ISSN 2278-3091

Volume 9, No.1.1, 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse6091.12020.pdf

https://doi.org/10.30534/ijatcse/2020/6091.12020

An Enhanced Concept Based Approach for User Centered Health Information Retrieval to Address Medical and Vocabulary Mismatched Issues

Ibrahim Umar Kontagora[1, 2], **Isredza Rahmi A. Hamid**[1], **Nurul Aswa Omar**[1]

[1]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn, Malaysia.
[2]Department of Computer Science, Niger State Polytechnic, Zungeru, Niger State, Nigeria.
*ibrosoftuk@yahoo.com[1,2], rahmi@uthm.edu.my[1], nurulaswa@uthm.edu.my[1]

ABSTRACT

With the huge growth in health related information online, user centered health information retrieval has become an obli-gation today. In spite of all its numerous benefits, laymen patients are faced with the challenge of finding relevant and authoritative information online due to vocabulary mismatched issues between laymen queries and expert vocabulary during query expansion. However, this study is focused on designing an enhanced concept based approach for user cen-tered health information retrieval that will address vocabulary mismatched issues. The Proposed Enhanced Concept Based Approach has two special modules fully designed and incorporated in it namely: vocabulary controlled module and the medical terms control module to specifically address vocabulary mismatched issues by ensuring that only the most specific terms in a layman's search query are searched for and their synonyms extracted from the dictionary of the selected dataset. We presented and discussed the performance of the proposed enhanced concept based approach with the existing information retrieval approaches. We concentrated on comparing and evaluating the proposed enhanced concept based approach with three existing approaches used in health information retrieval which are: Concept Based Approach, Query Likelihood Model and the Latent Semantic Indexing. The experimental results obtained using the Met-amap, MeSH, UMLS and Khresmoi Project 6 datasets shows that the Proposed Enhanced Concept Based Approach is the best approach to be applied in addressing vocabulary mismatched issues as it managed to obtain best results in re-spect to maxSim scores and idf weighting values using the Text Similarity Scoring Function with maxSim scored 0.89 (89%) and idf weighting value of 3.61.The Proposed Enhanced Concept Based Approach outshines the existing ap-proaches: Concept Based Approach, Query Likelihood Model and Latent Semantic Indexing with scores ranging be-tween 11% to 15% in respect to maxSim scores and idf weighting values. This better results were obtained by the en-hance concept based approach due to the fact that, it ensures that all the medical terms contained in both the layman's queries and expert vocabularies are vocabulary matched and same. Hence we conclude that the Proposed Enhanced Con-cept Approach is best suited to be deployed in addressing the Vocabulary Mismatched issues encountered during query expansion.

Key words : Medical Discharge Reports, Clinical Reports, Concept Based Approach, Query Likelihood Model, Latent Semantic Indexing

1. INTRODUCTION

Searching of health related information from a variety of biomedical data is a domain-specific information retrieval task [4]. The benefiting category of this task comprises of a huge community of medical professional, medical attendants, patients and their relatives, researchers and anyone with accurate health information related needs [11]. Thus, it has become inevitable to design user centered health information retrieval systems that provides the users with readable and precise information as preferred by laymen patients [7], [14]. Information retrieval performance is affected due to vocabulary mismatched issues between laymen queries and expert vocabularies during query expansion and that has prevented laymen patients from finding relevant and authoritative information [8], [7]. However, the failure of the current health information retrieval systems to integrate controlled vocabularies during query expansion and also not focusing more on labeled most specific medical terms also hinders information retrieval performance [9], [8]. Previous researches had targeted more on searching and retrieving general terms rather than most specific terms in a layman query and also not restricting the synonyms search to the extracted most specific terms from the layman's query which affects systems performance [1], [2].

[24] focused on adding more lights on the relationship between healthcare and information communication technology (ICT). It also aimed at identifying the most effective and popular ICT based healthcare technology that can address the challenges faced by healthcare systems today. However, [25] focused on how both clinical staffs and

patients can understand, access and author electronic health information systems on multilingual systems. Labeled concept terms on the search query were given more attention. [23] concentrated on how precision medicine could be employed to determine the prevention, treatment of diseases, management and pathology. It also stressed that the success of any user centered health information retrieval in improving diseases awareness, cancer diseases inclusive is by having a background knowledge about the genetic variation on the illness processes.

User centered health information relates to individual patients with the aim of presenting health professionals with information about the health status of a patient [14], [12], [27]. It typically involves the patient's medical history, present diagnostics and prescribed drugs record based on the specific format used [5], [14]. These records normally contain unstructured data or semi-structured data due to a combination of computer generated and free or narrative results [4]. When queries fails to reflect users' specific information needs due to the non-labeling of most specific concept terms, it leads to results that do not address their information needs [3], [17], [7].

The main objective of this study is to design an enhanced concept based approach for user centred health information retrieval that would address vocabulary mismatched issues that exist between laymen queries and expert vocabularies during query expansion, which affects the system performances and prevents laymen patients from finding relevant and authoritative information. However, the proposed enhanced concept based approach is expected to return relevant documents with no differences between the laymen queries and expert vocabularies. The remaining parts of the paper is organized as follows: Section 2 contains related work concerning Information retrieval approaches. Section 3 deliberates on the proposed new algorithm and its implementation, Section 4 describes performance analysis and Section 5 concludes the work and gives a direction for future work.

2. RELATED WORK

Mining of health related information for health advice has now become a domain-specific information retrieval task [17], [4]. The large community of medical practitioners, clinical attendants, researchers and patients are usually the beneficent of this task [5], [10], [16], [27]. Thus, it has become obvious to design user centered health information retrieval systems that provides the users with vocabulary mismatched free information as preferred in real time [8], [1].

