إستخدام تقنيات استرجاع المعلومات في مطابقة الأسماء العربية في قاعدة البيانات الوطنية

د. عبدالسلام منصور الشريف
جامعة طرابلس ، كلية تقنية المعلومات، طرابلس ، ليبيا
a.abdoessalam@uot.edu.ly

مها شعبان عمارة
جامعة طرابلس ، كلية العلوم، طرابلس ، ليبيا
eng.mhaomara@gmail.com

أ.صلاح خيري البسكري
المعهد العالي للمهن الشاملة ، طرابلس ، ليبيا
Eng.salaheddin@gmail.com

الملخص— هذا البحث جاء نتيجة للبحث عن حلول للمشاكل القائمة في مطابقة الأسماء وذلك لإكتشاف التكرار في بيانات المواطنين في الدوائر الحكومية والذي من شأنه تحسين أداء هذه الدوائر ومكافحة تكرار بيانات المواطن والذي ينتج عنه في كثير من الأحيان ضياع المال العام.

رغم أن استخدام الرقم الوطني في معاملات الدولة الليبية قلل بشكل كبير في عملية التكرار الواردة لمؤسسات الدولة ، إلا انه في بعض الحالات والتي لا يتم فيها الربط بقاعدة البيانات الوطنية تحدث الكثير من الأخطاء والتي من شأنها تضييع حقوق المواطنين أو استغلالها في تكرار البيانات.

ولهذا السبب قام الباحثون بالعمل على تحسين بعض الخوارزميات الخاصة بالبحث وذلك لاستخدامها في اكتشاف هذه التكرارات. ففي هذا البحث تم التركيز على تحسين أحد طرق البحث في النصوص وهي الطريقة التي تستخدم نظرية Edit Distance في عملية مطابقة النصوص العربية. تم في هذا البحث اختيار الأسماء العربية نظرا للمشكلة التي تعاني منها المؤسسة قيد البحث وهي مطابقة الأسماء العربية المجمعة بعدة طرق ببيانات قاعدة البيانات الوطنية.

الكلمات المفتاحية : مطابقة النصوص ، تحسين معدل الإسترجاع ، الأسماء العربية.

1. المقدمة

إن استخدام التقنية الحديثة في المجالات العملية هو الهدف الأسمى لمعظم الإختراعات والإكتشافات في عالمنا. فلطالما كانت المؤسسات الحكومية منها والخاصة تدفع أموال طائلة للأبحاث بشتى أنواعها لكي تقوم بحل المشاكل التي تواجهها. ومن أهم التقنيات العملية المستخدمة في الحاسوب هي تقنية استرجاع المعلومات. ولعل من أشهر المواقع على الانترنت هو موقع **Google** والذي يعتبر الملاذ الأول لأي بحث في الانترنت. فيعتبر هذا الموقع الأكثر استخداما للبحث عن أي معلومة في الانترنت. إن هذا الموقع يدار بتقنيات تعتمد على آخر ما توصلت إليه الأبحاث في مجال تقنيات استرجاع المعلومات.

يعمل نظام استرجاع المعلومات على الإستجابة لطلب المستخدم ، بتقديم الوثائق المرجعية حتى يتم الحصول على المعايير المطلوبة. هنالك عوامل مختلفة تعين ما اذا كانت الوثائق ذات صلة بسؤال المستخدم أم لا. وأحد هذه العوامل هى الوثائق نفسها ، مجال محتوياتها وكيفية كتابتها ، ويعتبر المستخدم ايضا عامل مهم لان الوثائق ذات الصلة باهتمام المستخدم تعتمد على حكم المستخدم او قرارته.

