Research Journal of Applied Sciences 6 (3): 184-193, 2011
ISSN: 1815-932X

Applying a Novel Query Reformulation Keywords Algorithm in A Mobile Healthcare Retrieval Context

[1]Kehinde Agbele, [2]Bayo Adetunmbi, [3]Segun Olajide and [4]Daniel Ekong
[1]Department of Computer Science, Soft Computing and Intelligent Systems Research Group, University of the Western Cape, Private Bag X17, 7535 Bellville, South Africa
[2]Department of Computer Science, Federal University of Technology, P.M.B. 704 Akure, Nigeria
[3]Department of Computer Science, Adeyemi College of Education, P.M.B. 520 Ondo, Nigeria
[4]Department of Mathematical Sciences, University of Ado-Ekiti, P.M.B. 5363 Ado-Ekiti, Nigeria

Abstract: Today, searching information in the web or in any kind of document collection has become one of the most regular activities. Though, user queries can be formulated in a way that hinders the recovery of the requested information. The objective of automatic query reformulation is to improve the predicted relevance of the retrieved documents. This study describes a new OPTRANDOC algorithm; an acronym for OPTmize RANking DOCuments procedure used to modify the set of keyword terms from the information sources that compose a user query with the use of a SMS based on HIV/AIDS content-related corpus. The proposed method extracts the frequency of SMS-query keyword terms that appears within the context of (FAQs) databases. This study presents a novel framework of Information Retrieval Systems (IRS). The developed OPTRANDOC procedure is used as an evaluation measure. Researchers apply TFIDF method to obtain the trial results for which the SMS corpus provides its different forms that is promising with its attractive results.

Key words: Information retrieval, information retrieval systems, ranking function, context awareness, genetic algorithm, relevance feedback, mobile information access, HIV/AIDS management

INTRODUCTION

Document retrieval is the problem of finding stored documents that contain helpful information. There exist a set of documents on a range of topics written by different researchers at different times and at varying levels of depth, detail, clarity and precision and a set of individual who at different times and for different reasons search for recorded information that may contain in some of the documents in this set. In each instance in which an individual seeks information, he or she will find some documents of the set helpful and other documents not useful. The documents find useful areresearchers say relevant otherwise not relevant.

There is a virtual explosion in the availability of electronic information. The advent of the internet or World Wide Web (WWW) has brought far more information than any human being can absorb. Furthermore with the emergent proliferation of mobile devices, users are increasingly using internet services on the go. According to searchenginewatch.com, major search engines such as Google, AltaVista and Yahoo, take delivery of millions of search request per day. This fact obviously demonstrates the significance of search engines in the daily life. The goal of Information Retrieval (IR) systems is to assist user to organize and store such information and retrieve useful information when a user submits a query to the IR systems. To resolve this problem, many research communities have implemented diverse techniques such as inverted index, keyword querying, boolean querying, knowledge-based, neural network, probabilistic retrieval, genetic algorithm and machine learning.

It is the responsibility of a user to formulate query and send the query to the search engine. Information retrieval system searches for the matches in the document databases and thus retrieves search results of the matching process. The user will then display the search results based on the relevance. The relevance of the document is very important to the user. If the user feels that it is a relevant document, he finishes the search else user continues to search in the document database by reformulating the query until the relevant documents are retrieved that will satisfy users information needs. But

Corresponding Author: Kehinde Kayode Agbele, Department of Computer Science, Soft Computing and Intelligent Systems Research Group, University of the Western Cape, Private Bag X17, 7535 Bellville, South Africa

according to Erba *et al.* (2011) the results returned by the search engine may not be relevant to the users information need and hence users need to modify and reformulate their queries. Baeza-Yates and Ribeiro-Neto (1999) stated that user is in need of information.

Context awareness is thus the ability of an entity to be aware of the surrounding situations and use the information to perform some tasks. An entity can be a person, a place or an object that is considered relevant to the interaction between a user and an application including the user and application themselves (Dey, 2001). This classification helps to understand the use of context in mobile applications. Prasannakumari (2010) develops a very simple efficient method for contextual information retrieval from multimedia databases to meet any individual user information needs. Prasannakumari method combines learning by feedback approach and improved relevant ranking to build a better database. In this regards, context information can be environmental, application or device-oriented or user-related. Based on the contextual information acquired, a mobile system reacts, adapts and responds accordingly but only within the parameters that determine the perceived context.

As discussed by Agbele *et al.* (2010) access to information has important benefit that can be achieved in many areas including social-economic development, education and healthcare. In healthcare for example, access to appropriate information can minimize visits to physicians and period of hospitalization for patients suffering from chronic conditions such as asthma, diabetes, hypertension and HIV/AIDS. Agbele method examines opening of health information system based on ICT as one fundamental healthcare application area especially within the context of the Millennium Development goals to improve the management and quality of healthcare for development at lower cost.