The essence of restricting the search for medical terms on labeled most specific terms is to allow easy access to relevant and authoritative information, and also to prevent medical terms in a laymen queries not matching with their equivalent extracted synonyms terms from the selected dataset during query expansion, which prevents laymen patients from finding relevant and authoritative information [1], [14], [8]. In the medical domain, querying the internet for useful information has become increasingly important owing to the huge amount of information available [11], [12], [14]. The outcome of previous campaigns on eHealth which focusses on addressing the differences that occur between medical terms in laymen queries and the extracted synonyms terms (expert vocabularies) had clearly shown that the task is challenging with a room for enhancement [10], [8], [16].

[22], [27] used the pre-marketing clinical trials and traditional post-marketing surveillance using voluntary and spontaneous report systems to determine the adverse drug reactions (ADRs) which has now become a major health challenge. Reports also shows that it causes more deaths in several parts of the world [12]. However, [18] focused mainly on how participatory design between the system users and stakeholders in the healthcare sectors have revolutionized the healthcare technologies and yet is not fully adopted due to the system performance being hindered by similarities issues between laymen queries and expert vocabularies. Think Aloud Protocol in the evaluation of users' involvement as a key vital tool to the success of the design and implementation of a medical collaboration and communication platform that aimed at improving medical care on critical hospitalized patients by team of clinicians [21].

The inability of the existing health information retrieval systems to fully incorporate controlled vocabularies during query expansion and focus only on labeled most specific medical terms affected their performance, as it prevents laymen patients from accessing authoritative information [8], [6], [9]. However, [8] proposed to focus more on controlled vocabularies and most specific medical terms rather than general terms which has affected information retrieval systems performance. Information retrieval performance is also affected due to vocabulary mismatched issues between laymen queries and expert vocabularies during query expansion which prevents laymen patients from finding relevant and authoritative information [7], [9], [8].

The Participatory Design technique was broadly applied as a means of enhancing information retrieval performance through interactive system users to stakeholders in the healthcare sector relationship [18], [27]. There are numerous factors that distresses query expansion results namely: re-weighting method, term selection and finally source of expansion which prevents end-users from accessing readable and authoritative information [19]. Several patients and their care givers still ask questions after reading their medical documents due to high rate of medical terms found in it [17]. Generally, search engines are widely used to extract more explanations related to specific medical terms, the problems of readability and vocabulary mismatched should be better addressed for the system to perform effectively [20].

The work by [26] used the Patient Centered Personal Health System Approach which aimed at identifying the factors that hinders the adoption of a patient-centered personal health records systems which when fully implemented would promote evidence-based decision-making and information sharing among clinicians and patients. The study also proposed a number of new approaches that could promote its adoption which tends to address the readability and vocabulary mismatched issues encountered by patients and their relatives. The outcome of the study resulted in the development of two prototype systems namely: My Clinical Record System (MCRC) that organizes, manages, stores, shares and retrieves personal health records in good time and the second which is Health Decision Support System (HDSS) that helps end-users to use SNOMED CT Codes and likely diseases as a diagnostic results.

3. THE PROPOSED ENHANCED CONCEPT BASED APPROACH

We proposed an Enhanced Concept Based Approach that would address the vocabulary mismatched issues between laymen queries and expert vocabularies during query expansion which prevents laymen patients from finding relevant and authoritative information. The proposed enhanced concept based approach is designed to fully incorporate two special modules namely: the Vocabulary Controlled Module and Medical Terms Control Modules. These two modules would ensure that only the most specific medical terms in the queries are found, labeled and extracted. And their identical terms extracted from the dictionary of the selected dataset. The functions of these two special modules would be fully incorporated to avoid vocabulary mismatched issues.

For the specific purpose of addressing the vocabulary mismatched issues encountered by end-users, the proposed enhanced concept based approach would be implemented in two phases. Firstly by fully implementing the two modules: the medical terms control module and the vocabulary controlled module incorporated in the proposed enhanced concept based approach to address vocabulary differences between laymen queries and expert vocabularies on retrieved documents. And secondly by severely ensuring that only the synonyms of the most specific medical terms are searched for, extracted and fully implemented in the new extended query.

The implementation strategy for the proposed enhanced approach is as shown in Fig 1 where $N\ is$ the number of concepts derived from the original query, C is concept terms and K are the expansion terms. SQ Represents search query and CC_n represents the nth concept, K denotes the number of expansion terms, and ET_k represents the kth expansion terms. The kth is the last expansion term in an expansion query and the symbol # represents space character (i.e., 0x20), and the double quotation marks indicate that the string in it must appear consecutively.

```
1  Input $Search\ Information$
2  [Medical Terms Control Module]
3  For all concepts $n \in [1, N]$ do
4  Set SQ = query
5  $SQ$ = “$SQ$” + “$CC_n$”
6  [Vocabulary Controlled Module]
7  For all the expansion terms $k \in [1, K]$, do
8  New $SQ$ = $SQ$ # “$ET_k$”
9  End;
10 Display Results
11 End
```