إن انظمة استرجاع المعلومات قد لا تكون دقيقة النتائج 100% ؛ ولكن ماذا يستطيع نظام استرجاع المعلومات عمله؟ إن باستطاعة أنظمة استرجاع المعلومات تقدير فعالية النتائج ومدى مقابلتها لمتطلبات المستخدم المؤسسة على اسئلته [1]. ولكن هذه العلاقة صعبة المعايرة ، فعدة عوامل تستطيع أن تُعين نتائج ذات صلة ولكن أكثرها هو رضى المستخدم. فاذا كان المستخدم يريد الوثيقة فالوثيقة فهي ذات صلة وإلا فلا ، لذلك يحاول الباحثون باستمرار ايجاد طرق حديثة لتعريف اهمية الوثائق لسؤال المستخدم.

إن أحد هذه التحديات التي تواجه الباحثين عندما يتعاملون مع النص العربي هو التطبيع (Normalization). فالعربية الفصحى الحديثة (Modern Standard Arabic) (MAS) عندها بعض التناقض فيما يتعلق بالتشكيل في الكلمة ، فالكلمة قد يكون لديها واحد او اثنين من علامات التلفظ. ونظرا لهذا الإختلاف في التلفظ لزم أن نقوم بالتطبيع وذلك بتحويل النص الممنهج الى شكل متحد. مثلا حرف (ا) لديه اشكال مختلفة (أ، إ، آ) معتمدة على مكان علامة (ء) تحت او فوق الحرف.

بدون التطبيع هذه الأشكال تُعامل مختلفة في الكلمة ، خاصة في نماذج استرجاع المعلومات مع أشكال نصية مختلفة. إن بعض النصوص تستعمل الفعل (إقرا) في نماذج مختلفة (أقرأ ،إقرأ،اقرا). فبدون التطبيع ، فإن نظام استرجاع المعلومات غير قادر على استرجاع نماذج مختلفة للكلمة الواحدة. وهذا من شأنه التقليل من دقة نظام استرجاع المعلومات ، كما سيزيد التطبيع من جهة أخرى مستوى الغموض ، مثلا بالتطبيع ستعامل الكلمات بنفس الشئ حتى ولو كانت ذات معاني مختلفة [2].

فمثلا كلمة (الشَعر) يصعب تمييزها عن كلمة (الشِعر) ، إن نظام استرجاع المعلومات الذى ينفذ التطبيع سوف يسترجع كل الوثائق التى تحتوي على الكلمات الغامضة بدون اعتبار للمعنى ، مثال اخر كلمة (كتب) لديها معنيين قد تعني فعل ماضى (كَتَبَ) او جمع (كُتُب) ، نتيجة لذلك فإن عدة كلمات عربية غامضة ستظهر تحدي مختلف لنظام استرجاع المعلومات. لايزال معنى تطبيع الكلمات يمكن استدلاله من قطعة او من نص ، إن احدى التقنيات التى تم اقتراحها لحل مشكلة الغموض هى تقنية الإحساس بعدم الغموض (Word Sense Disambiguation technique) و هذه التقنية تستعمل المعاجم والمفردات لاستخلاص المعاني من الكلمات المجاورة [3].

في الجزء 2 سيتم عرض الدراسات التي قامت على الخوارزمية المستهدفة. والجزء 3 يقدم العمل الذي أنجز في هذا البحث وطريقة التحسين ، بينما يعرض الجزء 4 النتائج المتحصل عليها بعد التعديلات المقترحة وتحليلها. وأخيراً يعرض الجزء 5 التوصيات والأعمال المستقبلية المكملة لهذا العمل.

2. الدراسات السابقة

أ. خصائص اللغة العربية

اللغة العربية واحدة من اقدم اللغات في العالم وهى من اللغات الرائدة اليوم وهى تنتمى الي عائلة اللغات السامية والتى تحوي الاكادية و الارامية و الاثيوبية والعبرية والفينيقية والسريانية و اوغاريتية [4]. إن اللغة العربية هى اللغة الرسمية في الوطن العربي ولغة الدين الاسلامى في العالم حيث انها لغة الكتاب الكريم (القرآن) اضافة الى ذلك هى من ست لغات الرائدة المعترف بها من قبل الامم المتحدة .