Adesina *et al.* (2010) used Short Messaging Service (SMS) messages as a tool in a health provision environment in different forms of communication to form a set of pre-formed questions related to HIV/AIDS. The SMS were provided for all group participant of 1st year Computer Science Department, University of the Western Cape to form the SMS-corpus. Therefore, an information retrieval system has its heart a collection database about certainty (Korfhage, 1997). Information Retrieval System (IRS) is a system used to store items of information that need to be processed, searched and retrieved corresponding to a user's query.

According to relevant literatures (Nyongesa and Maleki-Dizaji, 2006; Mauldin *et al.*, 1987; Chen *et al.*, 2010), most IRSs suffer from keywords barriers to convey

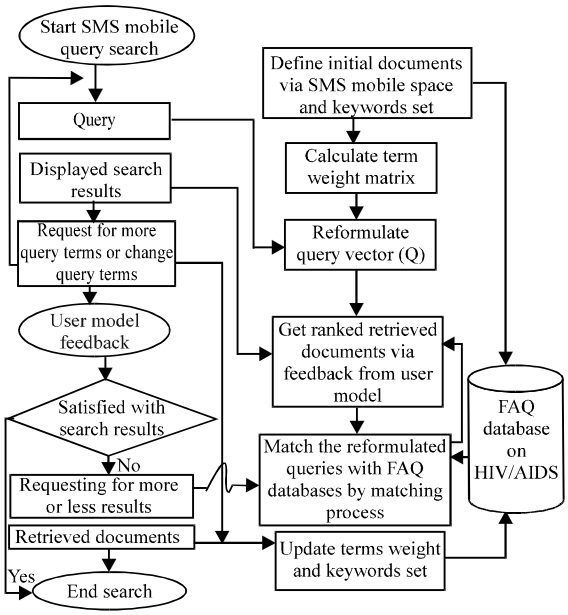

Fig. 1: Flowchart of the proposed algorithm procedures

the semantic context meaning of retrieve documents. Further, the system first extracts keyword terms by using different approaches. As a consequence such a system has two key problems; one is how to extract keywords specifically and the other is how to decide the weight of each keyword. Bani-Ahmad and Al-Dweik (2011) proposed a new term-ranking approach that give an approximation of the relative importance of the terms within the document where they are observed to improve similarity scores. Therefore, this study described a new OPTRANDOC algorithm procedure that modify the set of keyword terms via feedback from user using TF*IDF method. This method is aiming to effectively adapt SMS-query keywords weights in a health provision environment.

The proposed retrieval system incorporated frequency of keyword terms that appear in FAQ databases related to HIV/AIDS documents based on developed algorithm procedures (Fig. 1).

MATERIALS AND METHODS

Concept of the proposed OPTRANDOC algorithm procedures: Based on method 1, a weight matrix of all the documents with respect to the extracted significant keywords can be set up based on the frequency of a keyword appeared in documents. That is by consider each weight (w_{ij}) as an index calculated from the frequency of a keyword k_i across the entire documents d_j, where:

$$j = 1,..., M \text{ and } i = 1,... ,N$$

Table 1: Weight matrix of documents

	Keywords			
Documents	k_1	k_2	k_i	k_M
d_1	w_{11}	w_{12}	-	w_{1M}
d_2	w_{21}	w_{22}	-	-
d_i	-	-	w_{ij}	-
d_N	w_{N1}	-	-	w_{NM}

Then a weight matrix can be constructed as shown in Table 1. Then a query vector:

$$Q \equiv [q_j, j = 1,..., M]$$

corresponding to the keywords is defined by binary numbers of which the element $q_j = 1$ if any of the SMS-query terms given by a user matches the jth keyword in the keywords set otherwise $q_j = 0$. Researchers want to construct an Information Retrieval (IR) system using a weight-based approach on the proposed OPTRANDOC procedures; an acronym for OPTimize RANking DOCuments. The proposed procedure will extract the frequency of SMS-query keyword terms that appear in FAQs databases on HIV/AIDS content-related documents by optimizing the ranking of retrieved documents from the search engine. A new method of retrieval system to address the following issues is hereby proposed:

- The algorithm must provide the suitable amount of relevant information according to each user's request
- The algorithm must improve the ranking mechanism for the mobile search results in an attempt to adapt the retrieval environment for users
- The proposed algorithm must be self-learning that can automatically adjust its search structure to a user's query behaviour

Salton (1970) first studied weight collection by considering term frequency (tf) and inverse document frequency (idf). For searching user weight of each vector term, a weighting approach (semantic process) of FAQ document collection based on TFIDF method. tf_{ij} is defined as the number of occurrences of keyword term k_j in document d_i and idf_i defined as log (N/df_j) in which N is the total number of documents containing keyword k_j. Based on Salton model, diverse researchers further modified the formulas into various forms to adapt to their needs.