Figure 1: The Enhanced Concept Based Approach

3.1. The Medical Terms Controlled Module

This module is designed to ensure that only the most specific medical terms in a search query are searched for, extracted, and extended to the vocabulary controlled module for onward searching of their identical terms from the dictionary of the selected dataset used. Hence, it will reduce the rate of similarities issues that occur between medical terms contained in two text segments, that is T1 (Input Segment) and T2 (Output Segment). It comprises of lines no. 12 to 19 as shown in Figure 2.

```
12 Declare: getPatient-Terms (MST)
13 Search = Select “Only Most Specific Terms” in
   SearchQR.
14 If Terms = @Terms (Most Specific Terms) then
15 For all Terms (Most Specific Terms) n $\in$ [1, $N$] do
16 SearchQR = SearchQR + CCn
17 Fetch = Search (For Most Specific Terms)
18 Expand = Move “All extracted Most Specific
   Terms to Vocabulary Controlled Module”
19 End.
```

Figure 2: The Medical Terms Controlled Module

3.2. The Vocabulary Controlled Module

This module is designed and incorporated in the proposed enhanced concept based approach to ensure that only the synonyms of the most specific terms are searched for and extracted from the dictionary of the selected dataset during a particular search query. Hence, it will address the vocabulary mismatched issues which results to dissimilarities between the content of laymen queries and the expert vocabularies. It comprises of lines no 20 to 26 as shown in Figure 3.

```
20 Declare: getSynonyms-Terms (MST)
21 Fetch = Search (Synonyms of Most Specific
   Terms)
22 If Synonyms= @Synonyms(Most Specific Terms)then
```

```
For all Synonyms(Most Specific Terms) $k \in [1, K]$ do
    Extract = Retrieve (Synonyms of Most Specific Terms)
    New $SearchQR = SearchQR + ET_k$
End.
```

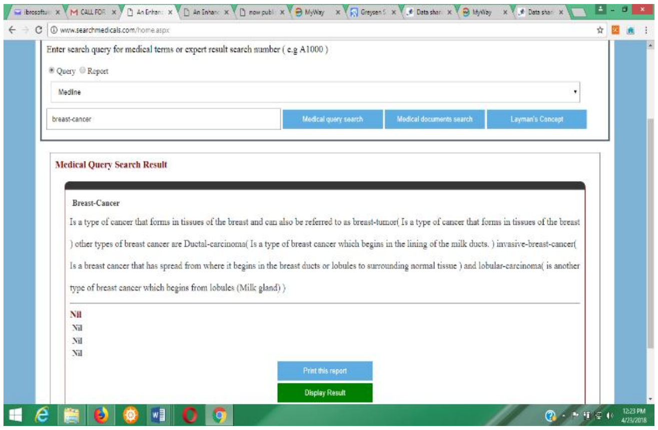
Figure 3: The Vocabulary Controlled Module

By fully incorporating these two modules in the proposed enhanced concept based approach, the medical terms control module would ensure that only the most specific medical terms in the layman's query are searched for and retrieved. While the vocabulary controlled module will ensure that only the synonyms of the extracted most specific terms from the layman queries are extracted and expanded into the new search query. With this, the challenge of medical terms in layman's queries not matching with their equivalent synonyms terms extracted from the dictionary of the selected dataset would be prevented.

Vocabulary mismatched issues between laymen queries and expert vocabularies during query expansion has also affected information retrieval systems performance which in turn prevented laymen patients from finding relevant and authoritative information. The failure of the current health information retrieval systems to fully integrate controlled vocabularies during query expansion and also not focusing only on labeled most specific medical terms during query expansion also hinders the information retrieval systems performance. Previous researches had targeted much on searching for general terms in a search query rather than most specific terms and also not restricting the synonyms search to only extracted most specific terms during query expansion.

Figure 4 is the Interface View of one of the many sample outputs generated by the proposed enhanced concept based approach as a result of the incorporation of the two special modules namely: medical terms control module and the vocabulary controlled module in it. The two special modules ensures that only the most specific terms in a launched search query are searched for, labeled and extracted, and their synonym terms being extracted from the dictionary of the selected dataset and extended into the new search query. This will ensure that the similarity score between the medical terms/concepts (W1, W2) contained in the two text segments T1 (Input Segment) and T2 (Output Segment) as shown in Table 1 and Figure 4 is significantly high and better addressed the vocabulary mismatched issues as compared to the existing approaches.

From Figure 4, the terms contained in T1 (Input Segment) is Breast Cancer which is related to all the terms Breast-tumor, Ductal-carcinoma, Invasive-breast-cancer and Lobular-carcinoma contained in the T2 (Output Segment). All the terms/concepts contained in the two texts segments T1 and T2 are similar and are related to the search term. That has shown the significant level at which the proposed enhanced concept based approach has gone in addressing vocabulary mismatched issues between laymen queries and expert vocabularies during query expansion.



www.searchmedicals.com/home.aspx

Enter search query for medical terms or expert result search number (e.g A1000)

● Query ○ Report

Medline

breast-cancer

Medical Query Search Result

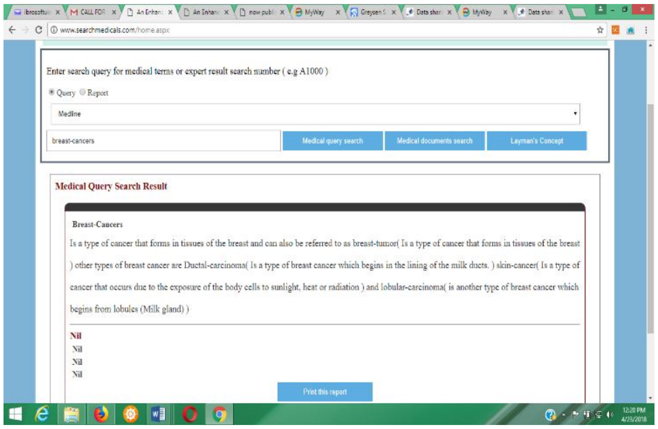
Breast-Cancer

Is a type of cancer that forms in tissues of the breast and can also be referred to as breast-tumor(Is a type of cancer that forms in tissues of the breast) other types of breast cancer are Ductal-carcinoma(Is a type of breast cancer which begins in the lining of the milk ducts.) invasive-breast-cancer(Is a breast cancer that has spread from where it begins in the breast ducts or lobules to surrounding normal tissue) and lobular-carcinoma(is another type of breast cancer which begins from lobules (Milk gland))

Nil
Nil
Nil
Nil

Figure 4: Sample output generated by the enhanced concept based approach in respect to addressing vocabulary mismatched issues.