يمكن تصنيف اللغة العربية الي لغة كلاسكية (**Classic Arabic CA**) و لغة عربية نموذجية (**Modern Standard Arabic (MSA)**). فاللغة

الكلاسيكية هى اللغة العربية القديمة المستعملة في القران والكتب الدينية الاخرى . واللغة العربية النموذجية هى لغة الكتابة الشائعة والرسمية في التعامل في العالم العربي اليوم وهى تستعمل في المدارس والاعلام والمجلات والمجلات المنهجية والجرائد [3].

إن اللغة الكلاسكية هى لغة ثابتة ولها نمط معين والفاظ ولها حصانة ثابتة لأي تغيرات حادثة لأكثر من 15 القرن. يستطيع العرب اليوم فهم اللغة الكلاسكية بسهولة نسبيا [2]. وفي نفس الوقت تعتبر العربية highly diglossic فالبلد العربي تجد له ثلاثة او اربع من لغات محدثة مختلفة تستعمل في غير الحالات الرسمية وكل تشكيل غير رسمي له تهجئة والفاظ خاصة به [5] .

إن للغة العربية 29 حرف تبدأ بالهمزة (مثل الحرف a في اللغة الانجليزية) وهى في بعض الحالات تتصرف بعلامات صوتية مميزة انظر جدول (1) والذي ويوضح نطق الحروف العربية بالانجليزي. يكتب النص العربي افقيا من اليمين الي اليسار بينما الارقام تكتب من الشمال الي يمين. إن الحروف العربية تخضع لتحويرات قليلة عندما تجمع في كلمة واحدة ، فبعض الحروف قد يكون لها شكل أو أكثر على حسب موقعه في الكلمة. بينما الحروف الاخرى قد يكون لها شكل واحد فقط مثلا الهمزة تكتب بنفس الشكل في أي موضع من الكلمة وبعض الحروف لها اثنين او ثلاثة او اربعة اشكال مثل حرف العين (ع)(عـ ،ع، ـعـ ،ـع).

جدول 1. الحروف العربية ونطقها بالحروف الانجليزية

ر	ذ	د	خ	ح	ج	ث	ت	ب	أ
/raa/	/thal/	/dal/	/khaa/	/haa/	/jeem/	/thaa/	/taa/	/baa/	/alif/
r	th	D	kh	h	j	th	t	b	a
ف	غ	ع	ظ	ط	ض	ص	ش	س	ز
/faa/	/ghain/	/'ain/	/thaa/	/taa/	/thad/	/sahad/	/sheen/	/seen/	/zaa/
f	gh	`	d	t	d	s	sh	s	z
	ء	ي	و	ه	ن	م	ل	ك	ق
	/a/	/yaa/	/waw/	/haa/	/noon/	/meem/	/laam/	/kaaf/	/qaaf/
	a	Y	o	h	n	m	l	k	q

تتكون الابجدية العربية من الحروف الساكنة. وهي علامات الحروف المتحركة تكتب فوق او تحت الحروف الساكنة. إن حروف العلامات الصوتية المميزة مثل الضمة والفتحة والكسرة والشدة تتحكم في تهجي الكلمة ؛ فاي استعمال غير صحيح لهذه العلامات يمكن ان ينتج كلمة ذات معنى دلالي مختلف. إن عملية كتابة هذه العلامات أعلي او تحت الحرف الساكن لمعرفة التهجئ الصحيح يدعى التلفظ. تعتبر اللغة العربية لغة غنية ومرنة نتيجة لخصائص الصرفية (morphological characteristics) فالعشرات والمئات من الكلمات يمكن اشتقاقها من نفس الجذور ولهذا السبب فإن العربية لها ثلاثة اضعاف كلمات اللغة الانجليزية وتقريبا 5 مليون كلمة اصلها من 11 الف واربعمائة جذر.منها الف ومئتان جذر تستعمل فعليا في لغة العربية الحديثة (MSA) [3].