On account of consideration of trade-offs between precision and time, Salton and Buckley (1988)'s technique discussed below has been adopted as a pilot guide in this study. The resulting ranking function for a query Q and document D can be represented as made known in Eq. 1:

$$w_{ij} = \frac{(\log tf_{ij} + 1.0) * idf_j}{\sum_{j=1}^{m} [(\log tf_{ij} + 1.0) * idf_j]^2} \quad (1)$$

Where:

$$j = 1,..., M \text{ and } i = 1, ..., N$$

The weight of importance of a document will be measured according to the degree of fitness DF of the document with respect to the query vector with a small-operator defined as matrix G indicated : $G = [g_{ij}]_{N*M}$ where:

$$g_{ij} = \min (w_{ij}, q_{ij}) \quad (2)$$

Where:

$$1 \le i \le N, 1 \le j \le M$$

To retrieve the documents, a specific weight threshold (ϖ) given by the system as specified and calculated from Eq. 3. Hence, there is defined the average weight value as Eq. 3 and will be taken as a default weight threshold:

$$\varpi = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} w_{ij}}{M * N} \quad (3)$$

where, $w_{ij} > 0$ (default weight threshold). Therefore, any weight element of matrix G greater than the threshold will be reserved to add to a matrix T as made known in Eq. 4:

$$T = [t_{ij}]_{N*M} \quad (4)$$

Where:

$$\begin{cases} t_{ij} = g_{ij}, if g_{ij} \ge \varpi \\ t_{ij} = 0, if g_{ij} \prec \varpi \end{cases} \quad 1 \le i \le N, 1 \le j \le M$$

In this regards based on the matrix T the system will calculate scores, sco_i of all documents which are as the largest weighting value of each corresponding vector as Eq. 5:

$$Sco_i = \max \{t_{ij}\}, 1 \le i \le N, 1 \le j \le m \quad (5)$$

Document d_i is retrieved if $sco_i > 0$ and added into the retrieved document set, D shown in Eq. 6. After a query is issued by a user, the system will rank the retrieved documents in order of Sco_i and will recommend in such other (In general, there is assume the utility of a relevant documents decreases with its ranking order):

$$D = \{d_i | \text{ if } Sco_i > 0, 1 \le i \le N\} \quad (6)$$

To provide the suitable amount of relevant information based on each user's needs; different information is varied from different users. In this regards, there is introduce parameter α's which will allow user to retrieve the desired amount of documents by the following operations. If more documents are desired, the

user can press more button. To enlarge the number of nonzero value in matrix T, the system will decrease the weight threshold by setting a value $\alpha>1$ to adjust ϖ by:

$$\varpi = \left(\varpi^{\alpha}\right) \quad (7)$$

The larger is α, the larger will be the reduced range of the threshold and the more are the increase amounts of retrieved documents. Set the default value. If less amount of documents are expected especially when users does not have much time to filter large amounts of documents, the user can press less button. For this purpose, the system will increase the weight threshold setting $0<\alpha<1$ in Eq. 7. The smaller is α, the larger will be the incremental range of threshold and the more are decreased amounts of retrieved documents. Set the default value.

The search engine is to provide automatic adjustment mechanism for the users based on their preference and for suitable amount of information to be retrieved as set condition in the proposed algorithm. The keyword set K provided by the documents and the weight values will be updated by the feedback of the users. Any new query term not belonging to K will be added and a new column of weight value will be computer and expanded for documents automatically. If any retrieved document d_i is retrieved by the users, the corresponding weight values with respect to the query keywords will be increased by Eq. 8. The default is set to increase the corresponding weight values:

$$w_{ij} = \left(w_{ij}\right)^{\beta}$$

Where:

$$0<\beta<1,\ i \in \{i|di \in D\} \text{ and } j \in \{j|q_j = 1\} \quad (8)$$

The search structure is to be modified if more users make queries and retrieved the documents. This will address the issue of self-learning set condition in the proposed algorithm. If a user gives the same query term, the rank of the previous retrieved documents will be raised to achieve the goal of ranking improvement as specified in the user preference function. This will address the issue of ranking mechanism set condition in the proposed algorithm. However, the new SMS-query keyword terms are selected from the SMS-Corpus as a user-defined parameter. The representation of the query vectors will be of the form:

Vector 1: {(AIDS, 0.301) (Treatments, 0.120), ...}

Vector 2: {(HIV, 0.397) (Symptoms, 0.099), ...}

Hence, user query reformulation applies by updating its profile. A user profile or model is a stored knowledge about a particular user. Simple model consists usually of keywords describing user's area of interest in context In this regards, weigh of ith user of jth vector is given by the following procedure: W_{ij} = tf * idf where tf = $freq_{i,j}$/max keywordcnt and idf = log (N/n_k). Hence, $w_{i,j}$ = ($freq_{i,j}$/max keywordcnt) * Log (N/n_k). Where $freq_{i,j}$ is the frequency of the Kith user in Djth query; max keywordcnt is the maximum keywordcnt of all keywords in the FAQ database; N is the total number of documents in the entire FAQ database collection and n_k is the total count of keywords in the entire FAQ database collection.