Figure 5 is the sample output generated by the existing concept based approach in respect to vocabulary mismatched issues. Here, the search is not restricted to only most specific terms but rather on the general terms in the search query. And the synonyms search is also spread across the entire general terms extracted from the dictionary of the selected dataset. Hence, the similarity between the medical terms/concepts contained in the two text segments T1 (Input Segment) and T2 (Output Segment) as shown in Figure 5 is significantly low as some terms contained in the two text segments T1 and T2 are not related to the search term/record. For example, skin-cancer found in a breast-cancer search as shown in Figure 5, has shown the high level of vocabulary mismatched issues that exists in the existing concept based approach and other existing approaches compared to the improved concept based approach.



www.searchmedicals.com/home.aspx

Enter search query for medical terms or expert result search number (e.g A1000)

● Query ○ Report

Medline

breast-cancers

Medical Query Search Result

Breast-Cancers

Is a type of cancer that forms in tissues of the breast and can also be referred to as breast-tumor(Is a type of cancer that forms in tissues of the breast) other types of breast cancer are Ductal-carcinoma(Is a type of breast cancer which begins in the lining of the milk ducts.) skin-cancer(Is a type of cancer that occurs due to the exposure of the body cells to sunlight, heat or radiation) and lobular-carcinoma(is another type of breast cancer which begins from lobules (Milk gland))

Nil
Nil
Nil
Nil

Figure 5: Sample output generated by the existing concept based approach in respect to addressing vocabulary mismatched issues

4. PERFORMANCE ANALYSIS

For the specific purpose of evaluating the performance of our proposed enhanced concept based approach, we present in this section our experimental setup, the dataset used for the study, the metrics used in evaluating the performance of our proposed improved approach with the existing approaches namely: the latent semantic indexing (LSI), concept based approach (CBA), and the query likelihood model (QLM) and finally the experimental results were explicitly discussed in this section.

4.1. Experimental Setup

The experimental setup for the research work was carried out using HTML 5.0, C#.net, JavaScript and CSSnc. Windows 7 operating system with Intel (R) Core i7 processor, 3.40GHz and 4GB RAM memory. HTML – Hypertext Mark-up Language was used for creating the website structure, C#.net – an object oriented programming language was used for creating and activating the functionalities of the proposed algorithm web application structures, JavaScript – was used for adding effects to the website through slides, animations and quick interactions on the web and finally the CSS – Cascading Style Sheet was used for beautifying the website through styles.

4.2. Dataset

The datasets used for this study comprised of set of medical-related Web documents, provided by UMLS (Unified Medical Language System), Khresmoi project 6, MeSH (Medical Subject Heading) and Metamap [10]. These datasets covers a wide range of medical and patient's information. All related and relevant documents in the databases were downloaded from numerous free online sources, including Genetics Home Reference, Health on the Net organization certified websites, Diagnosia7 and Clinical.gov [16].

4.3. Performance Metrics

In order to measure the performance of our proposed method in respect to similarity scores, we used the Text Semantic Scoring Function which was also used in previous work in measuring the similarity scores between the terms (W1, W2) contained in the two text segments. Below is the illustrations of how the performance analysis was carried out:

i. Text Semantic Similarity Scoring Function

Given two input text segments T_1 and $T_{2,}$ using the Text Semantic Scoring Function, it measures the semantic similarities of the terms or concepts (W1, W2) contained in the two text segments T_1 and T_2. The similarity between the terms (W1, W2) contained in the two text segments T_1 and T_2 can be determined using the scoring function:

$$\mathrm{Sim}\ (T_1, T_2) = \frac{1}{2}(\frac{\sum_{w\in\{T1\}}(maxSim(w1T1)*idf(w))}{\sum_{w\in\{T1\}}idf(w)} + \frac{\sum_{w\in\{T2\}}(maxSim(w2T2)*idf(w))}{\sum_{w\in\{T2\}}idf(w)}) \quad [15], [13] \qquad (1)$$

The similarity score has a value between 0 and 1, with score 1 representing that the two text segments are identical and score 0 representing that the two text segments are not identical.
The IDF weighting represents the specificity of a word: when the specificity weighting is higher, that means the word is very specific to that particular document and when the specifity weighting is low, that means that the word is common among many documents. The IDF values for a word Wi can be obtained using:

$$\mathrm{IDF\ (Wi) = Log\ (N/Ni)} \quad [13], [15] \qquad (2)$$

4.4. Results and Discussion

For the specific purpose of comparing and evaluating the performance of the proposed enhanced concept based approach with the existing approaches, the Text Semantic Similarity Function was used. The existing approaches used in comparing and evaluating the proposed enhanced concept based approach were Query likelihood Model (QLM), Latent Semantic Indexing (LSI) and the existing Concept Based Approach (CBA). The proposed enhanced concept based approach focused more on addressing vocabulary mismatched issues between laymen queries and expert vocabularies during query expansion which affects system performance.
In order to address the vocabulary mismatched issues faced by laymen patients and their relatives in exploring authoritative information from their medical search queries and medical discharge documents online, the proposed enhanced concept based approach was implemented in two segments. Firstly by implementing the two program modules: the medical terms control and the Vocabulary Controlled Modules incorporated in the proposed enhanced concept based approach to specifically restrict the medical terms search to most specific terms. And secondly by severely ensuring that only the synonyms of the most specific medical terms extracted from the laymen queries are searched for, extended and fully implemented in the new expanded query.