يتكون الكلام العربي في معظمه من الأفعال والأسماء حيث الأفعال تأتي عادة قبل الأسماء. إن الأفعال العربية لها زمنين الماضي وغير الماضي ولها لفظان صوتيان هما المبنى للمعلوم والمبنى للمجهول . لدى الأسماء في العربية 3 حالات من القواعد (الاسم والمضاف اليه والمفعول به) والجنسان (مذكر ومؤنث) .مثلا المؤنث من كلمة العربية "المعلم" هو"المعلمة" والمذكر من "المعلم" هو "المعلم" ، وللاسم في العربية 3 حالات هي المفرد والمثنى والجمع .

إن لحروف اللغة العربية أشكال مختلفة وتغيير إضافي يتم إضافته بالكتابة المختلفة بالمحادثة. فمثلا عند كتابة حرف ي الذى يظهر في نهاية الكلمة يتم استبداله بحرف (ى) .مثال اخر كحرف (ا) والذى يكتب (ا،أ،إ،آ) العديد من الكتاب يكتبون (ا) مجردة بينما اخرون يكتبونها بالتشكيل المناسب. هذا يسبب في أن تظهر الكلمة مختلفة ولديها مجموعة من حروف الترميز المختلفة مثلا (أشرب) و (اشرب) هى نفس الكلمة و لكن التهجي يختلف.

مثال أخر حرف (ة) تكتب (ه). يتم استعمال التشكيل قليلا في النص العربي لذا يتم ازالته لتوحيد الشكل الصوتي وغير الصوتي. إن معظم نظم استرجاع المعلومات العربية تقوم بعملية ماقبل التجذير و تطبيع النص العربي لتوحيد أسلوب كتابة النص العربي.

ب. تطبيع النص في اللغة العربية

اول خطوة في عملية التطبيع هي ازالة التشكيل وعمليات الترقيم وبعض الحروف غير العربية و الخطوة التالية هى تطبيع الاخطاء المطبعية المكتوبة كما موضح في الجزئية التالية :
يتعرض النص العربي لاساليب مختلفة من الكتابة والاخطاء الشائعة مثلا اقتران حرف الواو (و) و حرف الهمزه (ء) كلمة (المسؤل) تكتب فوق الواو الهمزه وتكتب (المسوءل) بجنب حرف الواو في الكلمة .وايضا اقتران حرف (ى) وحرف الهمزه (ء) تكتب مختلفة باختلاف الكتاب في بعض الكلمات تكتب كحرف واحد (ئ) وفي مواضع اخر كحرفان منفصلان (ىء). كما ان الحرف (ا) يكتب عادة في بداية الكلمة وفي بعض الحالات يتم اعادته اكثر من مرة وهذا يتعارض مع قوانين الصرف والكتابة العربية.

كما أن الحرف (ة) يمكن ان يظهر فقط في نهاية الكلمة اذا كان الفراغ بين الكلمة المنتهية بهذا الحرف والكلمة التى تليها قد يحذف عادة حرف (ة) ولا يأثر الحرف الذى يليه . يبدو ان الحرف (ة) هى الكلمة الوسطية مثلا (قناةالجزيرة) هى في الحقيقة مركب من كلمتان الاولي تنتهي عند حرف (ة) . بمثل الحروف ر،و،ز،ي،د،ذ في كثير من الاحيان يربط الكلمة بدون اي فراغ .يستطيع الانسان تمييز هذه الكلمات بدون اى مشكلة ولكن يجب تكييف برامج استرجاع الالي لمعرفة هذه الحالات [6].

