# PROJECT FINAL REPORT

**Research Project No. ( EO03/18 )**

| | | | | | |
|---|---|---|---|---|---|
| **ENGLISH TITLE:** | Enhanced Arabic Latent Semantic Indexing (LSI) for Search Engines | | | | |
| **ARABIC TITLE:** | فهرسة محوسبة كامنة محسنة للغة العربية لمحركات البحث | | | | |
| **TOTAL BUDGET *:** | 3250 | | | | |
| **START DATE:** | 15/10/2018 | **EXTENSION TAKEN:** | | **END DATE **:** | 14/10/2019 |
| **PAPER(S) SUBMITTED:** | 1 | **PAPER(S) ACCEPTED:** | | | |
| **PAPER(S) PUBLISHED:** | | **CONFERENCE PAPER(S):** | 2 | | |

## RESEARCH TEAM

| | | | |
|---|---|---|---|
| **First: Principal Investigator** | | | |
| **Name:** | FAWAZ S. AL-ANZI | **Rank:** | PROFESSOR |
| **Faculty:** | FACULTY OF ENGINEERING & PETROLEUM | **Department:** | COMPUTER ENGINEERING |
| **Second: Co-Investigator(s)** | | | |
| **1. Name:** | | **Rank:** | |
| **Faculty:** | | **Department:** | |
| **2. Name:** | | **Rank:** | |
| **Faculty:** | | **Department:** | |
| **3. Name:** | | **Rank:** | |
| **Faculty:** | | **Department:** | |
| **4. Name:** | | **Rank:** | |
| **Faculty:** | | **Department:** | |
| **Third: Contributor(s)** | | | |
| **1. Name:** | DIA EDDIN ABUZEINA | **Rank:** | ASST_PROF |
| **Faculty:** | | **Department:** | |
| **2. Name:** | | **Rank:** | |
| **Faculty:** | | **Department:** | |
| **3. Name:** | | **Rank:** | |
| **Faculty:** | | **Department:** | |
| **4. Name:** | | **Rank:** | |
| **Faculty:** | | **Department:** | |

* Including Supplementary Budget taken (if any)

** Including Extension(s) taken (if any)

***Prof. Fawaz S. Al-Anzi***
*Computer Engineering Department*
*College of Engineering and Petroleum*
*Kuwait University*
*E-mail: Fawaz.Alanzi@ku.edu.kw*

# Final Report

# EO03/18

# Enhanced Arabic Latent Semantic Indexing (LSI) for Search Engines

**By**

**Prof. Fawaz S. Al-Anzi (PI)**

**Dr. Dia AbuZeina (CoI)**

**15-10-2019**

## Table of Contents

## Executive Summary


### *Abstract*

As a common document representation model, the Vector Space Model (VSM) is that is widely used in data mining and information retrieval (IR) systems. Some challenges in this technique such as high dimensional space and semantic looseness of the representation. Consequently, the latent semantic indexing (LSI) was suggested to lessen the feature dimensions and to generate a richer semantic features that can represent conceptual term-document associations. LSI has been effectively employed in search engines and text classification applications. In this research, we proposed an innovative method to further enhance the quality of the retrieved documents in search engines for Arabic language in particular. Where we introduce a new extension of the LSI technique based on the documents evaluating cosine similarity measures in the term-by-document matrix. The evaluation of the performance was carried out using an Arabic language data collection that contains 500 documents with more than 30,000 unique words. A testing set comprises of keywords from specific domains is used to evaluate the quality of the top 20-30 retrieved documents using different singular values. The results show that the performance of the proposed new method is superior when compared to the standard LSI.



***Keywords:*** Arabic Text, Latent Semantic Indexing, Search Engine, Dimensionality Reduction

# Chapter 1:

# Introduction

Due to the tense development of online data presence, search engines have a noticeable title role in information retrieval (IR) and web data mining uses. The web is the main source of open data that certainly needs effective algorithms for retrieving and cleaning out the textual data as well as other kinds. Henceforth, search engines are suitable to become more intelligent in obtaining the desired contents. In broad-spectrum, textual data can be represented using the Vector Space Model (VSM) where each document is represented using a vector of attributes, many of which could be nil. However, VSM have some challenges such as there are overlong features and semantic losing representation. Hence, the Latent Semantic Indexing (LSI) method is suggested to ease such encounters and optimistically to enhance the presentation. LSI aims at converting the original textual vectors into conceptual vectors that are written off by two properties: reduced dimensions and semantic rich features. The essential nature that determines the feature of the LSI is found in the semantic property that carry out by returning semantically close documents without the restriction to have the exact search keywords.

LSI is based on a proposition from linear algebra that is named Singular Valued Decomposition (SVD). The SVD can transform the textual data, which is represented as a large term-by-document matrix into a lesser semantic space characterized as three matrices where the product of the created matrices equivalent the original term-by-document matrix. Therefore, the primary step of LSI is to decompose the term-by-document (A) matrix as follows: $A=USV^T$ where U is a matrix that gives the weights of terms, S makes available the eigenvalues for each principal component direction, and $V^T$ is a matrix that offers the weights of documents. $V^T$ is the matrix that comprises of the document feature vectors that are normally used in IR and text mining uses. Fig. 1 shows the decomposition procedure that truncates a term-by-document matrix (A) into the three matrices.