A. i. Analysis Using Text Semantic Similarity Scoring Function

We compared and evaluated the semantic similarity scores and idf weighting values between the words (W1 and W2) contained in the two texts segments T1 (Input Segment) and T2 (Output Segment) generated by the proposed enhanced concept based approach. In order to get the maxSim and idf values, 200 sample data were extracted randomly from UMLS (Unified Medical Language System), Khresmoi project 6, MeSH (Medical Subject Heading) and Metamap datasets and were used in validating this approach. The study also used same threshold value of 0.50 as used in all similar experiments reviewed. There are no words that appear in the two text segments as identical (e.g. Skin-Cancer – Skin-Cancer or Kidney-Tumor – Skin-Tumor) or non-identical (e.g. where the words in the two text segments are not identical). However, the outcome of the comparison shows that, the proposed enhanced concept based approach contains medical terms/concepts (W1 and W2) in the two texts segments T1 (Input Segment) and T2 (Output Segment) that are similar and related. This makes it to have high maxSim similarities scores and idf weighting values in all the displayed simulation results. As compared to the existing approaches namely: (CBA, QLM, LSI), which displays one or more term(s) (W2) in the T2 (Output Segments) that are not similar and related. This has made them to score low maxSim and idf weighting values, as compared to the simulation results scored by the proposed enhanced concept based approach as shown from 5 out of the 200 samples results generated and displayed by the proposed approach in Table 1.

Table 1: Similarity Scores for the Proposed Enhanced Concept Based Approach

TEXT 1 (W1)	**TEXT 2 (W2)**	**maxSim**	**idf**
Skin-Cancer	Skin-Tumor, Carcinoma-Cancer	0.86	3.52
Ductal-Carcinoma	Lobular-Cancer, Breast-Cancer	0.89	3.61
Sarcoma-Cancer	Skin-Tumor, Carcinoma-Cancer	0.89	3.61
Kidney-Tumor	Kidney-Cancer, Kidney-Failure	0.87	3.54
Lung-Metastasis	Lung-Cancer, Lung-Tumor	0.87	3.54

We compared and evaluated the semantic similarity scores and idf weighting values between the words (W1 and W2) contained in the two texts segments T1 (Input Segment) and T2 (Output Segment) generated by the existing concept based approach. In order to get the maxSim and idf values, 200 sample data were extracted randomly from UMLS (Unified Medical Language System), Khresmoi project 6, MeSH (Medical Subject Heading) and Metamap datasets and were used in validating this approach. The study also used same threshold value of 0.50 as used in all similar experiments reviewed. There are no words that appear in the two text segments as identical (e.g. Skin-Cancer – Skin-Cancer or Kidney-Tumor – Skin-Tumor) or non-identical (e.g. where the words in the two text segments are not identical). However, the outcome of the comparison shows that, the proposed enhanced concept based approach contains medical terms/concepts (W1 and W2) in the two texts segments T1 (Input Segment) and T2 (Output Segment) that are similar and related. This makes it to have high maxSim similarities scores and idf weighting values in all the displayed simulation results. As compared to the existing concept based approach, which displays one or more term(s) (W2) in the T2 (Output Segments) that are not similar and related. This has made it to score low maxSim and idf weighting values, as compared to the simulation results scored by the proposed enhanced concept based approach as shown from 5 out of the 200 samples results generated and displayed by the existing concept based approach in Table 2.

Table 2: Similarity Scores for the Existing Concept Based Approach

TEXT 1 (W1)	**TEXT 2 (W2)**	**maxSim**	**idf**
Skin-Cancer	Skin-Tumor, Cancer	0.78	2.68
Ductal-Carcinoma	Lobular-Cancer, Skin-Cancer	0.68	1.87
Sarcoma-Cancer	Skin-Tumor, Carcinoma	0.77	2.65
Kidney-Tumor	Cancer, Kidney-Failure	0.76	2.59
Lung-Metastasis	Lung-Cancer, Breast-Tumor	0.67	1.86

We compared and evaluated the semantic similarity scores and idf weighting values between the words (W1 and W2) contained in the two texts segments T1 (Input Segment) and T2 (Output Segment) generated by the Query Likelihood Model. In order to get the maxSim and idf values, 200 sample data were extracted randomly from UMLS (Unified Medical Language System), Khresmoi project 6, MeSH (Medical Subject Heading) and Metamap datasets and were used in validating this approach. The study also used same threshold value of 0.50 as used in all similar experiments reviewed. There are no words that appear in the two text segments as identical (e.g. Skin-Cancer – Skin-Cancer or Kidney-Tumor – Skin-Tumor) or non-identical (e.g. where the words in the two text segments are not identical). However, the outcome of the comparison shows that, the proposed enhanced concept based approach contains medical terms/concepts (W1 and W2) in the two texts segments T1 (Input Segment) and T2 (Output Segment) that are similar and related. This makes it to have high maxSim similarities scores and idf weighting values in all the displayed simulation results. As compared to the Query Likelihood Model, which displays one or more term(s) (W2) in the T2 (Output Segments) that are not similar and related. This has made it to score low maxSim and idf weighting values, as compared to the simulation results scored by the proposed enhanced concept based approach as shown from 5 out of the 200 samples results generated and displayed by the Query Likelihood Model in Table 3.