ت. التنوع في الكتابة باللغة العربية

يمكن مقارنة الحروف العربية مع بعضها البعض بثلاثة اوجه للتشابه ويمكن القول كم هى متشابهة مع بعضها البعض وفقا لمظهر التشابه [7] ومن هذه المظاهر :

1- تشابه الاحرف شكليا (**The Similarity of Letter's Form**)
فالحروف العربية تكون متصلة مع بعضها البعض في النص المطبوع بينما في اللغة الانجليزية تكتب الحروف منفصلة. إن الحروف في اللغة الانجليزية تكتب من اليسار الي اليمين شكليا ولكن في الحروف العربية تكتب من اليمين الي اليسار في شكليتها.

في اللغة العربية هناك 22 حرف من 28 حرف لديه 4 متغيرات للحرف مثل حرف (هـ) و حرف (ف) انظر الى الجدول (2) .

جدول 2. صيغ كتابة بعض الحروف

هـ	ﻫ	ﻬ	ه
غـ	ﻏ	ﻐ	غ
فـ	ﻓ	ﻔ	ف

عليه يمكن القول إن للحرف العربي 4 أشكال مختلفة :
1- الحرف يقف لوحده.
2- كأول حرف في الكلمة.
3- في داخل الكلمة بين حرفين آخرين.
4- الحرف الأخير في الكلمة

يعين شكل الحرف وفقا لوضعه في الكلمة مثلا حرف (غ) يكتب في أول الكلمة (غـ) وفي وسط الكلمة (ـغـ) وفي نهاية الكلمة (ـغ) و(غ) عندما يكون معزول لوحده. أما في بقية الحروف 6 للحروف العربية أ،و، د،ذ،ر،ز فلديها متغيران اثنان أما أن يكون الحرف في أول الكلمة وإما أن يكون الحرف في أخر الكلمة. إن هذه الحروف لا يمكن ربطها بالحرف التابع لها حتى وان كانت داخل الكلمة. هذا يعني بان الكاتب عليه أن يرفع قلمه او حتى يده داخل نفس الكلمة.

في دراسة قامت بتطبيق Edit Distance على اللغة العربية [7] تم تقسيم الأحرف إلى مجموعات تتشابه في شكل الحرف. فالجدول (3) يقسم الحروف المتشابهة في الشكل إلى 6 أصناف يحتوي كل صنف على مجموعات متشابهة من الحروف وتم إعطاء كل صنف وزنا يدل على مدى التشابه. يوضح الجدول مجموعات من الحروف العربية لكل صنف وفقا لفهرسة التشابه ابتدأ من التشابه 100 % الي التشابه 0 %. ولقد تم في هذا البحث استخدام نفس الجدول ولكن بقيم مختلفة أخذت من البيانات الحقيقية وتم تطبيق نفس الخوارزمية على نفس البيانات لإثبات وجود التحسين.

جدول 3. مجموعات الحروف المتشابهة في الشكل.

No	Similar group	Similarity index
1	{ا-أ-إ-آ},final form of{ة-ه}and{ي-ى}	**1**
2	{ء-آ}،{ء-ا}،{ء-أ}،{ء-إ}،{و-ؤ}،{ى-ئ}	**0.8**
3	{ع-},{ر-ز},{ذ-د},{ج-ح-خ},{ث-ن},{ت-ث},{ب-ن}, {ص-ض},{ظ-ط},{ق-ف},{غ-}, initial and medial form of {ب-ي}	**0.6**
4	{ش-س}،{ث-ب}, initial and medial form of {ت-ي},{ث-ي}	**0.4**
5	{هـ - م}،medial form of{ك-ل} Final form of	**0.2**
6	Any other combination of Arabic letters	**0**

2- التشابه الصوتي (**The Similarity of Phonetic**)

إن اللغة العربية غنية بالمصطلحات وذلك لمقدرتها الصوتية حيث تتهجي الحروف العربية في حوالي 17 وجه مختلف لذلك فأي انحراف في التهجئ سيقود الي تغير في المعنى. يعتمد التهجئ على انواع اللغة العربية مثل اللغة العربية الكلاسيكية و الحديثة والعامية. مثال على ذلك في العامية المصرية حرف (ق) يلفظ مثل حرف (ا) مثل (قدري) تصبح (ادري) ايضا في العامية المصرية العليا تلفظ (جدري). ويمكن تصنيف الحروف العربية الى 6 اصناف وفقا للتشابه الصوتي في الجدول (2) الذى يوضح مجموعة من الحروف العربية لكل صنف وفقا لفهرسة التشابه ابتدأ من التشابه 100% الي التشابه 0% [7].