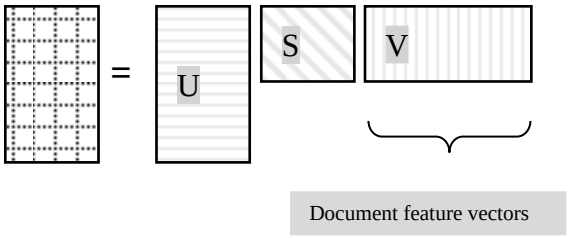



**Fig. 1.** SVD Decomposition Process.

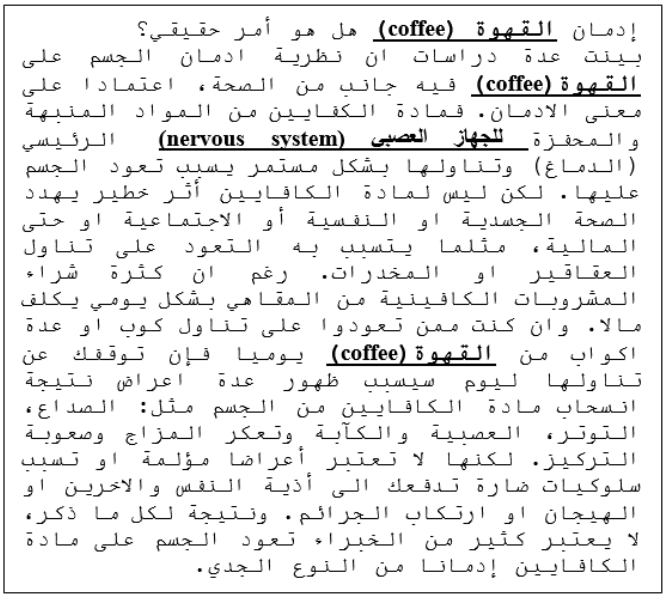
Consequently, a standard process of creating LSI starts with a term-by-document matrix to generate the required feature vectors that are used for classification. Nevertheless, a term-by-document matrix is in general formed by means of different values such as Boolean flags, counts, or weights. This is used to track the occurrences of terms in documents. For classification, the produced LSI feature vectors are generally developed using a similar measure such as Euclidian, Mahalanobis, Manhattan, cosine similarity, etc. The cosine similarity measure is known to be one of the popular distance measures in pattern recognition. In this research, we propose an extension of the LSI implementation by using the cosine measures instead of the standard word co-occurrences values. Hence, the proposed method forms the term-by-document matrix using the cosine measures between documents before employing the SVD process.

In the next chapter, we present the motivation followed by the literature review in chapter 3. In chapter 4, we present the proposed method followed by the experimental results in chapter 5. We conclude in chapter 6 and references used are presented in chapter 7. The publication results of this research is demonstrated in the appendix

# Chapter 2:

# Motivation

Text mining systems are intuitively in need for very efficient algorithms that intelligently understand the search engine's documents as well as the query's keywords. Moreover, huge online data requires considering the semantic relationships between both the documents and the words that are also called co-occurrences. Unlike trivial searching methods that are based on traditional text matching, LSI is characterized by semantic rich value that enable the system to return useful results without having exact matching words between the document and the query's keywords. For example, if we search for the word "coffee," it is expected that the system will return many documents related to this word, however, it might return other related documents that have no "coffee" word in it, but are semantically related to the word "coffee." That is, it is possible to obtain documents that belong to the topic, such as stimulant effects, caffeine, etc. Fig. 2 shows an example of an article that contains the word "coffee." The figure also shows that the document has other words such as "nervous system." Therefore, the searching process for the word "coffee" might return documents related to "nervous system" topics that, at the same time, have no "coffee" word in it. This is the strength of the LSI method and one reason for its popularity.



إدمان **القهوة (coffee)** هل هو أمر حقيقي؟
بينت عدة دراسات ان نظرية ادمان الجسم على **القهوة (coffee)** فيه جانب من الصحة، اعتمادا على معنى الادمان. فمادة الكفايين من المواد المنبهة والمحفزة **للجهاز العصبي (nervous system)** الرئيسي (الدماغ) وتناولها بشكل مستمر يسبب تعود الجسم عليها. لكن ليس لمادة الكافايين أثر خطير يهدد الصحة الجسدية او النفسية أو الاجتماعية او حتى المالية، مثلما يتسبب به التعود على تناول العقاقير او المخدرات. رغم ان كثرة شراء المشروبات الكافينية من المقاهي بشكل يومي يكلف مالا. وان كنت ممن تعودوا على تناول كوب او عدة اكواب من **القهوة (coffee)** يوميا فإن توقفك عن تناولها ليوم سيسبب ظهور عدة اعراض نتيجة انسحاب مادة الكافايين من الجسم مثل: الصداع، التوتر، العصبية والكآبة وتعكر المزاج وصعوبة التركيز. لكنها لا تعتبر أعراضا مؤلمة او تسبب سلوكيات ضارة تدفعك الى أذية النفس والاخرين او الهيجان او ارتكاب الجرائم. ونتيجة لكل ما ذكر، لا يعتبر كثير من الخبراء تعود الجسم على مادة الكافايين إدمانا من النوع الجدي.

**Fig. 2.** An Example of word co-occurances of "coffee" and "nervous system".

Using the enhancing searching process with thorough overwhelming digital data requires an endless research effort to satisfy the users' requests. In fact, text mining is a challenging task since documents usually have mixed contents that make it difficult to digitally understand the document's category. For an illustration, Fig. 2 shows a document that has different words, such as "headaches": "صداع", "addiction": "ادمان", "drugs": "مخدرات", "tension": "التوتر", "frenzy": "الهيجان", and "crimes": "الجرائم". Such diverse words make it vague for IR algorithms to search correctly for the required data. By nature, medical documents require precise algorithms that can adequately find the proper documents for the user.

The LSI has been proven to be a valuable tool that reveals the semantical relationship between data objects (i.e: the words in this research). Based on detected underlying semantic distinctions, LSI is able to bring out the relevant documents that do not contain the searching keyword at all. Fig. 3 shows a medical article that is related to "التنفس":"breathing." If a user searches for the word "الاوكسجين" : "oxygen," then the semantic loss methods (i.e plain keyword search) will fail since there is no exact match between the searching word and the document's words. However, LSI does support the semantic search and this is what a user is looking for in the search.