Issues to be resolved by the concept: The ability of the search engines to return useful and relevant documents is not always satisfactory. Often users need to refine the search query several times and search through large document collections to find relevant information. In this regards, these issues have been discussed in literature with the thought of using optimization techniques according to Glover *et al.* (1999, 2001). However, the necessary amount of relevant information is varied from diverse users. According to Erba *et al.* (2011), explored explicit relevance feedback to measure the variability in judgements and behaviour for the same query from individual users for ranking. The explicit relevance feedbacks give room to observe the consistency in relevance assessments across different individual users. The major challenge of this study includes how to gather satisfactory data and it is burdensome for users to provide explicit judgements. Thus, how to provide suitable amount of relevant information according to individual user information needs is what to be addressed in this study.

It is important to lay emphasis on how to improve the ranking mechanism for the searching results of FAQ on HIV/AIDS content-related documents from the search engine. According to satisfying, the users' preference, genetic algorithms have been helpful by many researchers to improve the search queries (Salton and Buckley, 1988; Yang and Korfhage, 1994). Though, their systems failed to offer a satisfactory evaluation to score and rank the retrieved information constantly.

As discussed by Lin and Wang (2006), Billerbeck *et al.* (2003) and Kim *et al.* (2001) query expansion afforded system users with relevant results from online users' feedback. However, highlighted below are the major flaws for these methods:

- Their system reformulate processes require users' additional preference based on the previous retrieved result
- Their system cannot make use of users' query experience to help the new users

- The existing search systems cannot change the search structure whenever a user takes some actions for instance, retrieving a correct relevant documents

Thus, self-learning search engine (or IRS) that can automatically adjust its search structure to user query behaviour is both valuable and essential. Hoque and Avery (2010), proposed and designed concept that support faster query execution. The results perform quicker and efficient having both time and space complexities reduced considerably. In this study, a new method is proposed based on the three issues evaluated from the existing IRS of effectiveness, ranking mechanism and self-adjustment to the users to improve mobile retrieval performance results in a health provision environment.

The proposed optrandoc algorithm procedures: Based on the promise concepts described in the study, this study review the proposed procedure with the evaluation of the OPRANDOC algorithm procedures effectiveness is shown by a demonstrated example.

The OPTRANDOC algorithm: The OPTRANDOC algorithm is described below with its flowchart.

Stage 1
(Initialization of SMS query):

- Set the initial document set, $D_0 = \{d_1, d_2, d_3 \ldots, d_N\}$ and obtain the initial query keywords set, $K = \{k_1, k_2, k_3..., k_M\}$
- Define a set B with the features of the documents as $B = \{B_1, B_2\}$ where, B_1 is the publishing year and B_2 is the properties of documents including journal, thesis, conference, seminar study, patent, textbooks, technical reports

Stage 2
(Calculate term weights matrix from the FAQ database):

- Calculate the term frequency (tf)
  - For each k_j in K
  - For each d_i in D_0
  - Find the number tf_{ij} = ($freq_{i,j}$/max keywordcnt)
- Calculate the inverse document frequency (idf)
  - For k_j = 1-M
  - put $n_j = 0$
  - For d_i = 1-N
  - If $tf_{ij}>0$ then $n_j = n_j+1$
  - Find the number idf = log (N/n_j)
- Get term set W_{ij}
  - For k_j = 1-M
  - For d_i = 1-N
  - Calculate W_{ij} = (tf * idf) as Eq. 1
- Calculate the threshold ϖ as Eq. 3

Stage 3
Reformulate a query:

- If a user selects the features of B, filter N documents by B_1 and B_2 and obtain N documents
- Define query vector Q
  - For each k_j in K
  - If (K_J matches the query terms) Then $q_j = 1$
  - Else $q_j = 0$

Stage 4
Get feedback via user model from documents to be retrieved:

- Create matrix G
  - For i = 1-N
  - For j = 1-M
  - g_{ij}= min (w_{ij}, q_{ij})
- Create matrix T by the weight threshold ϖ
  - For i = 1-N
  - For j = 1-M
  - If g_{ij}> = w then $g_{ij} = f_{ij}$
  - Else $t_{ij} = 0$
- Calculate the scores and generate D for the sets of retrieved documents via SMS mobile query search
  - For d_i = 1-N
  - Sco_i = max (default value; t_{ij}) for j = 1-M
  - If $sco_i>0$ then
  - Add d_i into D
- Display the two sets of retrieved documents according to the rank of the related scores i.e., Retrieval Status Value (RSV)
  - For d_i in D
  - Sort sco_i and display results

Stage 5
(Match the reformulated queries with FAQ documents):

- If a user is requesting for more documents, the system will decrease ϖ (go to stage 4)
- Else if a user is requesting for less documents is for increasing ϖ in the system (go to stage 4)
- Else if a user wants to reformulate query or stop querying (go to stage 6)

Stage 6
(Update term weights and keywords set):

