collected from various sources which can be categorized as primary and secondary sources. Visiting libraries, publishers, searching websites for Sindhi documents, books and other related material and seeking friends help. Local publishers from the districts Khairpur, Shikarpur and Hyderabad are also visited individually and asked for the soft copies of their Sindhi published books. Also, numbers of Sindhi books were downloaded from various Sindhi websites available freely online for everyone. Next step is the storing all the collected material into the database. In our collection, we found most of the books are related to Sindhi literature and Sindhi poetry. Where, two most considerable books we use in our research were the Sindhi translation book of Holy Quran and a well-known Shah-jo-Risalo book. The IR system we developed is completely based and worked with the developed database. The document retrieving and ranking process performed on these stored documents based on user query.

Table 3 provides the statistical information of all the collected Sindhi books. We have collected 31 books in total number consisting of 4466 pages. The number of paragraphs and lines are also calculated which were 72798 and 123715 respectively. A table formed which contained all the computed words and characters of the books. These classifications of data help in result calculation and evaluation.

TABLE 3. **Statistical information collected from books.**

Book Name	No. of Pages	No. of Paragraphs	No. of Lines	No. of Words
Holy Quran	239	886	7809	169522
Shah Jo Risalo	312	6820	14136	38065
DilJiKhanaBadoshi	193	1762	5097	22349
Dais 2	27	172	697	8931
Zindagy	101	1193	2827	22201
Sindh jiTarikh	97	376	1900	25480
Shairi Ali dost 2	42	980	1010	3670
ShairyKousar	264	1212	7127	26698
Kadahn b Acheen	65	1278	1283	7085
Short Articles	165	1274	4091	63969
NethShaKaje	297	1597	6447	94576
JaniShairi	163	1866	4247	10711
Shairi Ali Dost	25	629	634	3706
Islamabad 4 Denh	47	322	1309	24027
Hawai j Hawaun M	28	456	3392	36057

SirajSahab	7	203	1049	6777
Chehra 2	82	541	2351	30071
M A Sial	48	519	1752	24308
Pardais	29	172	812	10052
Jail JaDeenh	6	33	162	3090
OadBrothery	136	2368	2449	25566
OadIbhyas 2	174	2547	2623	25806
Dictionary	1553	40353	40353	63545
Shabir Poetry	112	2057	2434	13516
Nirdosh (Poetry)	36	574	574	3742
AdarshInsaan	88	810	3472	53604
Las Vegas	35	208	1014	12684
Qaidi	8	59	248	4575
ShairMaloomat	43	415	980	6348
OadIbhyas 1	27	680	691	7211
Storeroom	19	436	745	5068
Total	**4466**	**72798**	**123715**	**853010**

5. EXPERIMENTATION AND RESULTS

A database of 31 collected Sindhi books were used for the creation of Sindhi document corpus. Further the corpus categorized into 20 sub parts. For the training and testing of developed IR system, we use separate datasets of Sindhi documents. All the created datasets are shown in Table 4. Moreover, 185 male and female students of Shah Abdul Latif University were selected to perform experimentations following PD approach and to carry out the evaluation process of developed Sindhi IR system. The input queries to the system were formatted as a single word, double word and sentence based scenario. Based on the selected students, we received 5550 input queries for the system as we allow each student to submit up to the 30 queries where not more than 10 queries for one scenario. Furthermore, we selected 50 queries randomly for each scenario as an input to the IR system. For better evaluation of results, we calculated individual as well as collective results. Furthermore, the availability of hardware affects on the performance of the IR system. The Intel Core i5 computer running with 8 GB of RAM was used so that the execution time and other parameters can be recorded accurately. The Internet connection was used as the 10Mbps LAN card with a link of 2 Mbps as a DSL connection.