الغُبار قد يضرّ المصابين بمشاكل في **التنفّس (breathing)**

جنيف - رويترز - قالت منظمة الصحة العالمية ان الغبار الناجم عن ثورة بركان ايسلندا قد تضر أيضا الاشخاص الذين يعانون مشاكل في التنفس (**breathing**)، لأن هذه الجزيئات عند استنشاقها يمكن ان تصل الى المناطق المحيطة من القصيبات التنفسية (**breathing**) والرئتين (**lungs**)، ويمكن أن تسبب مشاكل، خاصة للأشخاص الذين يعانون الربو أو مشاكل بالجهاز التنفسي (**breathing**).
وفي جانب متصل، أكدت الهيئة البريطانية للوقاية الصحية أن الرماد البركاني لا يشكّل خطورة كبيرة على الصحة العامة، ومن غير المرجّح أن يسبب ضررا كبيرا، حيث يتطلب الامر التعرض بشكل كبير جدا للغبار المنخفض السَّمية حتى يكون هناك تأثير في الناس.
وقال كين دونالدسون أستاذ علم السموم التنفسية (**breathing**) في جامعة أدنبره لـــ «رويترز»: «هناك تأثير ضعيف بشكل كبير في الغلاف الجوي، حيث يتشتت بفعل الرياح، ما يعني أن الكمية التي تصل إلى الأرض صغيرة للغاية». واتفق دونالدسون على ان الناس المصابين بأمراض بالرئة (**lung**) بالفعل يجب ان يبقوا في اماكن مغلقة إذا كان هناك تغيير ملموس فى مستويات الجسيمات.

**Fig. 3.** An Example of "breathing" in a medical document.

# Chapter 3:

# Literature Review

In the literature, there are many studies that discuss the LSI technique. In particular, LSI is used for the text mining task, such as text classification, text summarization, text clustering, search engines, etc. LSI initially was presented by Deerwester in [1] as a standard dimension reduction technique in IR. Reference [2] presents an algorithm to enhance the results of search engines. The algorithm combines common phrase discovery and LSI techniques to separate search results into meaningful groups. Reference [3] presents a new implementation of the standard LSI. The new implementation aims to provide efficient, extensible, portable, and maintainable LSI. Reference [4] presents a theoretical model for understanding the performance of LSI in retrieval applications. Reference [5] presents an LSI based method for fully automated cross-language document retrieval in which no query translation is required.

Reference [6] describes a word clustering approach that is based on LSI. Reference [7] proposes a local LSI method called "Local Relevancy Weighted LSI" to improve text classification by performing a separate SVD on the transformed local region of each class. Reference [8] uses LSI to automatically identify the conceptual gene relationships from titles and abstracts in a database citation. Reference [9] proposes and empirically tests the feasibility and utility of post-retrieval clustering of digital forensic text string search results – specifically by using Kohonen Self-Organizing Maps (SOM) as a self-organizing neural network approach. Reference [10] proposes a hybrid term frequency – inverse document frequency (TF-IDF) that is based on algorithm and a clustering based algorithm for obtaining multi-post summaries of Twitter posts along with the detailed analysis of Twitter post domain. Reference [11] uses LSI for automatic software clustering. LSI was used as the basis to cluster software components, source code, and its accompanying documentation. Reference [12] proposes two text summarization approaches: the modified corpus-based approach (MCBA) and the LSI-based approach.

LSI has been widely documented as a retrieval method that employs SVD for semantic rich reduced feature vectors. Nevertheless, utilizing LSI and SVD requires understanding which values in the reduced dimensional space contain the word relationships (latent semantic) information. Hence, many studies in the literature have discussed this important aspect. Reference [13] presents an empirical study of the required dimensionality for large-scale LSI applications. Reference [14] was developed as a model for understanding which values in the reduced dimensional space contain the term relationship (latent semantic) information.

Regarding cosine similarity, it is a well-known similarity measure that has been widely mentioned in the literature. Reference [15] indicates that cosine similarity dominants have similar measures in IR and text classification. This measure is based on the cosine of the angle between two vectors. Reference [16] demonstrates that the similarity between two documents can be measured using the cosine of the angle between the two document feature vectors, which are represented by using VSM. Theodoridis and Koutroumbas in [17] defines the cosine similarity measure as: $Scosine(x,y) = x^T y/\|x\|\|y\|$ where $\|x\|$ and $\|y\|$ are the lengths of the vectors x and y, respectively. Reference [18] proposes that the cosine similarity is a robust metric for scoring the similarity between two strings. Reference [19] demonstrates that the cosine similarity is used to find the vectors neighborhood. Reference [20] demonstrates that the cosine similarity is easy to interpret and simple to compute for sparse vectors, this indicates that it is widely used in text mining and IR. Reference [21] used the cosine similarity measure for the Arabic language text summarization. Reference [22] uses the cosine measure for the language identification problem.

# Chapter 4:
# The proposed method

This The proposed enhanced method for the standard LSI starts initially, by the preprocessing step where it is performed by declaring the stop words and ignoring the characters' list. In addition, all small words that were less than three characters in length were discarded. A normalization process was also performed to change some Arabic characters such as (أ→ا) and (إ→ا). As shown in Fig. 2, the term-by-document (A) matrix was created using the unique words in the used corpus. The term-by-document (A) matrix weighted used TF-IDF. For comparison purposes between the proposed method and the standard LSI, A was decomposed into three matrices (U: Term by dimension; S: Singular values; and $V^T$: Document by dimension). The diagonal of matrix S contains singular values so that one can choose the desired reduced dimensions. In general, not all singular values were considered; instead, only the most important values were considered starting from the first singular values up to the desired value (k).