- Update term weights (tf and idf) values
  - For d_i = 1-N and $d_i \in D$
  - If d_i is retrieved
  - For k_j = 1- M and $q_j = 1$
  - Update w_{ij} = (tf * idf)
- Update keywords set, K
  - For any query term q_k not in K then
  - Add q_k into K
  - For d_i = 1-N, $k_j = M + 1$

- Calculate w_{ij} = (tf * idf) as Eq. 8
- If user want to reformulate query then (go to stage 3); Else, stop

Testing the validity of the proposed OPTRANDOC algorithm using TFIDF method: This study describes the effectiveness of optimizing ranking terms via OPTRANDOC procedure including 2 FAQ databases and extracted SMS-query keywords set on HIV/AIDS content-related documents using TF*IDF method to demonstrate the proposed OPTRANDOC procedure effectiveness.

Stage 1: The initial SMS-query keywords were first collected into the set K = {HIV, AIDS, monitoring, symptoms, awareness, medication, eligibility, criteria, reminder, treatments} in the initial stage.

Stage 2: The number of each keyword term occurred in each FAQ database was counted as keyword frequency and listed as shown in Table 2. Thus, $w_{i,j}$ = ($freq_{i,j}$/total wordcnt) * Log (N/n_k) based on TF*IDF approach, the term weight set can be obtained as shown in Table 3 below. Hint: N = 2, n_k = 32, d_1 total keywordcnt = 20, d_2 total keywordcnt = 12 and $freq_{i,j}$ is the frequency of Kith user in Djth query.

Stage 3: A particular user model may have numerous and different query request of varying interests. In this regards, this study apply context awareness to reformulate queries in order to improve the predicted relevance of the retrieved documents on mobile devices. Suppose a user makes SMS-query including two keywords AIDS and treatments. Hence, the query vector will be Q = [0, 1, 0, 0, 0, 0, 0, 0, 0, 1] and the equivalent extended SMS-query weight set is obtained as shown in Table 4. 1 being relevant and 0 being non-relevant.

Stage 4: Therefore, the retrieved query keywords set is K = {k_2 and k_{10}} via feedback from user model or human assessment about FAQ on HIV/AIDS content-related documents.

Stage 5: At this stage if a particular user wants more SMS-query keywords set from FAQ documents and select the more function to include the following two terms, HIV and AIDS, HIV and awareness, medication and reminder, HIV and monitoring. Therefore, Table 5 shows the retrieved query keywords set for K = {k_1, k_2, k_3, k_5, k_6, k_9, k_{10}} where, k_1, k_3, k_5, k_6 and k_9 are added SMS-query keywords. On the contrary, a particular user who wants less SMS-query keywords from FAQ documents can select the less function to include the following two terms AIDS and symptoms, HIV and eligibility. Therefore, Table 6 shows the retrieved query keywords set for K = {k_1, k_2, k_4, k_7} where k_{10} is eliminated from the SMS-query keywords set.

Stage 6: After the reformulated queries if the user retrieves the 2 FAQ documents d_1 and d_2 then the information retrieval system will update (increase) the term weight value according to the keywords provided by d_1 and d_2. That is $w_{1,2}$, $w_{1,10}$, $w_{2,2}$ and $w_{2,10}$ will be recomputed as stated in the proposed algorithm formula and the updated term weight matrix is obtained as shown in Table 7.

Table 2: Extracted significant keywords in each FAQ document

Document	HIV	AIDS	Monitoring	Symptoms	Awareness	Medication	Eligibility	Criteria	Reminder	Treatments
d_1	7	5	1	1	0	2	1	0	1	2
d_2	4	3	0	1	1	0	0	0	2	1

Table 3: Term weight set

Document	HIV	AIDS	Monitoring	Symptoms	Awareness	Medication	Eligibility	Criteria	Reminder	Treatments
d_1	0.421	0.301	0.006	0.012	0.000	0.120	0.060	0.000	0.060	0.120
d_2	0.397	0.301	0.000	0.099	0.100	0.000	0.000	0.000	0.201	0.100

Table 4: SMS-query keyword extension set

Document	HIV	AID	Smonitoring	Symptoms	Awareness	Medication	Eligibility	Criteria	Reminder	Treatments
d_1	0	1	0	0	0	0	0	0	0	1
d_2	0	1	0	0	0	0	0	0	0	1

Table 5: More retrieved SMS-query keywords

Document	HIV	AIDS	Monitoring	Symptoms	Awareness	Medication	Eligibility	Criteria	Reminder	Treatments
d_1	1	1	1	0	1	1	0	0	1	1
d_2	1	1	1	0	1	1	0	0	1	1

Table 6: Less retrieved SMS-query keywords

Document	HIV	AIDS	Monitoring	Symptoms	Awareness	Medication	Eligibility	Criteria	Reminder	Treatments
d_1	1	1	0	1	0	0	1	0	0	0
d_2	1	1	0	1	0	0	1	0	0	0