It is important to evaluate the IR system using some evaluation techniques. The Precision and Recall are the two most frequent, common and well-known metrics through which the effectiveness and authenticity of the developed IR system can be measured. Precision can be termed as the friction of retrieved document. It can be represented as division of the number of correct retrieved documents with the total number of retrieved documents. While the term Recall is the retrieved fraction of the relevant documents. The formula of precision and recall given below were used to measure the performance of developed IR system.

$$Precision = \frac{\#(RelevantDocumentRetrieved)}{\#(RetrievedDocuments)}$$

$$Recall = \frac{\#(RelevantDocumentsRetrieved)}{\#(RelevantDocuments)}$$

TABLE 4. **Datasets of Sindhi documents.**

Topics (Categories)	**Document Number**
Poetry	431
Science	324
Politics	278
History	487
Novel	279
Story	345
Social	313
Legal	403
Education	422
Psychology	301
Computer	426
Business	212
Media	536
Religion	437
Sociology	425
Literature	441
Housing	228
Translations	224
Health	319
Miscellaneous	452
Total	**7283**

A. SENARIO 1: SINGLE WORD QUERIES

The experiments were performed separately with each scenario to compare the performance of Sindhi IR system which is based on our proposed research methodology. Hence, initially single word query scenario was given to the selected students of department of Computer Science where they all are asked to write queries based on a single word only. Experiments done on the randomly selected 50 queries and based on them the performance of developed IR system for Sindhi language was evaluated. Based on inputted queries of single word, we have calculated the precision and numbers of all the retrieved documents from the documents stored in the database.

The graphical representation of retrieved precisions with single word queries based on selected 50 queries is depicted in Figure 2. The minimum and maximum received precisions are discussed to evaluate the performance of Sindhi IR system. 0.33 was the minimum recorded precision. While the maximum recorded precision was at 0.77. An average precision is also calculated with all the 50 input queries which were recorded as 0.6538%. The overall percentage of retrieved document is at acceptable level. Also, the number of retrieved documents with respect to each query was calculated to measure the efficiency of the developed Sindhi IR system. Figure 3 is the illustration of analysis of all the retrieved documents based on selected 50 queries. The lowest documents were received with query No. 10 which is 48 and

the maximum documents were retrieved with query No. 8 which is 6651.The average of 2896 documents was retrieved with single word queries.

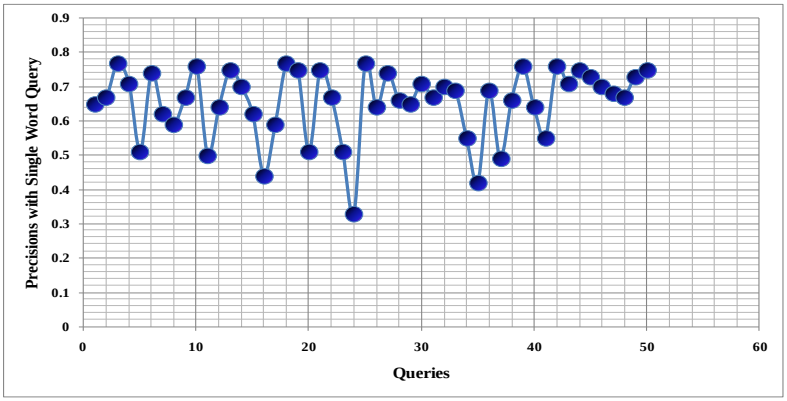

FIGURE 2. **Calculated precisions with single word queries.**

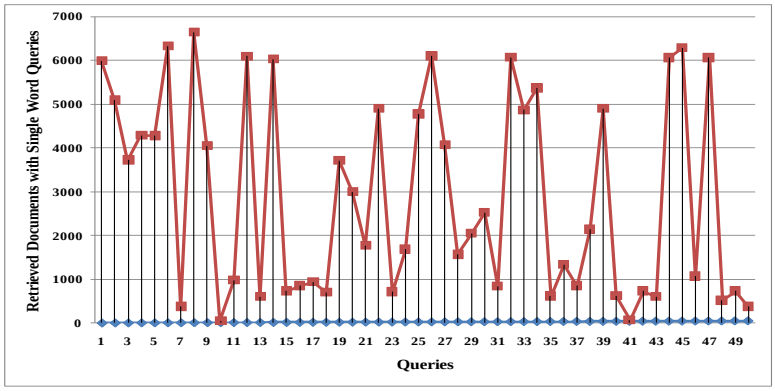

FIGURE 3. **Retrieved documents with single word queries.**

B. SENARION 2: DOUBLE WORD QUERIES

After performing experiments on single word queries, we also performed experimentation on double words queries in order to measure and evaluate the performance. Based on the situation given, students were advised for ranking the documents and feedback. After performing the experiments on double words-based queries, retrieved documents were ranked into the levels from less relevant to the most relevant documents.