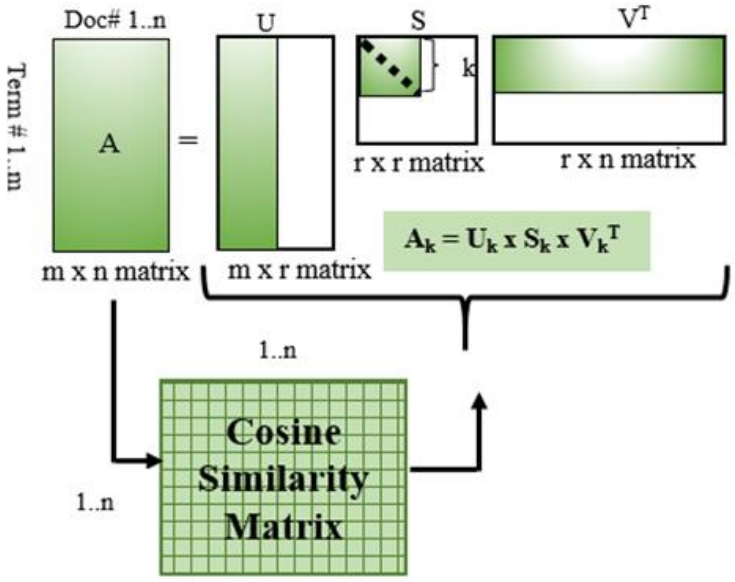



**Fig. 4.** Forming cosine similarities matrix after SVD.

The proposed method has an extension of the standard LSI by creating a new matrix called the cosine similarity matrix. This new matrix has used the cosine similarities between all documents in the corpus instead of the co-occurrences (i.e. instead of the frequency of a word in a document) that usually are used when creating term-by-document matrices. Hence, the enhanced method is summarized by using four main steps as follows: 1) creating the standard term-by-document matrix using word co-occurrences; 2) the matrix is weighted using TF-IDF; 3) based on the standard term-by-document matrix, we formed a new matrix called cosine similarity matrix that contains the cosine measures between each two vectors in the standard term-by-document matrix ; and then finally, 4) the SVD is used to truncate the cosine similarity matrix to generate the enhanced feature vectors that are used in the search engine. Of course, different singular values (k) might be investigated to find the optimal performance.

Fig. 5 demonstrates an example on how to create a cosine similarity matrix for three documents. The diagonal entries contain 1.0 as the cosine of angle zero, which is 1.0 (i.e. the document itself). Hence, as our corpus contains 800 documents, the cosine similarity matrix of the used corpus is of the size 800 × 800. Of course, the cosine similarity matrix is a symmetrical matrix.

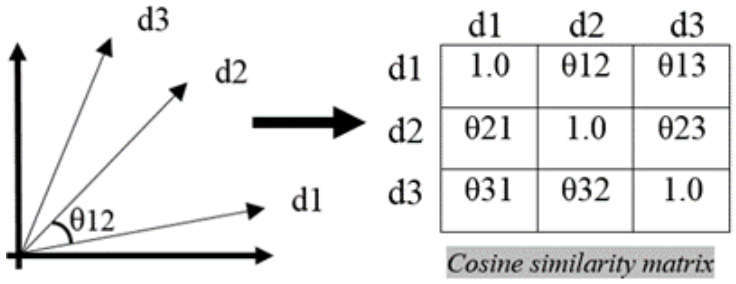



**Fig. 5.** A matrix of cosine similarities for three vectors.

In both cases, the standard LSI or the proposed method, the query's keywords have to transfer to the LSI space. For the standard LSI, the query's feature vectors transform into the new reduced space that is called "folding-in." This is done by using the following formula: $V'^{T}=AUS^{-1}$. Hence, $V'^{T}$ contains the reduced query's feature vectors that are used along with $V^{T}$ in the classification process. For the proposed method, the query's feature

vectors have two transformation steps. The first is regarding the cosine measures against all feature vectors of the training documents before using the folding-in technique as a second step (i.e. similar to the standard LSI but for the cosine measure instead of the word co-occurrences). Fig. 6 shows how to generate the query's vector in terms of cosine similarity. Hence, the cosine similarity matrices of the training and the testing set are generated for the new SVD implementation.

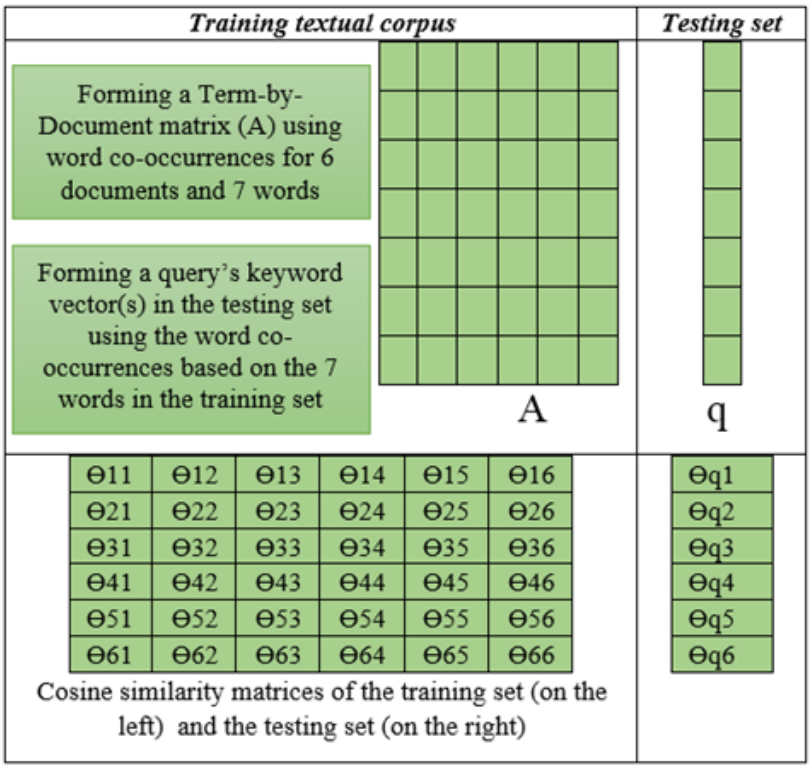



**Fig. 6.** An example of creating cosine similarity matrices.

# Chapter 5:
# The experimental results

The proposed method was evaluated using an Arabic textual corpus containing 800 documents, 353,888 words, and 47,222 unique words. The data collection was in regards to medical stories obtained from Alqabas [23], a Kuwaiti newspaper. A testing set contained five medical keywords that were used as queries for the developed search engine. Hence, the testing set arbitrarily contained {"الزهايمر" : "Alzheimer", "فيروس" : "virus", "الاوكسجين" : "oxygen", "القهوة" : "coffee", "اشعة" : "rays"}. However, a query could have more than one word (i.e. a sentence of many words or an article). Table 1 shows more information regarding the testing set and its appearance in the training corpus. Table 1 shows that the word "القهوة" : "coffee" appeared 143 times in 36 different documents.