Table 3: Similarity Scores for the Query Likelihood Model

TEXT 1 (W1)	**TEXT 2 (W2)**	**maxSim**	**idf**
Skin-Cancer	Tumor, Carcinoma-Cancer	0.76	2.84
Ductal-Carcinoma	Lobular-Cancer, Ductal	0.73	2.80
Sarcoma-Cancer	Skin-Tumor, Carcinoma	0.77	2.83
Kidney-Tumor	Brain-Cancer, Kidney-Failure	0.64	1.81
Lung-Metastasis	Lung-Cancer, Kidney-Cancer	0.66	1.82

We compared and evaluated the semantic similarity scores and idf weighting values between the words (W1 and W2) contained in the two texts segments T1 (Input Segment) and T2 (Output Segment) generated by the Latent Semantic Indexing. In order to get the maxSim and idf values, 200 sample data were extracted randomly from UMLS (Unified Medical Language System), Khresmoi project 6, MeSH (Medical Subject Heading) and Metamap datasets and were used in validating this approach. The study also used same threshold value of 0.50 as used in all similar experiments reviewed. There are no words that appear in the two text segments as identical (e.g. Skin-Cancer – Skin-Cancer or Kidney-Tumor – Skin-Tumor) or non-identical (e.g. where the words in the two text segments are not identical). However, the outcome of the comparison shows that, the proposed enhanced concept based approach contains medical terms/concepts (W1 and W2) in the two texts segments T1 (Input Segment) and T2 (Output Segment) that are similar and related. This makes it to have high maxSim similarities scores and idf weighting values in all the displayed simulation results. As compared to the Latent Semantic Indexing, which displays one or more term(s) (W2) in the T2 (Output Segments) that are not similar and related. This has made it to score low maxSim and idf weighting values, as compared to the simulation results scored by the proposed enhanced concept based approach as shown from 5 out of the 200 samples results generated and displayed by the Latent Semantic Indexing in Table 4.

Table 4: Similarity Scores for the Latent Semantic Indexing

TEXT 1 (W1)	TEXT 2 (W2)	maxSim	idf
Skin-Cancer	Skin-Tumor, Lung-cancer	0.74	2.86
Ductal-Carcinoma	Carcinoma, Breast-Cancer	0.72	2.88
Sarcoma-Cancer	Brain-tumor, Skin-Cancer	0.73	2.83
Kidney-Tumor	Kidney-Cancer, Kidney	0.74	2.86
Lung-Metastasis	Metastasis, Lung-Tumor	0.73	2.84

Table 5 gives a summary of the extracted results of various approaches (Proposed Enhanced Concept Based Approach, CBA, QLM and LSI) from Tables 1 – 4 showing their highest similarities scores in respect to maxSim and idf values obtained by each approach using the Text Similarity Scoring Function. The outcome of the results as shown in the table 5 and Figure 6 shows that the proposed approach had the best results of up to (0.89) 89% in relation to maxSim Scores and idf (specificity weighting scores) of up to 3.61 which shows clearly that the proposed approach better address the similarity issues encountered by laymen patients and their relatives as shown in Table 5 and Figure 6:

Table 5: Summary of the results for the Proposed Enhanced Concept Based Approach and the Existing Approaches: CBA, QLM & LSI.

Approaches	Similarity Measures	
	maxSim	idf
Proposed Approach	0.89	3.61
CBA	0.78	2.89
QLM	0.77	2.84
LSI	0.74	2.88

The diagrammatical representation of all the scores obtained in the table containing the summary of results for the proposed enhanced concept based approach and the existing approaches CBA, QLM & LSI is as shown in Figure 6:

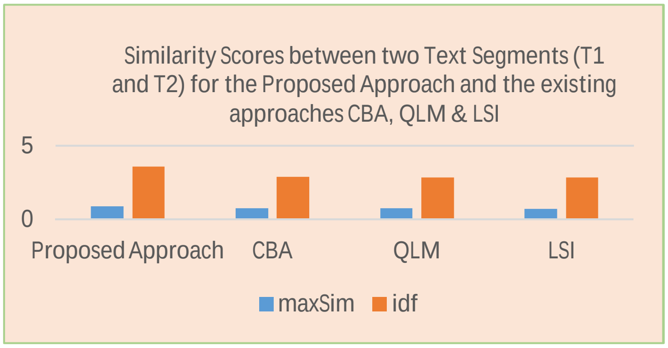


Figure 6: Similarity Scores between two text segments (T1 and T2) for the proposed enhanced approach and the existing approaches CBA, QLM & LSI.

When these two special modules: the medical concept free module and vocabulary controlled modules are fully incorporated into the proposed enhanced concept based approach, the medical concept free module will always ensure that only the most specific medical terms in the layman query are searched and retrieved while the vocabulary controlled module will also ensure that only the synonyms of the extracted most specific terms from the layman search queries are extracted and expanded into the new search query. Hence, the challenge of medical terms in layman queries not matching with their equivalent synonyms terms extracted from the dictionary of a selected dataset would be prevented.