Likewise in single word-based queries, also 50 double words-based queries were randomly selected for evaluating the performance of Sindhi IR system where rising and falling ratios depicts different performance averages at different levels. The received precision rate of 50 experimented queries is depicted in Figure 4. The received results proved that the highest precision is calculated in between query 20 to 25. The least precision value we found was 0.33 and maximum was 0.88 while 0.67 was the average precision recorded value. Beside this, we also calculated minimum, maximum and average cases of documents retrieved with double words-based queries. The inferior retrieved documents of 6 and highest percentage of 116 documents were retrieved with the average value of 74 documents. Initiation of double words-based queries and documents retrieved against them are illustrated in Figure 5. It is observed that at different places more than 100 documents were retrieved from the query no 11 to 43.

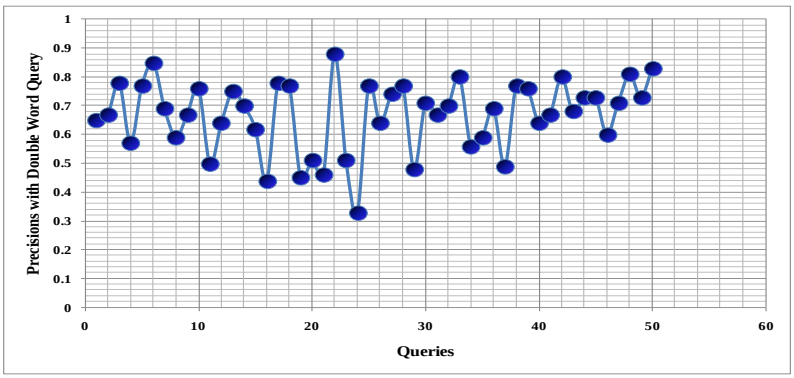

FIGURE 4. **Calculated precision with double word queries.**

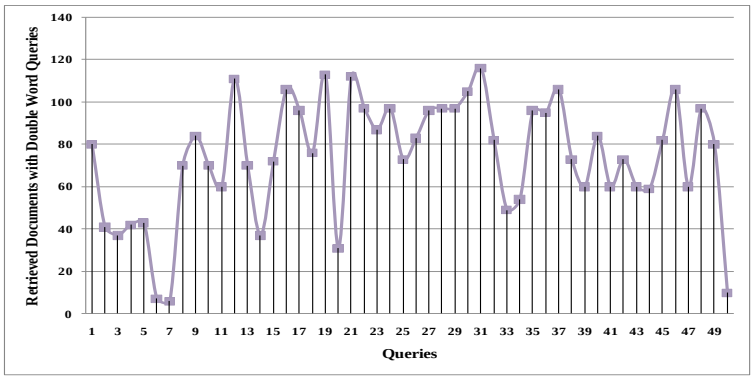

FIGURE 5. **Retrieved documents with double word queries.**

C. SENARION 3: SENTENCE QUERIES

Third scenario to be performed with the developed IR system for Sindhi language was Sentence based queries where selected students were guided to write their queries in a sentenced structure. Reducing the number of documents to be retrieved increased the execution time of developed IR system while performing experiments on sentence-based queries. Table 5 is the complete picture of queries values, precision rate and number of retrieved documents.

The graphical representation of the calculated precision values with sentence-based queries is illustrated in Figure 6. Minimum and maximum precision values for sentence-based queries were recorded as 0.32 and .067 respectively and 0.54 was the average precision value record for the same scenario.

Analytical representation of retrieved documents with sentence-based queries is shown in Figure 7. In Figure 7, it can be seeing that minimum and maximum, maximum and average retrieved documents with query no.7 are 3, 82 and 27 respectively. Moreover, we also performed precision and recall experiments individually on the extracted dataset of documents. To obtain accuracy in results with developed IR system for Sindhi language, we also created query set. Collective results of precision and recall achieved with different scenarios such as Single-Words, Double-Word and Sentence-Word are depicted in Table 6. Minimum and maximum averages of precision rate were recorded as 0.11 and 0.77 respectively. While with recall rate, 0.4 and 0.8 were the recorded minimum and maximum rates respectively.