3- التشابه في لوحة المفاتيح (**The Similarity of keyboard**)

يرتبط التشابه في لوحة المفاتيح بحرفان مع بعضهما البعض وفقا لقرب المفاتيح في لوحة المفاتيح وهذا يقود الى اختلاف في الحرف ويمكن النظر الى لوحة المفاتيح العربية والتى تستعمل نافذة مايكروسوفت وكيفية توزيع الحروف عليها. ويكون المشغل اكثر احتمالية لفعل الخطأ على لوحة المفاتيح لقرب الحروف من ناحية اخر التشابه لمفتاحين يزيد اذا ما كانا قريبان على لوحة المفاتيح.[7]

ث. تطابق الأسماء العربية

تطابق الأسماء هى عملية لتعيين ما اذا كان الاسمان متطابقان. هناك طرق متعددة تم تطويرها لملائمة الاسماء والتى تعكس العدد الكبير للاخطاء والتحورات التي قد تحدث في البيانات اليومية المعتادة. إن الهدف الاساسي لهذه التقنيات هو ملائمة او مطابقة الاسماء او السلاسل التى تكون متماثلة أو متشابهة بدلا من التطابق التام (exact match) ، فمقياس التشابه عادة يتم حسابه للحصول على قيمة بين 0 و 1 (1 - عندما تكون الاسماء متطابقة) ، (0 - عندما تكون الاسماء مختلفة كليا). وهناك عدة طرق معروفة لتقدير التشابه بين السلاسل والتى نستطيع فصلها الى مجموعتان هما : التقنية الحرفية (**character-based techniques**) والتقنية الرمزية (**token-based techniques**) **[8]**.

تعتبر خوارزمية ليفينشتن (Levenshtein algorithm) [9] ومتغيراتها تعتمد على التقنية الحرفية والمؤسسة على مقاييس مسافة التحرير (edit distance metrics) و مسافة التحرير لليفينشتن تعرف أساسا بتطابق سلسلتان ذات أطوال مختلفة (arbitrary lengths). انها تحسب اقل اختلاف بين السلاسل على مبدأ عدد المدخلات والمحذوفات او المستبدلات المطلوبة لتحويل سلسلة واحدة إلى أخرى. فالصفر يبين التلائم التام. إن طريقة ليفينشتن الأساسية تم تمديدها لعدة اتجاهات [10] ؛ فمثلا لدينا تمديد لاعتبار التنظيم المنعكس (تبديل الحروف transposition of characters) مباشرة في عملية مسافة التحرير. اتجاه اخر للتعميم هو بالسماح لأوزان مختلفة عند مستوى الحرف. يمكن أن تعتمد أوزان إستبدال الحروف على لوحة المفاتيح او التشابه الصوتي [11]. في الفقرة القادمة نقدم طريقة عملية لتحسين أداء خوارزمية مسافة التحرير المعتمدة على تشابه الحروف العربية.

3. خوارزمية مسافة التحرير المطورة

للعمل على تحسين الخوارزمية تم إجراء التجارب اللازمة وكانت على النحو التالي :

1- تجهيز البيانات :

تم تنفيذ هذه الدراسة على بيانات حقيقية للأسماء العربية والتي تتكون من فئتين : الفئة الأولي تمثل أسماء مواطنين تم تجميعها عن طريق مركز معلومات مصلحة العمل ، والتي تم إدخالها يدويا للحاسوب. أما الفئة الثانية فتمثل بيانات نفس المواطنين في قاعدة البيانات الوطنية. تم إزالة كل البيانات المطابقة 100% والتي يتطابق فيها الإسم في كلا الفئتين. بحيث تم إختيار 324209 اسم من كلا الفئتين لا يتطابقان بشكل كامل.