**Table 1. Testing set information**

| Query word | Total appearance | Total documents |
|---|---|---|
| **"الزهايمر" : "Alzheimer"** | **32** | **14** |
| **"فيروس" : "virus"** | **204** | **66** |
| **"الاوكسجين" : "oxygen"** | **55** | **36** |
| **"القهوة" : "coffee"** | **143** | **36** |
| **"اشعة" : "rays"** | **324** | **103** |

Since the number of singular values is important in LSI applications, we considered a wide range of singular values to measure the performance for both cases (i.e. standard LSI and the proposed method). Hence, the search engine was evaluated using the different number of feature vectors dimensions (k). That is, a series of experiments were performed using the following k :{ k=10, 20, 30, 40, 50, 60, 70, 80, 90, 100,150,200, 250,300, 350,400, 500}. At each singular value, we analyzed the top-20 retrieved documents to investigate the query's keyword occurrences.

Table 2 shows the performance of the word "الزهايمر" : "Alzheimer." In the table, the first row indicates the medical query keyword that is the first word in the testing set. The first column indicates the singular values k that starts at 10 and ends at 500. At k=10, the word "الزهايمر" : "Alzheimer" is found in the first document zero times while it was found in the top-20 retrieved documents three times, using the standard LSI as shown. On the other hand, it was found one time in the first document and 4 times in the top-20 retrieved documents using the proposed method. The results show that the retrieved document using the proposed method is of a high quality compared to the standard LSI even with lower dimensions. For an illustration, at k=80, the standard LSI retrieved a document that contained 1 occurrence of the searched word with 11 occurrences in the top-20 documents, while the proposed method returned a document that contained 6 occurrences with 20 occurrences in the top-20 documents. Table 2 also shows that the maximum occurrences of the word "الزهايمر" : "Alzheimer" is 27 times using standard LSI, however, it scored 29 occurrences using the proposed method. The results presented in Table 2 did not require the

exact match cases as we considered the word “الزهايمر" : "Alzheimer” to be the same as saying “the Alzheimer” “زهايمر" and “بالزهايمر ", etc. Hence, different variations of the same word were counted.

**Table 2. Searching results for different dimensions of “alzheimer”**

| **"Alzheimer" : "الزهايمر** | | | | |
|---|---|---|---|---|
| | **The Standard LSI** | | **The Proposed Method** | |
| **k** | **First doc.** | **Top-20 doc.** | **First doc.** | **Top-20 doc.** |
| **10** | **0** | **3** | **1** | **4** |
| **20** | **0** | **6** | **6** | **13** |
| **30** | **0** | **7** | **6** | **13** |
| **40** | **0** | **10** | **6** | **18** |
| **50** | **1** | **3** | **6** | **18** |
| **60** | **1** | **7** | **6** | **18** |
| **70** | **1** | **11** | **6** | **18** |
| **80** | **1** | **11** | **6** | **20** |
| **90** | **1** | **7** | **6** | **16** |
| **100** | **1** | **11** | **6** | **16** |
| **150** | **1** | **19** | **6** | **21** |
| **200** | **1** | **21** | **6** | **23** |
| **250** | **1** | **18** | **6** | **23** |
| **300** | **1** | **18** | **6** | **25** |
| **350** | **3** | **24** | **6** | **25** |
| **400** | **3** | **27 (max)** | **6** | **26** |
| **500** | **3** | **25** | **6** | **29 (max)** |

Table 3 shows the performance of the word "فيروس" : "virus." Using k=40, the proposed method had 114 occurrences of the searched word while it returned only 106 occurrences at k=150. Hence, with lower dimensions, the proposed method demonstrated better results. The proposed method also gave better results for the first retrieved document as it had 19 occurrences of the searched word, while it had nothing related to the searched word in the standard LSI.

**Table 3. Searching Results for different dimensions of "virus"**

| "فيروس" : "virus" | | | | |
|---|---|---|---|---|
| | The Standard LSI | | The Proposed Method | |
| k | First doc. | Top-20 doc. | First doc. | Top-20 doc. |
| 10 | 0 | 36 | 19 | 74 |
| 20 | 14 | 77 | 19 | 81 |
| 30 | 14 | 57 | 19 | 99 |
| 40 | 19 | 68 | 15 | 114 (max) |
| 50 | 19 | 74 | 19 | 92 |
| 60 | 19 | 79 | 15 | 87 |
| 70 | 19 | 82 | 19 | 91 |
| 80 | 19 | 94 | 19 | 94 |
| 90 | 19 | 91 | 19 | 96 |
| 100 | 19 | 78 | 19 | 96 |
| 150 | 19 | 106 (max) | 19 | 96 |
| 200 | 19 | 99 | 15 | 95 |
| 250 | 19 | 91 | 15 | 91 |
| 300 | 19 | 91 | 15 | 100 |
| 350 | 19 | 87 | 15 | 101 |
| 400 | 19 | 89 | 15 | 92 |
| 500 | 15 | 94 | 15 | 83 |

Table 4 shows the performance of the word "الاوكسجين": "oxygen." This word did not appear in the first retrieved document for both the standard LSI and the proposed method.

However, for the top-20 list, the proposed method had 29 occurrences of this word while it had just 17 occurrences using the standard LSI.

**Table 4. Searching results for different dimensions of "oxygen"**

| "oxygen" : "الاوكسجين" | | | | |
|---|---|---|---|---|
| | The Standard LSI | | The Proposed Method | |
| k | First doc. | Top-20 doc. | First doc. | Top-20 doc. |
| 10 | 0 | 2 | 0 | 7 |
| 20 | 0 | 7 | 0 | 8 |
| 30 | 2 | 10 | 0 | 11 |
| 40 | 2 | 11 | 0 | 10 |
| 50 | 2 | 5 | 6 | 11 |
| 60 | 2 | 6 | 6 | 13 |
| 70 | 2 | 10 | 6 | 13 |
| 80 | 2 | 10 | 2 | 13 |
| 90 | 2 | 8 | 2 | 13 |
| 100 | 2 | 9 | 2 | 13 |
| 150 | 2 | 14 | 2 | 18 |
| 200 | 2 | 8 | 2 | 14 |
| 250 | 2 | 9 | 2 | 17 |
| 300 | 2 | 16 | 2 | 16 |
| 350 | 2 | 16 | 2 | 19 |
| 400 | 2 | 17 (max) | 2 | 19 |
| 500 | 2 | 17 | 2 | 26 (max) |

Table 5 shows that the first document had 60 occurrences (it is a relatively long document) of the word "القهوة" : "coffee," while it did not return any query word in the first

document using the standard LSI. It is worthwhile to observe that the standard LSI retrieved this long document at k=90 whereas it retrieved it at k=10 using the proposed method.