TABLE 5. **Calculated precisions and retrieved documents based on sentence queries.**

Query No.	Precision	Documents Retrieved	Query No.	Precision	Documents Retrieved
1	0.45	5	26	0.62	4
2	0.58	13	27	0.45	41
3	0.61	14	28	0.48	38
4	0.45	46	29	0.47	41
5	0.63	6	30	0.64	28
6	0.6	6	31	0.67	41
7	0.38	3	32	0.4	35
8	0.5	61	33	0.61	46
9	0.67	13	34	0.54	35
10	0.64	6	35	0.59	48
11	0.5	10	36	0.63	35

12	0.37	12	37	0.49	9
13	0.48	25	38	0.41	12
14	0.6	13	39	0.48	36
15	0.62	9	40	0.62	13
16	0.66	4	41	0.67	31
17	0.58	10	42	0.61	33
18	0.55	20	43	0.65	34
19	0.57	19	44	0.59	4
20	0.51	82	45	0.36	49
21	0.49	60	46	0.6	21
22	0.64	58	47	0.53	21
23	0.56	25	48	0.59	22
24	0.39	70	49	0.44	27
25	0.4	20	50	0.55	28

The developed IR system offers three types of query-based scenarios i.e., single word, double word, and sentence. The purpose of above given Table 5 is to describe the average calculated precision with sentence-based queries and the total number of retrieved documents against the specific query number. During the experiments, we have assigned a separate number to each inputted query.

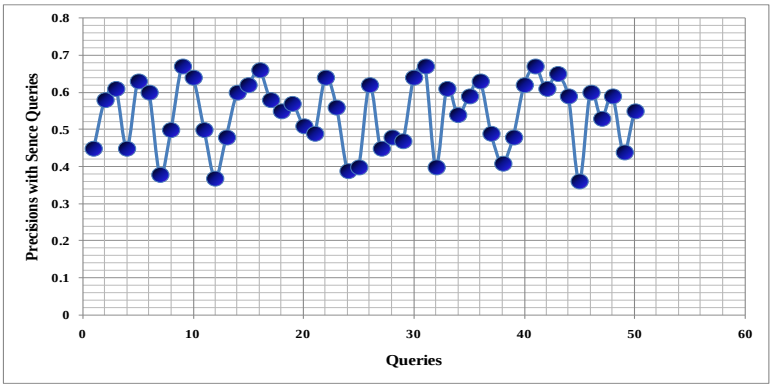

FIGURE 6. **Calculated precision with double word queries**

In the same way results with double words based scenario were also calculated where we got 0.11 and 0.89 as minimum and maximum average precision rates respectively. 0.36 was the recorded recall rate as minimum average and 0.87 as maximum average recall rate. Similarly with sentence based queries, the developed system provide 0.09 as minimum precision rate and 0.67 as maximum precision rate along with 0.21 and 0.69 as minimum and maximum recall rates respectively.

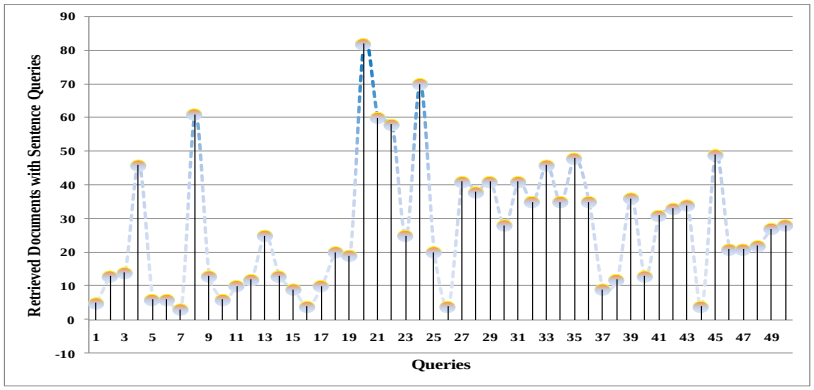

FIGURE 7. **Retrieved documents with sentence queries.**

TABLE 6. **Precisions and recall on collected Sindhi documents.**

Scenario	**Precision**		**ReCall**	
	Min (avg.)	Max (avg.)	Min (avg.)	Max (avg.)