2- إيجاد الأوزان المناسبة :

في هذه المرحلة تم إجراء خوارزمية AEDA على البيانات وذلك للتحصل على مدى مطابقة الأوزان في الجدول (3) للبيانات قيد التجربة. وبعد إجراء التجارب تم تعديل جدول أوزان المجموعات لينا سب والبيانات قيد التحليل. وكان ذلك بتغيير بعض المجموعات والتي لوحظ فيها أن معدل تكرارها (أي وزنها) أكبر من مما هو ممثل في الجدول الأصلي. والمجموعات هي :

أ. المجموعة (ب-ي) تم ترفيعها الى الصنف 2 بدلا من 3.
ب. المجموعة (ه-م) تم ترفيعها الى الصنف 3 بدلا من 5.
ت. المجموعة (ت-ي) تم ترفيعها الى الصنف 3 بدلا من 4.
ث. المجموعة (ك-ل) تم ترفيعها الى الصنف 3 بدلا من 5.

ولقد تمت المحافظة على باقي المجموعات في مكانها حتى ولو كانت أقل في تصنيفها وذلك لتوضيح الفرق الحقيقي وذلك لأن هذه التغييرات تمت على بيانات خاصة وليس نص عام. والذي من شأنه أن يقلل أهمية النتائج في حالة إنقاص أوزان المجموعات الأخرى.

3- إجراء التجارب النهائية

في هذه المرحلة تم إجراء الخوارزميات قيد المقارنة على البيانات وذلك لتبيين الفرق بين أدائها. ولقد تم حساب الدقة وهي مجموع ما تم استرجاعه من البيانات الصحيحة من مجمل البيانات التي تم استخدامها. علما بأن المجموع الكلي للبيانات هو 324209 اسم.

جدول 4. يوضح فرق الإسترجاع بين الخوارزميتين

M.D	ADEA	Modified	Diff.
0	157846	157846	0
1	242756	242843	87
2	276070	276192	122
3	295585	295684	99
4	321717	321793	76

4. النتائج والتحليل

يوضح الجدول (4) الفرق أو التحسين الذي طرأ على الخوارزمية **ADEA** وذلك بعد التعديل الذي تم على الأوزان لبعض المجموعات. فيمثل العمود **M.D** مسافة التحرير المسموح بها بينما يمثل **ADEA** عدد مرات الإسترجاع الصحيح للبيانات ، ويمثل **Modified** عدد مرات الإسترجاع الصحيح للخوارزمية المحسنة. ويتضح أن في حالة أن تكون مسافة التحرير صفر فإن التحسين الذي طرأ على الخوارزمية ليس له أي تأثير نظرا لأنه تم إلغاؤه. وذلك أنه لم يتم ترفيع أي مجموعة من صنف الى الصنف 1.

في باقي الحالات والتي يسمح فيها بإستخدام المجموعات المرفعة وأوزانها الجديدة فنجد التباين بين خوارزمية **ADEA** وخوارزمية **Modified**. وهذا يرجع لأن الأوزان مختلفة بين الخوارزميتين. كما يلاحظ أنه كلما ازدادت قيمة مسافة التحرير كلما اقترب نتائج الخوارزميتين لبعضهما. علما بأن زيادة قيمة مسافة التحرير سيشكل تناقصا في معدل الأداء ؛ فالقاعدة تقول إنه كلما قلت مسافة التحرير كلما دقت النتائج وكلما كبرت مسافة التحرير المسموح بها كلما تباعدت النتائج المرجعة.