The outcome of the summary of the experimental results in relation to Text Similarity Scoring Function obtained in Table 5 and Figure 6 shows that the Proposed Enhanced Concept Based Approach obtained a similarity score of 89% in respect to maxSim value and idf weighting value of 3.61 while LSI (Latent Semantic Indexing) obtained a maxSim score of 74% and idf weighting value of 2.88, CBA (Concept Based Approached) obtained a maxSim score of 89% and idf weighting value of 2.89 and finally the QLM (Query Likelihood Model) obtained a maxSim score of 77% and idf weighting value of 2.84 which clearly shows that the

Proposed Enhanced Concept Based Approach has better addressed the vocabulary mismatched issues encountered by laymen patients and their relatives by returning relevant and authoritative information.

The novelty of our proposed enhanced concept based approach is that, it has two separate and independent modules (medical terms controlled module and the vocabulary controlled module) that are specifically designed and incorporated in it to fully ensure that, only the most specific terms in the search queries are searched, labelled, extracted and expanded to the vocabulary controlled module. While the vocabulary controlled module also ensures that only the synonyms of the extracted most specific terms are searched for, extracted and expanded into the new search query. Hence it better address vocabulary mismatched issues encountered by laymen patients and their relatives in extracting information from their searched results.

The scientific reasons for the results obtained by the Proposed Enhanced Concept Based Approach, Concept Based Approach, Latent Semantic Indexing and Query Likelihood Model could be explained in reference to Tables 1, 2, 3, 4, 5 and Figure 6. The proposed approach concentrated more on implementing the two special modules: Medical Concept Free and Vocabulary Controlled Module incorporated in it and also limiting it search to most specific medical terms rather than general terms as applicable to the existing approaches. This has made the proposed approach to better address the information needs of laymen patients and their relatives in respect to addressing vocabulary mismatched issues.

The Proposed Enhanced Concept Based Approach works statistically based on the integrating strategy: For all concepts $n \in [1, N]$ do, Set SQ = query that is, all the medical concepts terms found in the search query would be labeled and extracted. $\mathrm{SQ} = $ “SQ” + “$\mathrm{CC_n}$”, For all the expansion terms $\mathrm{k} \in [1, \mathrm{K}]$, select all most specific concept terms. New $\mathrm{SQ} = \mathrm{SQ}$ # “$\mathrm{ET_k}$”. N is the number of concepts derived from the original query, SQ represents search query and CC_n represents the n th concept, K denotes the number of expansion terms, and ET_k represents the kth expansion terms. The kth is the last expansion term in an expansion query and the symbol # represents space character (i.e., 0x20), and the double quotation marks indicate that the string in it must appear consecutively.

5. CONCLUSION

The experimental results obtained in Table 5 and Figure 6 revealed that the Proposed Enhanced Concept Based Approach had the best results in respect to maxSim scores and idf weighting values using the Text Similarity Scoring Function with maxSim scored 0.89 (89%) and idf weighting value of 3.61, as against existing Concept Based Approach (CBA) which scored a maxSim score of 0.78 (78%) and idf weighting value of 2.89, Query Likelihood Model (QLM) scored a maxSim score of 0.77 (77%) and idf weighting value of 2.84 and finally the Latent Semantic Indexing (LSI) scored a maxSim score of 0.74 (74%) and idf weighting value of 2.88. This better results as shown from the experimental results obtained by the proposed enhanced concept based approach were as a result of the two special modules: medical concept free module and the vocabulary controlled modules that were designed, incorporated and fully implemented that provides readable and authoritative information to end-users.

The restriction of the search for medical terms by the proposed enhanced approach on only labeled most specific terms has increased the easy access to relevant and authoritative information. It has also prevented medical terms in a laymen queries not matching with their equivalent extracted synonyms terms (expert vocabularies) from the dictionary of the selected dataset used during the query expansion.

In addition, the experimental results in Table 5 and Figure 6 also shows that the Proposed Enhanced Concept Based Approach is the better approach to be applied in addressing the vocabulary mismatched issues faced by laymen patients and their relatives in exploring authoritative information from their medical search queries and medical discharge documents online. As every 89% of retrieved information by the Proposed Approach are Vocabulary Mismatched free and Readable information as against LSI (74%), QLM (77%) and CBA (78%). We recommend further work on this research study to include the design of algorithm that will address vocabulary mismatched issues encountered from retrieved videos, audios and images.

ACKNOWLEDGEMENT

The authors wishes to express their profound appreciation and gratitude to the management of Universiti Tun Hussein Onn Malaysia (UTHM) for funding the research. The research was funded under UTHM TIER-1 Grant with vot number H107 and Ministry of Education Malaysia under FRGS Grant with vot number K047.

REFERENCES

[1] Z. Xiaoshi, X. Yunqing, X. Zhongda, N. Sen, H. Qinan and H. Yaohai, "Concept-based medical document retrieval: THCIB at CLEF eHealth lab 2013 task 4." In Proceedings of the ShARe/CLEF eHealth Evaluation Lab, 2013.

[2] H. Suominen, L. Kelly, L. Goeuriot, L. Hanlen, A. Névéol, C. Grouin and G. Zuccon, “Overview of the CLEF eHealth evaluation lab 2015,” In International Conference of the Cross-Language Evaluation Forum for Languages (pp. 429- 443). Springer, Cham. 2015. https://doi.org/10.1007/978-3-319-24027-5_44

[3] B. Koopman, G. Zuccon, P. Bruza, L Sitb and M. Lawley, “An evaluation of corpus-driven measures of medical concept similarity for information retrieval,” In: Proceedings of CIKM. 2012, https://doi.org/10.1145/2396761.2398661

[4] N. Ksentini, M. Tmar and F. Gargouri, “Miracl at CLEF 2014: eHealth information retrieval task,” In:

Proceedings of the ShARe/CLEF eHealth Evaluation Lab.2014.