Single Word	0.11	0.77	0.4	0.8
Double Word	0.11	0.89	0.36	0.87
Sentence	0.09	0.67	0.21	0.69

D. EXECUTION TIME AND RETRIEVED PAGES

During the experiments, the execution time taken by the system was calculated with each adopted process. Experiments were performed in order to retrieve required documents and to calculate the execution time as well. The intervened time taken by the system while executing multiple queries such as single word based, double words based and sentence based queries was recorded and illustrated in Table 7. However the elapsed time taken by the system to execute sentence based query was inferior then words. Also minimum and maximum time taken by the system is shown in Figure 8.

TABLE 7. **Execution time against retrieved documents.**

Scenario	Retrieved Documents	Time Elapsed
Single Word	144820	0.34-7.9
Double Word	3698	0.29-6.3
Sentence	1342	0.17-3.8

Same method was used to accomplish the task of evaluation of execution process of inputted query and to represent pages fetched during the time of execution. Figure 9 shows the query number with respect to the particular retrieved document and its time of execution as well. Based on all the mentioned scenarios, the total number of single words, double words and sentences were quite huge. Hence, with respect to the requested queries, retrieved documents are also many in numbers. So it is impossible to show all the retrieved results in one figure. Execution time of only 130 queries can be shown in Figure 9.

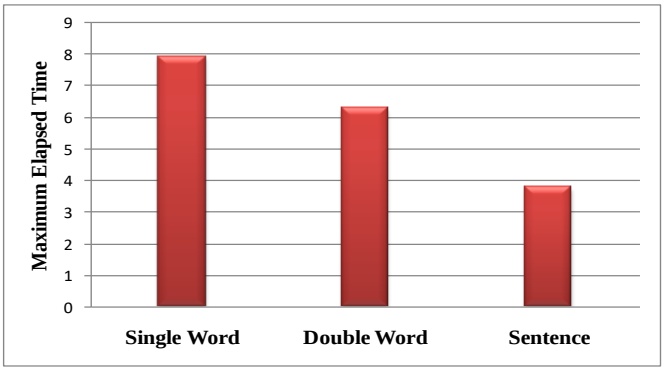

FIGURE 8. **Elapsed time of various experiments.**

We have compared the results of our proposed approach with existing approaches. The comparison of the approaches in terms of calculated precision is shown in Table 8. The received results are analyzed with the selected approaches and the datasets. The best results of information retrieval are reported with the approach of features extraction in which web data is used [23]. The inferior precision is reported with the approach of RCNN using MAHE datasets [9]. On the other hand, the calculated precision value of our proposed approach and the cluster-based approach described in [8] is same but the datasets are different.

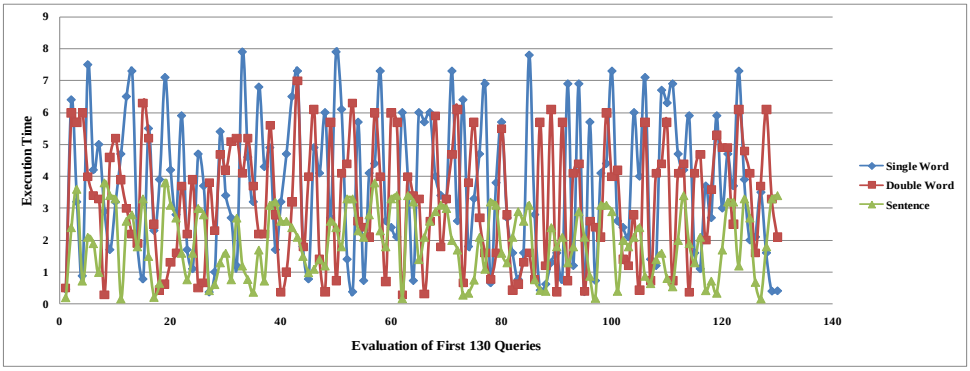

FIGURE 9. **Evaluation of queries with all scenarios and execution time.**

TABLE 8. **Comparison of the precision value of proposed and existing approaches.**