**Table 5. Searching results for different dimensions of "coffee"**

| "القهوة" : "coffee" | | | | |
|---|---|---|---|---|
| | **The Standard LSI** | | **The Proposed Method** | |
| **k** | **First doc.** | **Top-20 doc.** | **First doc.** | **Top-20 doc.** |
| **10** | **0** | **68** | **60** | **60** |
| **20** | **8** | **80** | **60** | **88** |
| **30** | **8** | **84** | **60** | **93** |
| **40** | **60** | **84** | **60** | **94** |
| **50** | **8** | **85** | **60** | **105** |
| **60** | **8** | **85** | **60** | **105** |
| **70** | **8** | **95** | **60** | **111** |
| **80** | **8** | **95** | **60** | **110** |
| **90** | **60** | **105** | **60** | **115** |
| **100** | **60** | **107** | **60** | **116** |
| **150** | **60** | **105** | **60** | **119** |
| **200** | **60** | **111** | **60** | **121** |
| **250** | **60** | **114** | **60** | **122 (max)** |
| **300** | **60** | **118** | **60** | **122 (max)** |
| **350** | **60** | **117** | **60** | **122 (max)** |
| **400** | **60** | **116** | **60** | **122 (max)** |
| **500** | **60** | **120 (max)** | **60** | **119** |

Table 6 shows that the first document returned three occurrences of the word "اشعة" : "rays" using both methods. Table 6 also shows that the performance started decreasing after k=200. Therefore, each LSI based application had a particular range of singular values (k) where it gave the optimal performance.

**Table 6. Searching results for different dimensions of "rays"**

| "اشعة" : "rays" | | | | |
|---|---|---|---|---|
| | The Standard LSI | | The Proposed Method | |
| k | First doc. | Top-20 doc. | First doc. | Top-20 doc. |
| 10 | 3 | 64 | 3 | 40 |
| 20 | 3 | 39 | 4 | 108 |
| 30 | 2 | 34 | 13 | 104 |
| 40 | 3 | 56 | 4 | 104 |
| 50 | 10 | 101 | 5 | 105 |
| 60 | 8 | 104 | 8 | 113 |
| 70 | 8 | 93 | 8 | 112 |
| 80 | 8 | 93 | 8 | 112 |
| 90 | 8 | 100 | 8 | 112 |
| 100 | 8 | 100 | 8 | 114 |
| 150 | 8 | 112 (max) | 22 | 116 (max) |
| 200 | 8 | 108 | 22 | 111 |
| 250 | 22 | 110 | 22 | 108 |
| 300 | 10 | 105 | 22 | 106 |
| 350 | 10 | 104 | 22 | 102 |
| 400 | 10 | 105 | 22 | 108 |
| 500 | 2 | 100 | 22 | 97 |

In addition, we evaluated the performance by measuring the percentage of the matched words among all occurrences in the training set. For example, the word "اشعة" : "rays" appears 324 times in the corpus. The standard LSI showed this word 112 times in the top-20 list as indicated in table 6. However, the proposed method listed it 116 times. Hence, the percentage for the standard LSI is 112/324=0.346. For the proposed method, the percentage is 116/324=0.358. These percentages are shown in table 7 for all words of the testing set. The table also shows that the average of the percentages for the standard LSI is 0.571 and for the proposed method is 0.629. This means that the proposed method outperforms the standard LSI by 5.83% for the top-20 retrieved documents.

**Table 7. The percentage of the retrieved searching words**

| # | Word | The Standard LSI | The Proposed Method |
|---|---|---|---|
| 1 | "الزهايمر" : "Alzheimer" | 0.844 | 0.906 |
| 2 | "فيروس" : "virus" | 0.520 | 0.559 |
| 3 | "الاوكسجين" : "oxygen" | 0.309 | 0.473 |
| 4 | "القهوة" : "coffee" | 0.839 | 0.853 |
| 5 | "اشعة" : "rays" | 0.346 | 0.358 |
| | Average | 0.571 | 0.629 |

In fact, other evaluation methods should be required since we only compared the match word while the semantic quality of the retrieved documents should also be evaluated. Fig. 7 shows the graphical representation of the performance differences between the standard LSI

and the proposed method. The graph's information is based on the percentages calculated in Table 7.