[5] R. White and E. Horvitz, "Cyb erchondria: Studies of the escalation of medical concerns in web search," In Technical report, Microsoft Research. 2015.

[6] E. M. Voorhees and R. M Tong, "Overview of the TREC 2011 medical records track," In: Proceedings of TREC, NIST. 2011.

[7] L. Goeuriot, L. Kelly, W. Li, J. Palotti, P. Pecina, G. Zuccon, A. Hanbury, G. Jones and H. Mueller, "ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User centered health information retrieval," In Proceedings of CLEF 2014 (2014).

[8] K. Drame, F. Mougin and G. Diallo, "Query expansion using external resources for improving information retrieval in the biomedical domain," In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. 2014.

[9] C. J Kalpathy, H. Muller, S. Bedrick, I. Eggel, A. G.S de Herrera and T. Tsikrika, "The CLEF 2013 medical image retrieval and classi_cation tasks," In: Working Notes of CLEF 2013, Cross Language Evaluation Forum. 2013.

[10] H. Thakkar, G. Iyer, K. Shah, and P. Majumder, "Team IRLabDAIICT at ShARe/CLEF eHealth 2014 Task 3: User-centered Information Retrieval system for Clinical Documents," In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab, 2014.

[11] S. Fox, "Health topics: 80% of internet users lo ok for health information online," In Technical Report, Pew Research Center, 2011.

[12] W. Shen, J. Y. Nie, X. Liu and X. Liui, "An investigation of the effectiveness of concept-based approach in medical information retrieval GRIUM @ CLEF2014eHealthTask 3," In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. 2014.

[13] S. Xie and Y. Liu, "Using Corpus and Knowledge Based Similarity Measure in maximum marginal relevance for meeting summarization," In Acoustics, speech and signal proceeding, 2008.

[14] I.U. Kontagora and I.R.A. Hamid, "Comparative Studies of Information Retrieval Approaches in User-Centered Health Information System," In: Ghazali R., Deris M., Nawi N., Abawajy J. (eds) Recent Advances on Soft Computing and Data Mining. SCDM 2018. Advances in Intelligent Systems and Computing, vol 700. Pp 171 – 180. Springer, Cham.2018.

[15] M. Rada, C. Courtney and S. Carlo, "Corpus-based and knowledge-based Measures of Text Semantic Similarity. In American Association for Artificial Intelligence (www.aaai.org). All right reserved @ 2006.

[16] H. Thakkar, G. Iyer and P. Majumder, "A comparative study of approaches in user-centered health information retrieval". In ArXiv preprint arXiv: 1505.01606. 2015.

[17] L. Goeuriot, L. Kelly, W. Li, J. Palotti, P. Pecina, G. Zuccon, A. Hanbury, G. Jones and H. Mueller, "ShARe/CLEF eHealth Evaluation Lab 2013, Task 2: User centered health information retrieval," In Proceedings of CLEF 2013 (2013).

[18] A. H. Pollack, A. Miller, S. R. Mishra and W. PD-atricians Pratt, "leveraging physicians and participatory design to develop novel clinical information tools. In AMIA Annual Symposium Proceedings/AMIA Symposium AMIA Symposium. 2016, 1030– 1039. 2016.

[19] S. R. Greysen, R. R. Khanna, R. Jacolbia, H. M. Lee and A. D Auerbach, "Tablet computers for hospitalized patients: a pilot study to improve inpatient engagement," In Journal of Hosp. Med. 9 (6) (2014) 396–399. 2014. https://doi.org/10.1002/jhm.2169

[20] C. Jaruskulchai, O. Thesprasith and M. Lawley, "Csku gprf-qe for medical topic web retrieval," In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. 2015.

[21] Y. P Yen, D. M. Walker, J. M. G. Smith, M. P. Zhou, T. L. Menser, and A. S. McAlearney, "Usability evaluation of a commercial inpatient portal," In International journal of medical informatics, 110, 10-18. 2018.

[22] Y. Liu, J. Shi and Y. Chen, "Patient-centered and experience-aware mining for effective adverse drug reaction discovery in online health forums," In Journal of the Association for Information Science and Technology, 69(2), 215-228. 2018.

[23] C. H. Halbert and B. W. Harrison, "Genetic counseling among minority populations in the era of precision medicine," In American Journal of Medical Genetics Part C: Seminars in Medical Genetics. 2018. https://doi.org/10.1002/ajmg.c.31604

[24] G. Aceto, V. Persico and A. Pescapé, "The role of Information and Communication Technologies in healthcare: taxonomies, perspectives, and challenges," In Journal of Network and Computer Applications. 2018.

[25] L. Kelly, L. Goeuriot, H. Suominen, A. Névéol, J. Palotti and G. Zuccon, "Overview of the CLEF eHealth evaluation lab 2016. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 255-266). Springer, Cham. 2018. https://doi.org/10.1007/978-3-319-44564-9_24

[26] M. A. Alyami, "Toward patient-centered personal health records systems to promote evidence-based decision-making and information sharing. 2018.

[27] Kontagora, I.U., Hamid, I.R.A. and Omar, N.A., "An Enhanced Concept based Approach for User Centered Health Information Retrieval to Address Presentation Issues". In International Journal of Advanced Computer Science and Applications (IJACSA), 10(2), pp 232 – 242, 2019. http://dx.doi.org/10.14569/IJACSA.2019.0100131.