Selected Approach	Datasets	Precision
Cluster-Based [8]	Wikipedia	89%
RCNN [9]	MAHE	87%
Features Extraction [23]	Web Data	90%
Pattern Discovery, This Work	Sindhi Text Data	89%

6. CONCLUTION AND DISCUSSION

The developed IR system is completely based on and works only with the Sindhi language. UTF8 is the writing character code scheme that we used in this research work. The word indexing strategy is the method that we used for our IR system in order to index the documents. 50 queries were randomly selected for precision throughout the scenarios and are shown in Figure 10. In Figure 11, we illustrate all the retrieved documents against the processed queries. The outcomes of this IR system showed that the Sindhi IR system's performance is good enough and acceptable with the queries based on single words, double words, and sentences. while some minute differences were also recorded during the experimentation process. With the received percentage of the IR system, it is analyzed that the system will continuously increase its performance as long as the documents are jointly used for processing. Hence, appending more documents will ultimately increase the performance of the developed IR system. The developed IR system provided much more accurate results. A PD technique is used in this IR system for performing experiments. With the obtained results, the performance of the IR system can be justified. 0.77 is the acquired value with single word-based queries. The unclear and uncertain structure of prefix and suffix-based words and the similarity of the characters in terms of shape are the two basic reasons behind the less accuracy in results with single word-based queries. Though the acceptable value of the maximum recall average was received, it was 0.8. The system provided more accurate results with double word-based queries due to the high recall and precision rate, which were recorded at 0.89 and 87, respectively. This betterness in results proved that the developed IR system worked more accurately with double word-based queries.

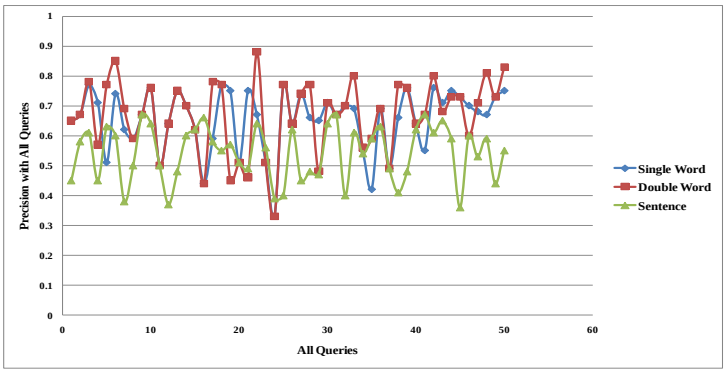

FIGURE 10. ***Queries and precisions in all scenarios.***

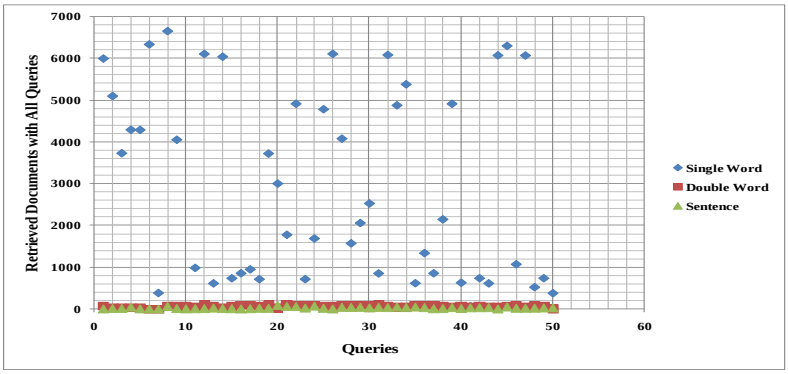

FIGURE 11. ***Queries and retrieved documents with all scenarios.***

Furthermore, the system provided deprived results with sentence-based queries as expected, specifically in terms of average precision and average recall values. The concluded reasons behind this failure are the complex, semantic, homographic, and compound-structured words of the Sindhi language. However, adding more documents to the developed database for the Sindhi language will surely improve the overall results. Elapsed and execution time for each query are calculated separately. Also, the same time intervals are calculated for retrieved documents. 144820 documents were

retrieved by the developed IR system with single word-based queries with an estimated elapsed time of 0.34 to 7.9. With the double word-based queries, the system successfully retrieved 3698 documents and 1342 documents with sentence-based queries. The calculated elapsed time in both cases was between 0.29 and 6.3 and 0.17 to 3.8, respectively. The obtained results proved that a large number of documents were retrieved with a minimum time investment by the processor.