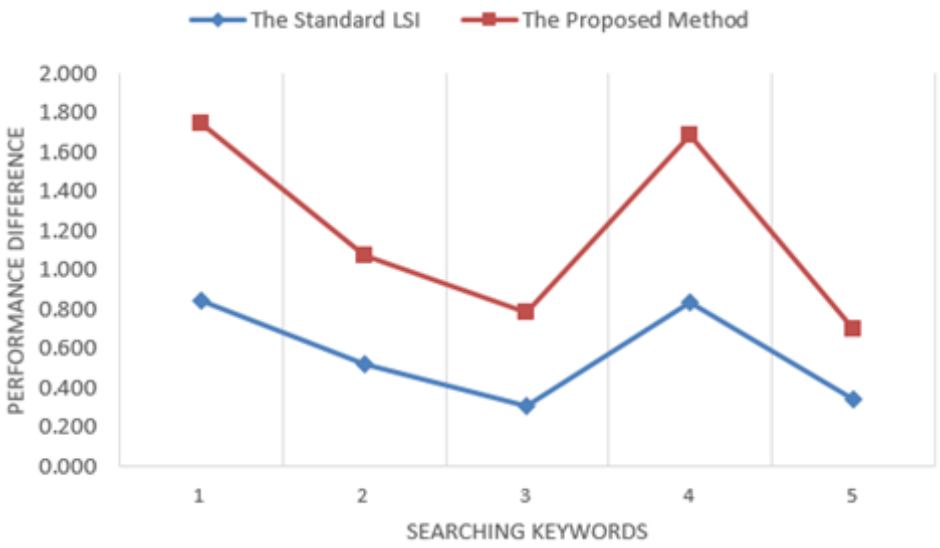



**Fig. 7.** The performance enhancement using the proposed method

Finally, the proposed method is suitable for relatively small data collections. However, it might not be very efficient for very large corpora that contains millions of documents. In fact, creating the cosine similarity matrix $O(n^2)$ where n is the total number of documents in the corpus is very complex and requires an extensive amount of time. Nevertheless, this method shows a possible enhancement especially when we look for precise results for highly mixed contents as medical documents. In addition, the recently available high speed machines might help to improve query searches over time with greater complexity. Moreover, as quickly as technology is evolving, in the future, we will be able to utilize these advancements with greater accuracy, faster speeds, and more enhanced data.

# Chapter 6:
# Conclusion

This research presents a new variant of the LSI technique for search engines. A comprehensive experimental evaluation shows the feasibility of the LSI technique as well as the enhancements of the new method over the standard LSI technique. The results showed that using the documents' cosine similarities instead of just word co-occurrences enhances the performance of search engines. The proposed method shows that the top-20 retrieved documents are of a higher quality than the top-20 documents retrieved using the standard LSI. As a future work, we propose to investigate the proposed method for larger data collections as well as investigating the time and space complexities of the proposed method. In addition, the evaluation should include the semantic quality and not just matched words.

# Chapter 7:
# References


[1] Deerwester, Scott, et al. "Indexing by latent semantic analysis." Journal of the American society for information science 41.6 (1990): 391.

[2] Osiński, Stanislaw, and Dawid Weiss. "A concept-driven algorithm for clustering search results." Intelligent Systems, IEEE 20.3 (2005): 48-54.

[3] Letsche, Todd A., and Michael W. Berry. "Large-scale information retrieval with latent semantic indexing." *Information sciences* 100.1 (1997): 105-137.

[4] Kontostathis, April, and William M. Pottenger. "A framework for understanding Latent Semantic Indexing (LSI) performance." Information Processing & Management 42.1 (2006): 56-73.

[5] Dumais, Susan T., et al. "Automatic cross-language retrieval using latent semantic indexing." *AAAI spring symposium on cross-language text and speech retrieval*. Vol. 15. 1997.

[6] Bellegarda, J. R., Butzberger, J. W., Chow, Y. L., Coccaro, N. B., & Naik, D. (1996, May). A novel word clustering algorithm based on latent semantic analysis. In Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on (Vol. 1, pp. 172-175). IEEE.

[7] Liu, T., Chen, Z., Zhang, B., Ma, W. Y., & Wu, G. (2004, November). Improving text classification using local latent semantic indexing. In Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on (pp. 162-169). IEEE.

[8] Homayouni, R., Heinrich, K., Wei, L., & Berry, M. W. (2005). Gene clustering by latent semantic indexing of MEDLINE abstracts. Bioinformatics, 21(1), 104-115.

[9] Beebe, Nicole Lang, and Jan Guynes Clark. "Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results." Digital investigation 4 (2007): 49-54.

[10] Inouye, David, and Jugal K. Kalita. "Comparing twitter summarization algorithms for multiple post summaries." Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. IEEE, 2011.

[11] Maletic, Jonathan, and Naveen Valluri. "Automatic software clustering via latent semantic analysis." Automated Software Engineering, 1999. 14th IEEE International Conference on.. IEEE, 1999.

[12] Yeh, J. Y., Ke, H. R., Yang, W. P., & Meng, I. H. (2005). Text summarization using a trainable summarizer and latent semantic analysis. Information processing & management, 41(1), 75-95.

[13] Bradford, Roger B. "An empirical study of required dimensionality for large-scale latent semantic indexing applications." Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008.

[14] Kontostathis, April. "Essential dimensions of latent semantic indexing (lsi)."System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on. IEEE, 2007.

[15] Elberrichi, Z., Rahmoun, A., & Bentaallah, M. A. (2008). Using WordNet for Text Categorization. Int. Arab J. Inf. Technol., 5(1), 16-24.

[16] Beil, Florian, Martin Ester, and Xiaowei Xu. "Frequent term-based text clustering." Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.

[17] Theodoridis, S. and K. Koutroumbas (2008). Pattern Recognition, Fourth Edition, Academic Press.

[18] Tata, S., & Patel, J. M. (2007). Estimating the selectivity of tf-idf based cosine similarity predicates. ACM Sigmod Record, 36(2), 7-12.

[19] Rajan Chattamvelli, Data Mining Algorithms, Published by Alpha Science International Ltd., 2011

[20] Dhillon, I. S., & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. Machine learning, 42(1-2), 143-175.

[21] Sobh, I., Darwish, N., & Fayek, M. (2006). A trainable Arabic Bayesian extractive generic text summarizer. In Proceedings of the Sixth Conference on Language Engineering ESLEC (pp. 49-154).

[22] Takçı, H., & Güngör, T. (2012). A high performance centroid-based classification approach for language identification. Pattern Recognition Letters,33(16), 2077-2084.

[23] Alqabas. (2016, October). Retrieved from http://www.alqabas.com.kw/Default.aspx

# Appendix:

# Journal & Conference Papers

## Conference Papers:

1. [**Accepted**] Al-Anzi, Fawaz S., and Dia AbuZeina ,“ Enhanced Latent Semantic Indexing for Arabic Language: In the Domain of Big Data Analytics”, ICAIAME 2019, International Conference on Artificial Intelligence and Applied Mathematics in Engineering 2019, Antalya, Turkey, 20-22 April 2019.
2. [**Accepted**] Al-Anzi, Fawaz S., and Dia AbuZeina, “Enhanced Search for Arabic Language Using Latent Semantic Indexing (LSI)”, 2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC, Mon Trésor, Plaine Magnien, Mauritius, 6-7 December 2018.