Table 7: Updated term weight set

Document	HIV	AIDS	Monitoring	Symptoms	Awareness	Medication	Eligibility	Criteria	Reminder	Treatment
d_1	0.421	0.344	0.006	0.012	0.000	0.120	0.060	0.000	0.060	0.172
d_2	0.397	0.370	0.000	0.099	0.100	0.000	0.000	0.000	0.201	0.185

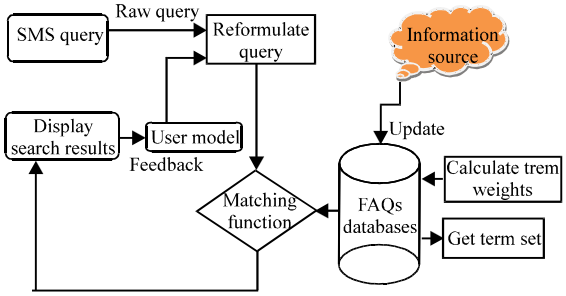

Fig. 2: Information retrieval system-proposed framework

Therefore, researchers then found that the term weight values of the keywords, AIDS and treatments, corresponding to d_1 and d_2 have increased from 0.301-0.344 and 0.120-0.172, respectively for d_1 and 0.301-0.370 and 0.100-0.185, respectively for d_2 which satisfies the conditions on stage 6 of the OPRANDOC algorithm procedure. However, if the user then again makes an SMS-query using the same keywords then the ranking of d_1 and d_2 could be changed accordingly by optimizing the ranking of retrieved documents from the search engine.

An information retrieval system-a proposed framework based on the developed optrandoc algorithm: In the proposed framework for information retrieval as showed in Fig. 2, user gives a mobile SMS-query (Raw query) and the query is reformulated in order to improve the predicted relevance of the retrieved document. The reformulated query is searched against the databases. The proposed retrieval system incorporates the frequency of keyword terms that appear in FAQs databases related to HIV/AIDS content related-documents using term weighting TFIDF method by optimizing the ranking order of retrieved documents from the search engine. The Information retrieval system searches for the matches in the document databases and thus retrieves search results of the matching process.

Based on the relevance, the user will then display the search results. The relevance of the document is very important to the user. If the user feels that it is a relevant document, he finishes the search else user continues to search in the document database by reformulating the query until the relevant documents are retrieved that will satisfy users' information needs. Hence, user query reformulations will apply by updating its model. A user model is a stored knowledge about a particular user. Simple model consists usually of keywords describing user's area of interest. Sort those documents according to TFIDF method. The documents which have the high Retrieval Status Value (RSV) are considered as the top ranked documents.

The two main components in the proposed information retrieval system framework are Document Databases and Reformulated Query Processing System. The Document Databases stores the databases related to documents and the representations of their information contents based on TFIDF method. A SMS-query keyword term is also associated with this component which automatically generates a representation for each document by extracting the frequency of the SMS-query keyword terms from the document contents. The reformulated query processing system consists of two subsystems: Searching-Matching unit and Displaying-Ranking unit.

Searching unit allows user to search the documents from the document database and matching unit does a comparison of all documents against the user's query. To improve the predicted relevance of the retrieved document, the reformulated query is searched against the databases. Searching-Matching unit does a thorough search and finds out which documents match the user query. This unit retrieves almost all the documents that match either part or whole of the entire query that is the unit retrieves relevant amid non relevant documents.

Displaying unit displays the search results based on relevance of the documents to user information needs and ranking unit ranks the document according to the relevance of the user query. Displaying-Ranking unit does a detailed display of search results and find out which documents have high RSV are considered as the top ranked documents. Therefore, Information Retrieval (IR) system ranks the documents according to the RSV between document and the query. If a document has got high RSV that document is closer to the query. In other words the document is relevant to the query.

Generally IR system ranks the list of documents in the descending order. After processing the query effectively, the top most relevant documents are retrieved and it is given to the user. Though, relevance feedback is one of the processes in an information retrieval system that seeks to improve the system's performance based on a user's feedback. It modifies queries using judgments of the relevance of a few, highly-ranked documents and has historically been an important method for increasing the performance of information retrieval systems.

Specifically, the user's judgments of the relevance or non-relevance of some of the documents retrieved are used to add new terms to the query and to reweight query

terms. For example if all the documents that the user judges as relevant contain a particular term then that term may be a good one to add to the original query. It is made known by Salton (1970) that relevance feedback has improved the system's overall performance by 60-170% for different document collections. Given the apparent effectiveness of relevance feedback techniques, it is important that any proposed model of information retrieval include these techniques. In the proposed system rather than modifying the matching function, researchers will modify the query vector using genetic algorithm to adapt the query vectors and to reflect a user's feedback about relevance.