This paper presented intelligent approaches for the retrieval of Sindhi textual information. The novelty of this research work is the use of document indexing and pattern discovery approaches using complex morphological structures of Sindhi text. The obtained outcomes proved that the proposed state-of-the-art based approach can be applied to any Arabic script-based language like Urdu and Persian. The contributions of this research work are complete understanding and identification of complex issues and challenges in the development of the Sindhi IR system. The pattern discovery approach is tested on Sindhi text. A Sindhi document database is developed and a list of prefix and suffix morphemes and stop words is prepared.

7. REFERENCES

[1] A. Jain, G. Kulkarni, and V. Shah, "Natural language processing", *Int. J. of Comput. Sci. and Eng.*, vol. 6, no. 1, pp. 161-167, 2018.

[2] K. B. Vijaya, and M. M. M. Fuad, "A new method to identify short-text authors using combinations of machine learning and natural language processing techniques", *Procedia Comput. Sci.*, vol. 159, pp. 428-436, 2019.

[3] W. A. Narejo, J. A. Mahar, S. A. Mahar, F. A. Surahio, and A. K. Jumani, "Sindhi morphological analysis: An algorithm for Sindhi word segmentation into morphemes", *Int. J. of Comput. Sci. and Inf. Sec.*, vol. 14, no. 6, pp. 293-302, 2016.

[4] J. A. Mahar, and G. Q. Memon, "Automatic diacritics restoration for Sindhi", *Sindh Univ. Res. J.*, June 2011, vol. 43, no. 1, pp. 43-50, 2011.

[5] M. Madankar, M. B. Chandak, and N. Chavhan, "Information retrieval system and machine translation: A review", *Procedia Comput. Sci.*, vol. 78, pp. 845-850, 2016.

[6] S. Kumar, and R. Kumar, "A study on different aspects of web mining and research issues", *in Proc. ICCRDA*, vol. 1022, Rajpura, India, 2020, pp. 1-10.

[7] N. Zhong, Y. Li, and S. T. Wu, "Effective pattern discovery for text mining", *IEEE Trans. on Knowled. and Data Eng.*, vol. 24, no. 1, pp. 30-44, 2012.

[8] Y. Djenouri, A. Belhadi, D. Djenouri, and J. Chun-Wei Lin, "Cluster-based information retrieval using pattern mining", *Appl. Intell.*, vol. 51, pp.1888-1903, 2021.

[9] H. S. Chiranjeevi, K. Manjula, and Shenoy, "Advanced text documents information retrieval system for search services", *Cogent Eng.*, vol. 7, no. 1, pp. 1-16, 2020.

[10] I. Safiulin, N. Butakov, D. Alexandrov, and D. Nasonov, "Ensemble-based method of answers retrieval for domain specific questions from text-based documentation", *ProcediaComput. Sci.*, vol. 156, pp.158-165, 2019.

[11] M. Esposito, E. Damiano, A. Minutolo, G. De Pietro, and H. Fujita, "Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering", *Inf. Sci.*, vol. 514, pp. 88-105, 2020.

[12] M. Roostaee, M. H. Sadreddini, and S. M. Fakhrahmad, "An effective approach to candidate retrieval for cross-language plagiarism detection: A fusion of conceptual and keyword-based schemes", *Inf. Proces. &Manag.*, vol. 57, no. 2, pp. 102-150, 2020.

[13] M. M. Qureshi, and M. Shoaib, "An efficient indexing and searching technique for information retrieval for Urdu language", *Pakistan J. of Sci.*, vol. 62, no. 3, pp. 172-176, 2010.

[14] F. S. Alotaibi, and V. Gupta, "A cognitive inspired unsupervised language-independent text stemmer for information retrieval", *Cog. Sys. Res.*, vol. 52, pp. 291-300, 2018.