## Journal Papers:

1. [**Submitted**] Al-Anzi, Fawaz S., and Dia AbuZeina, “Enhanced Latent Semantic Indexing Using Cosine Similarity Measures”, The International Arab Journal of Information Technology, (Submitted).

# Submitted Paper(s)

# [IAJIT] [IAJIT] Paper submission

IAJIT <iajit@ccis2k.org>

Tue 25/12/2018 13:34

**To:** Fawaz Alanzi <fawaz.alanzi@ku.edu.kw>

Dear Author,https://emea01.safelinks.protection.outlook.com/?url=http%3A%2F%2Fwww.ccis2k.org%2Fiajit%2Fopenconf%2Fchair%2F&data=02%7C01%7Cfawaz.alanzi%40ku.edu.kw%7Ccda0f9ac65eb4e5523fb08d66a5493a1%7Cf9258092e3624609bea875884d326920%7C0%7C0%7C636813308831081765&sdata=uPS9%2FyEiDjA1xP4W1SLOF5t2GSZ5B1IV8aDTF6SQTVY%3D&reserved=0
Thank you for submitting your paper to The International Arab Journal of Information Technology.

Your Paper with id 18141 entitled Enhanced Latent Semantic Indexing Using Cosine Similarity Measures has been submitted to our system. You are able to track its progress via logging into https://emea01.safelinks.protection.outlook.com/?url=http%3A%2F%2Fiajit.org%2Fopenconf%2Fauthor%2Fsignin.php&data=02%7C01%7Cfawaz.alanzi%40ku.edu.kw%7Ccda0f9ac65eb4e5523fb08d66a5493a1%7Cf9258092e3624609bea875884d326920%7C0%7C0%7C636813308831081765&sdata=sEiNl0%2B3WSWUR8JKJIc28xQPNgsUXjGJC3TIvtD0iUA%3D&reserved=0.
The initial process will take about 4 weeks before the paper assign to the reviewers to ensure IAJIT standard and scope.

If you have any problem please contact us at iajit@ccis2k.org.

IAJIT Secretariat
https://emea01.safelinks.protection.outlook.com/?url=www.iajit.org&data=02%7C01%7Cfawaz.alanzi%40ku.edu.kw%7Ccda0f9ac65eb4e5523fb08d66a5493a1%7Cf9258092e3624609bea875884d326920%7C0%7C0%7C636813308831081765&sdata=3sxq0mqk5K5gcHUIeRhnFaZ3MtCgbZC%2FgholMSFvnjw%3D&reserved=0
Tel: 00962-5-3821100 ext. 1452

IP Address: 139.141.11.139

# CONFERENCE PAPER(S)

Society of Information Technologists and Entrepreneurs
Address: Port Louis, Mauritius
Website: https://sitemauritius.wordpress.com
Email: sitemauritius@gmail.com
Tel.: (+230) 213 0032
Registered Number: 14128

# Invitation for Participation in the 2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC)

01 November 2018

Dear **Fawaz Al-Anzi**,
Kuwait University
Kuwait

The **Society of Information Technologists and Entrepreneurs (SITE)**, in collaboration with the **IEEE Mauritius Subsection** and the Organising Committee of **Mauricon ICONIC 2018** have the immense pleasure to inform you that, **after a double-blind peer review process**, your paper entitled *'Paper 112: Enhanced Search for Arabic Language Using Latent Semantic Indexing (LSI)"* has been accepted for oral presentation at the **2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC)** which will be held at Holiday Inn, Mon Trésor, Plaine Magnien, Mauritius on the 6th and 7th of December 2018. The paper will be submitted for potential inclusion in the **IEEE Xplore Digital Library** if all formalities and guidelines are followed. The IEEE conference number is **#44423** and the **ISBN** of the proceedings is **978-1-5386-6477-3**.

This letter is to officially invite you to attend the conference and give an oral presentation of your research and to communicate with other researchers and scholars from around the globe who have similar interests.

If you have any queries about the conference, please email us on mauricon2018@gmail.com or secretariat@mauricon.org.

Looking forward to seeing you at the conference.

Thank and regards.

Society of Information Technologists & Entrepreneurs
www.sitemauritius.org
Email:sitemauritius@gmail.com
Port Louis, Mauritius

Hoshiladevi Ramnial
Secretariat
Republic of Mauritius

ICAIAME'19
ICAIAME 2019 / UMYMK 2019 / 20-22.04.2019

**Dear Fawaz Al-Anzi;**
*Kuwait University, Kuwait*
*Fawaz.alanzi@ku.edu.kw*

**After the three-reviewer based evaluation process, your paper titled as *"Enhanced Latent Semantic Indexing for Arabic Language"* has been accepted for 'oral presentation' at ICAIAME 2019 (International Conference on Artificial Intelligence and Applied Mathematics in Engineering 2019), which will be held Antalya, Turkey (20-22 April 2019). Congratulations!**

**The paper will be published under the Springer Series: Lecture Notes on Data Engineering and Communications Technologies (There is also an opportunity of suggesting selected papers to some international journals, after the end of the event).**

**Thank you very much for your great contribution to the ICAIAME 2019.**

**Prof. Dr. Tuncay YİĞİT**
**Conference President**

**Paper Title:** Enhanced Latent Semantic Indexing for Arabic Language
**Paper ID:** 2
**Authors:** Fawaz Al-Anzi, Dia Abuzeina

# ACKNOWLEDGEMENT

*I hereby acknowledge the support of Kuwait University Research Sector in granting the Project and facilitating the research implementation.*

*And I also agree to the best of my knowledge that the information herein are true and complete.*

**Principal Investigator (Applicant)**

**Name:** FAWAZ S. AL-ANZI

**Project Code:** EO03/18

**Faculty:** FACULTY OF ENGINEERING & PETROLEUM

**Department:** COMPUTER ENGINEERING

**Date:** 29/10/2019

**KUWAIT UNIVERSITY**
***RESEARCH SECTOR***

**Name: Fawaz S. Al-Anzi**

**Date: 29/10/2019 03:01:01 p.m.**