A generalized genetic model approach for information retrieval: Information Retrieval (IR) is concerned primarily with finding and returning information stored in computers that is relevant to a user's needs (query). With the advent of the Internet, IR has acquired remarkable practical impact as well as theoretical importance. Therefore, IR is an area committed to the management of large collections of information and to the retrieval of helpful information from the documents collection for users. Baeza-Yates and Ribeiro-Neto (1999), tries to provide the user with easy access to the document databases that will satisfy his information needs and which is relevant in a suitable time interval. IR may be defined, in general as the problem of selection of information from documents collection (storage) in response to search queries issued by a user.

Therefore, the goal of an IRS is to estimate the relevance of information needs to a user's information expressed in a query. In this regards, IRS process user queries trying to allow the user to access relevant information in a suitable time interval. According to (Bookstein, 1983), states that basically, people want an IRS to provide lists of documents ordered by the importance of documents to them. Relevant documents are to be ranked first than the non-relevant documents so that the user need not search the entire documents collection to look for documents of interest. In general, documents with higher similarity to query are judge more relevant to the query and should be retrieved first.

The information retrieval approach shown in Fig. 3 can be applied to Genetic algorithm for retrieving information. In this case, query and documents are represented as gene (chromosome). An initial population of query is generated. The query is sent to the IRS (search engine), a matching process is carried out between the query chromosome and the document chromosome and then the document is considered as relevant. If non-relevant documents are found then the query is reformulated. The query is reformulated until a relevant document is retrieved. Besides, query will be seen as a vector of a user in the context of the query. The search vectors will be expressed as keyword terms against its associated term weights value. The key primary concern in representation is how to select proper keyword terms. Hence, representation commences by extracting the keyword terms that are considered as content identifiers and classifying them into the arrangement comprised of pairs (k, w) where k is a keyword term and w is the term's weight.

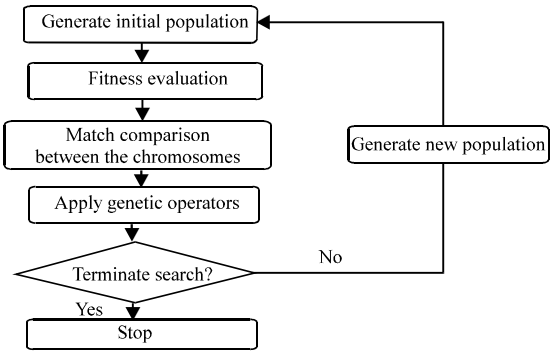

Fig. 3: A genetic algorithm approach to information retrieval

RESULTS AND DISCUSSION

Experimental evaluation performance results: To evaluate the performance of the OPTRANDOC algorithm procedures based on TF*IDF method shown in Fig. 4, the satisfactory level of the user were evaluated in Offline mode. While there is no information for analysis on precision and recall, testing system's effectiveness by self satisfaction was an alternative way adopted to include how relevant is the retrieved documents? And is the user satisfies with the function of adding personal new SMS-query keywords? To address the issue of self-learning, the keywords set k provided by the documents and its weight values will be updated by the feedback from human assessment. This can be achieved by; first any new SMS-query keyword not belonging to k will be added and a new column of weight will be computed and expanded for document automatically as demonstrated in the OPTRANDOC procedures (Table 7). Then if any

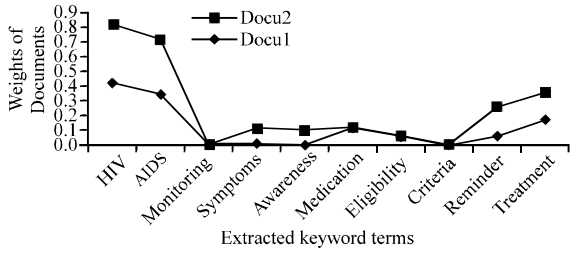

Fig. 4: Graph of experimental retrieval results of AIDS and treatment (Table 7)

retrieved document d_1 or d_2 is retrieved by the users, the corresponding weight values with respect to the mobile SMS-query keywords will be increased as shown in the OPTRANDOC procedures (Table 7). Hence, the search structure is modified if more users make SMS-queries and retrieve the document. As a consequence, the issue of self-learning is resolved. Also, if a user gives the same SMS-query keywords, the rank of the previous retrieved document will be raised to achieve the goal of ranking improvement and solve the issue of improving ranking mechanism. Finally, the information source will provide an automatic adjustment for the users based on their preferences which allow users to retrieve the relevant amount of documents as specified in stage 5 of the OPTRANDOC procedure. Thus, the issue of effectiveness is resolved.

CONCLUSION

In this study, a new method is proposed based on three issues evaluated from the existing systems of effectiveness, self adjustment and improving ranking mechanism to the users. Subsequently, this study has outlined a flowchart for analyzing, understanding and investigating extracted SMS-query keyword terms based on mobile information access in a healthcare provision environment. The effectiveness of the system performance was evaluated numerically based on the self satisfaction of the relevance feedback from the users' using TF*IDF method. The algorithm has demonstrated the ability of providing satisfactory functions for users to retrieve the relevant information according to user's preference for ranking. The subsequent task of this research will focus on how Genetic Algorithms (GA) can be adopted as an approach to get optimal or near optimal solutions (Goldberg, 1989; Holland, 1975) in online mode using Java-script for implementation. The retrieval effectiveness will be evaluated in terms of recall and precision measurements and the proposed IRS is allied to mobile healthcare information access.