[15] F. C. Fernandez-Reyes, and S. Shinde, "CV retrieval system based on job description matching using hybrid word embeddings", *Comput. Speech & Lang.*, vol. 56, pp.73-79, 2019.

[16] Y. Xu, H. Nguyen, and Y. Li, "A semantic based approach for topic evaluation in information filtering", *IEEE Access*, vol. 8, pp. 66977-66988, 2020.

[17] G. Yang, and B. Lee, "Utilizing topic-based similar commit information and CNN-LSTM algorithm for bug localization", Symmetry, vol. 13, no. 406, pp. 1-18, 2021.

[18] R. Yu, R. Tang, M. Rokicki, U. Gadiraju, and S. Dietze, "Topic-Independent modeling of user knowledge in informational search sessions", *Inf. Ret. J.*, pp. 1-29, 2021.

[19] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach", *Data Min. and Knowl. Disc.*, vol. 8, pp.53-87, 2004.

[20] H. Peng, "Discovery of interesting association rules on web usage mining," *in Proc. IEEE ICMC*, Hong Kong, 2010, pp. 272-275.

[21] N. Sael, A. Marzak, and H. Behja, "Web usage mining data preprocessing and multi level analysis on model", in Proc. AICCSA, Ifrane, Morocco, 2013, pp. 1-7.

[22] N. Lakshmi, R. S. Rao, and S. S. Reddy, "An overview of preprocessing on web log data for web usage analysis", *Int. J. of Innov.e Technol. and Explor. Eng.*, vol. 2, no. 4, pp. 274-279, 2013.

[23] C. S. Kumar, and R. Santhosh, "Effective information retrieval and feature minimization technique for semantic web data", *Comput. & Electrical Eng.*, vol. 81, 106518, pp. 1-14, 2020.

[24] M. Aggarwal, and A. Bhatia, "Pattern discovery techniques in online data mining", *Int. J. of Eng. and Techni. Res.*, vol. 3, no. 7, pp.28-31, 2015.

[25] C. Dipali, Sonawane, P. Tejal, Shirole, D. Kajal, Patil, V. Priyanka, Patil, K. Amol, and Patil, "Effective pattern discovery for text mining", *Int. Res. J. of Eng. and Techno.*, vol. 4, no. 4, pp. 194-197, 2017.

[26] S. S. Eldin, A. Mohammed, H. Hefny, and A. S. E. Ahmed, "An enhanced opinion retrieval approach on Arabic text for customer requirements expansion", *J. of King Saud Univ.-Comput. and Inf. Sci.*, vol. 33, no. 3, pp. 351-363, 2019.

[27] T. Hayashi, and Y. Ohsawa, "Information retrieval system and knowledge base on diseases using variables and contexts in the texts", *Procedia Comput. Sci.*, vol. 159, pp.1662-1669, 2019.
[28] H. Shaikh, J. A. Mahar, and M. H. Mahar, "Instant diacritics restoration system for Sindhi accent prediction using n-grams and memory-based learning approaches", *Int. J. of Adv. Comput. Sci. and Appl.*, vol. 8, no. 4, pp. 149-157, 2017.
[29] J. A. Mahar, H. Shaikh, and G. Q. Memon, "A model for Sindhi text segmentation into word tokens", *Sindh Univ. Res. J.*, vol. 44, no. 1, pp. 43-48, 2012.
[30] M. R. Shah, H. Shaikh, J. A. Mahar, and S. A. Mahar, S. A., "Sindhi stemmer for information retrieval system using rule-based stripping approach", *Sindh Univ. Res. J.*, vol. 48, no. 4, pp. 891-897, 2016.
[31] W. Gan, Lin, JC-W, H. C. Chao, H. Fujita, and S. Y. Philip, "Correlated utility-based pattern mining", *Inf. Sci.*, vol. 504, pp. 470-486, 2019.
[32] U. Yun, D. Kim, E. Yoon, and H. Fujita, "Damped window based high average utility pattern mining over data streams", *Knowledge-Based Sys.*, vol. 144, pp.188-205, 2018.
[33] Hon, Wing-Kail; Shah, Rahul; Vitter, and S. Jeffrey, "Ordered pattern matching: Towards full-text retrieval", Department of Computer Science Tech. Rep. Paper 1651, pp.1-10, 2006.