5. الخلاصة والعمل المستقبلي

لقد تم في هذه الورقة إستعراض النتائج الجزئية لطرق تحسين خوارزمية **Levenshtein distance** المعدلة للغة العربية **ADEA** والتي تستخدم الخوارزمية الأصلية مع تعديلها لتتناسب والكشف عن النصوص المتماثلة. ولقد تم التعامل مع خاصية تماثل أشكال الحروف. والتي بعد تحسينها بتغيير الأوزان الخاصة بمجموعات الحروف في المجموعة الواحدة أعطت نتائج أفضل من نتائج الخوارزمية الغير محسنة.

سيركز العمل المستقبلي على إستخدام المجموعات الأخرى ومحاولة تحسينها حتى تتلائم والبيانات قيد البحث. وخصوصا أن إحدى المجموعات التي تُعنى بالصوتيات والتي عادة ما تختلف من لهجة عربية الى أخرى كما بينت الورقة المقدمة لخوارزمية **ADEA**. ومن المتوقع أن يكون تحسين أداء هذه الخوارزمية أفضل مع استخدام المجموعات كلها ، كما من المتوقع أن خصوصية نطق بعض الحروف باللهجة الليبية والمدخلة بها البيانات المستعملة سيكون له أثر في هذا المجال وخصوصا أن الخوارزمية الأصلية كانت مرتكزة على العامية المصرية.

6. المراجع

[1] Raghavan, Vijay V., and SK Michael Wong. "A critical analysis of vector space model for information retrieval." Journal of the American Society for information Science 37.5 (1986): 279-287.

[2] Farghaly, Ali and KhaledShaalan. 2009. Arabic natural language processing: Challenges and solutions. ACM Transactions on Asian Language Information Processing, 8(4) article 14.

[3] Al-Maimani, Maqbool R., A. A. Naamany, and Ahmed Zaki Abu Bakar. "Arabic information retrieval: techniques, tools and challenges." GCC Conference and Exhibition (GCC), 2011 IEEE.IEEE, 2011.

[4] Moukdad, H., ―Lost In Cyberspace: How Do Search Engines Handle Arabic Queries?‖ In: Access to Information: Technologies, Skills, and Socio-Political Context. University of Manitoba, Winnipeg, Manitoba. June 3 - 5, 2004.

[5] Albalooshi, Noora, Nader Mohamed, and Jameela Al-Jaroodi. "The challenges of Arabic language use on the Internet." Internet Technology and Secured Transactions (ICITST), 2011 International Conference for. IEEE, 2011.

[6] A. F. Nwesri, S. Tahaghoghi, and F. Scholer, "Finding Variants of Out-of-Vocabulary Words in Arabic," inProceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues andResources, pp. 49–56, Association for Computational Linguistics, 2008.

[7] H. H. A. Ghafour, A. El-Bastawissy, and A. F. A. Heggazy, "AEDA: Arabic Edit Distance Algorithm TowardsA New Approach for Arabic Name Matching," in IEEE International Conference, IEEE Trans. Pattern Analysisand Machine Intelligence, vol. 15, pp. 926–932, 2011.

[8] Elmagarmid, A., Ipeirotis, P., &Verykios, V. (2007). Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering, 19(1), 1-16.

[9] Algorithms and Theory of Computation Handbook, CRC Press LLC, 1999,"Levenshtein distance", http://www.nist.gov/dads/HTML/Levenshtein.html in Dictionary of Algorithms and Data Structures, Paul E. Black, ed., U.S. National Institute of Standards and Technology, 14 August 2008 (accessed 24 March 2017).

[10] Hall, P., & Dowling, G. (1980).Approximate string matching. ACM Computing Surveys (CSUR), 12(4), 381-402.

[11] Snae, C. (2007).A comparison and analysis of name matching algorithms. International Journal of Applied Science, Engineering and Technology, 4(1), 252-257.