Though, this research project is at development and implementation stage. It is the strong belief that the full implementation and evaluation of the proposed information retrieval systems will assist users in documents ranking order according to their relevance. Therefore, HIV/AIDS content-related documents with higher similarity query are to be judged more relevant to the SMS-query keyword terms and should be retrieved first using genetic algorithm to adapt the query vectors via feedback of the users. This will in turn help HIV/AIDS managements and lower the cost of healthcare provision.

ACKNOWLEDGEMENTS

Researchers wish to thank Prof. Henry Nyongesa for his kind and helpful discussions and comments about the OPTRANDOC algorithm procedures. Also, researchers appreciate cooperative remarks from the colleagues in the research group.

REFERENCES

Adesina, A., K. Agbele and H. Nyongesa, 2010. Text messaging: A tool in E-health services. Proceedings of the SATNAC 2010, Sept. 5-8, Stellenbosch, South Africa.

Agbele, K., H. Nyongesa and A. Adesina, 2010. ICT and information security perspectives in E-health systems. J. Mobile Commun., 4: 17-22.

Baeza-Yates, R. and B. Ribeiro-Neto, 1999. Modern Information Retrieval. Addison Wesley, London.

Bani-Ahmad, S. and G. Al-Dweik, 2011. A new term-ranking approach that supports improved searching in literature digital libraries. Res. J. Inform. Technol., 3: 44-52.

Billerbeck, B., F. Scholer, H.E. Williams and J. Zobel, 2003. Query expansion using associated queries. Proceedings of the 12th International Conference on Information and Knowledge Management, Nov. 3-8, New Orleans, LA. USA., pp: 2-9.

Bookstein, A., 1983. Outline of a general probabilistic retrieval model. J. Document., 39: 63-72.

Chen, M.Y., H.C. Chu and Y.M. Chen, 2010. Developing a semantic-enable information retrieval mechanism. Expert Syst. Applic., 37: 332-340.

Dey, A.K., 2001. Understanding and using context. Personal Ubiquitous Comput., 5: 4-7.

Erba, F.G., Z. Yu and L. Ting, 2011. Using explicit measures to quantify the potential for personalizing search. Res. J. Inform. Technol., 3: 24-34.

Glover, E.J., S. Lawrence, M.D. Gordon, W.P. Birmingham and C.L. Giles, 1999. Recommending web documents based on user preferences. SIGIR 99 Workshop on Recommender Systems. Berkeley.

Glover, E.J., S. Lawrence, M.D. Gordon, W.P. Birmingham and C.L. Giles, 2001. Web search-your way. Commun. ACM, 44: 97-102.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning. 1st Edn., Addison-Wesley Publishing Company, New York, USA., ISBN: 0201157675, pp: 36-90.

Holland, J.H., 1975. Adaptation in Natural and Artificial Systems. 1st Edn., University of Michigan Press, Ann Arbor, Michigan, ISBN: 0472084607.

Hoque, M.T. and V.M. Avery, 2010. Novel strategies to speed-up query response. Res. J. Inform. Technol., 2: 11-20.

Kim, B.M., J.Y. Kim and J. Kim, 2001. Query term expansion and reweighting using term co-occurrence similarity and fuzzy inference. Proceedings of the Joint 9th IFSA World Congress and 20th NAFIPS International Conference, July 25-28, Vancouver, BC Canada, pp: 715-720.

Korfhage, R., 1997. Information Storage and Retrieval. John Wiley and Sons, USA., ISBN-10: 0471143383, pp: 368.

Lin, H.C. and L.H. Wang, 2006. Query expansion for document retrieval based on fuzzy rules and user relevance feedback techniques. Expert Systems with Appli., 31: 397-405.

Mauldin, M., J. Carbonell and R. Thomason, 1987. Beyond the keyword barrier: Knowledge-based information retrieval. Inform. Services Use, 7: 103-117.

Nyongesa, H.O. and S. Maleki-Dizaji, 2006. User modelling using evolutionary interactive reinforcement learning. Inform. Retrieval, 9: 343-355.

Prasannakumari, V., 2010. Contextual information retrieval for multi-media databases with learning by feedback using vector space model. Asian J. Inform. Manage., 4: 12-18.

Salton, G. and C. Buckley, 1988. Term-weighting approaches in automatic text retrieval. J. Inform. Proc. Manage., 24: 513-523.

Salton, G., 1970. Automatic text analysis. Science, 168: 335-342.

Yang, J. and R. Korfhage, 1994. Query modifications using genetic algorithms in vector space models. Int. J. Expert Syst., 7: 165-191.