دراسة تأثير مقاييس التشابه على خوارزمية K-Means في عنقدة النصوص العربية بالاعتماد على الكلمات المفتاحية

سهاد مهجر كريم

جامعة البصرة\ كلية العلوم \ قسم علوم الحاسبات

المستخلص **Abstract**

يعد تجميع (عنقدة) النصوص أحد الوسائل الهامة والفعالة في تنقيب النصوص ، حيث يهدف تجميع النصوص إلى تقسيم المجاميع الكبيرة للنصوص إلى مجاميع صغيرة تدعى تجمعات(عناقيد) تحتوي على كيانات تتشابه فيما بينها وتكون مختلفة عن الكيانات في التجمعات الأخرى. قدمنا في هذا العمل طريقة لتجميع النصوص العربية باستخدام إحدى تقنيات التجميع المشهورة والمتمثلة بخوارزمية k-means . تضمنت الطريقة تحليل النص كخطوة أولية لتهيئته إلى خوارزمية العنقدة التي طبقت على 100 نص عربي بأربع أصناف مختلفة شملت (رياضة ، فن ، جرائم ، طبية) ، حيث طورنا الطريقة باعتماد قاعدة بيانات من الكلمات المفتاحية الخاصة بكل مجال لاختيار مراكز التجميع بدلا من اختيارها بشكل عشوائي ثم استخدمنا مقياسين للتشابه هما (مقياس المسافة الاقليدية ومقياس تجييب الزاوية) لحساب المسافات بين مركز التجميع والمستندات لبناء العناقيد . قيمنا تاثير مقاييس التشابه (المسافة الاقليدية، تجييب الزاوية) على نتائج خوارزمية k-means باستخدام مقياس التقييم F-Measure وكانت النتائج عبارة عن مقارنة بين انجازيه الخوارزمية باستخدام مقياس تشابه المسافة الاقليدية ومقياس تجييب الزاوية وعلى عدد من العوامل منها عدد التجمعات وعدد الأصناف. وأخيرا لاحضنا بان انجازيه خوارزمية k-means باستخدام مقياس تجييب الزاوية هو افضل من انجازيتها باستخدام المسافة الاقليدية.

الكلمات المفتاحية **Keywords**

تنقيب النصوص ، تجميع النصوص العربية ، خوارزمية K-Means ، المسافة الاقليدية (Euclidian)، مقياس تجييب الزاوية (Cosine Similarity).

1. المقدمة Introduction

أصبح الحصول على المعلومات المفيدة من أهم وسائل النجاح والتطور في الدول النامية وعلى مختلف المجالات . وقد أدى التطور السريع الذي يشهده العالم اليوم والتضخم الكبير في حجم النصوص الالكترونية المتوفرة مثل المقالات والكتب والأخبار إلى زيادة التحديات وبالتالي الحاجة لإيجاد وسائل فعالة تُعنى بترتيب وتصنيف وتحليل المعلومات لمساعدة المستخدمين بإيجاد وتصفية وتنظيم تلك النصوص[1،2]. ظهر تنقيب النصوص كمجال جديد في علم الحاسوب الذي يتبنى دراسة الارتباطات القوية للنص باستخلاص المعلومات المفيدة من البيانات الغير منتظمة من خلال التعريف والاستكشاف للأنماط المهمة ، يوجد العديد من التقنيات المتنوعة في تنقيب النصوص منها تصنيف النصوص(text classification) ، عنقدة النصوص (text clustering) ، تلخيص النصوص (text summarization) وغيرها[3،4].

اخترنا في هذا العمل تقنية عنقدة النصوص وهي عبارة عن معالجة لتنظيم المستندات واستخلاص المفاهيم الأساسية للنص من خلال تقسيم البيانات (بإيجاد الهيكل الداخلي لها) إلى أجزاء من الكيانات المتشابهة تدعى العناقيد(clusters) بحيث تكون البيانات داخل العنقود(التجمع) الواحد تمتلك درجة تشابه فيما بينها، بينما يكون التشابه بين البيانات في العناقيد المختلفة قليل جدا. تقسم تقنيات العنقدة عادة إلى قسمين هما التقنية الهرمية (heretical clustering) وتقنية التجزئة (partition clustering). التقنية الهرمية تقسم البيانات المعطاة إلى أجزاء صغيرة في أسلوب هرمي أو شجري وذلك بإنتاج سلسلة متداخلة (nested) للأجزاء مع تجمع شامل في الأعلى وتجمعات مفردة للنقاط الثانوية في الأسفل، بينما تعتمد تقنية التجزئة على تجزئة تجمعات النصوص إلى مجموعة من العناقيد الغير متداخلة لكي تقلل من قيمة التقييم للتجميع. ونلاحظ أن هنالك فرق بين الطريقتين فعلى الرغم من أن التقنية الهرمية تعطي جوده أفضل إلا إنها لا تحتوي إمكانية إعادة تخصيص للكيانات الذي ربما يكون تصنيف سيء في المراحل المتقدمة من تحليل النص أي أنها تعاني من عجزها لانجاز التعديلات عند إجراء عملية الدمج مما يؤدي إلى انخفاض في دقة التجميع ، ويضاف إلى ذلك التعقيد بحساب التشابه بين كل زوج من التجمعات مما يجعلها غير ملائمة للتجمعات الكبيرة من المستندات. تمتاز خوارزمية التجزئة بأنها ملائمة جداً لعنقدة المجاميع الكبيرة من المستندات بسبب تناسبها لمتطلبات حسابية منخفضة وأيضا قليلة التعقيد مما يجعلها تستعمل بصورة واسعة [4،5،6] ولهذا استخدمنا تقنية التجزئة في عملنا.

ومن أشهر خوارزميات عنقدة التجزئة هي خوارزمية k-means وهي التي استخدمناها في هذا البحث حيث تمتاز بالبساطة ويستند عملها على تحليل التباينات وتعتمد على عدد من البراميترات مثل عدد العناقيد المطلوبة واختيار قيمة المركز الأولية، وتعد من الخوارزميات الأكثر مرونة حيث تبدأ الخوارزمية باختيار قيمة ابتدائية للمركز ثم تبنى العناقيد وذلك بحساب المسافة بين المركز والبيانات حيث الأقل مسافة توضع بنفس العنقود[7،8] . ولكون هنالك العديد من المستندات المتوفرة بعدد كبير من اللغات ومنها اللغة العربية التي تعد واحدة من ست لغات دولية التي تستخدم من قبل أكثر من 30 مليون شخص عبر العالم ، وتمتاز اللغة العربية بصعوبة استرجاع المعلومات لعدة أسباب منها امتلاك اللغة العربية تشكيلة لغوية معقدة ومزيج من الرموز تكتب بطرق مختلفة لتكون الكلمات [9] . لذا حاولنا في هذا العمل تقديم طريقة لعنقدة النصوص العربية في مجالات مختلفة باستخدام خوارزمية k-means .

2. الأعمال السابقة Related Works

توجد العديد من الأبحاث في هذا المجال إذ العديد من الباحثين قدموا عددا من خوارزميات عنقدة النصوص بعضها تم تطبيقها على اللغة العربية والبعض الآخر طبق على لغات أخرى، سنتناول في هذا الجزء عددا من البحوث السابقة التي تتضمن عرض دراسات مختلفة.

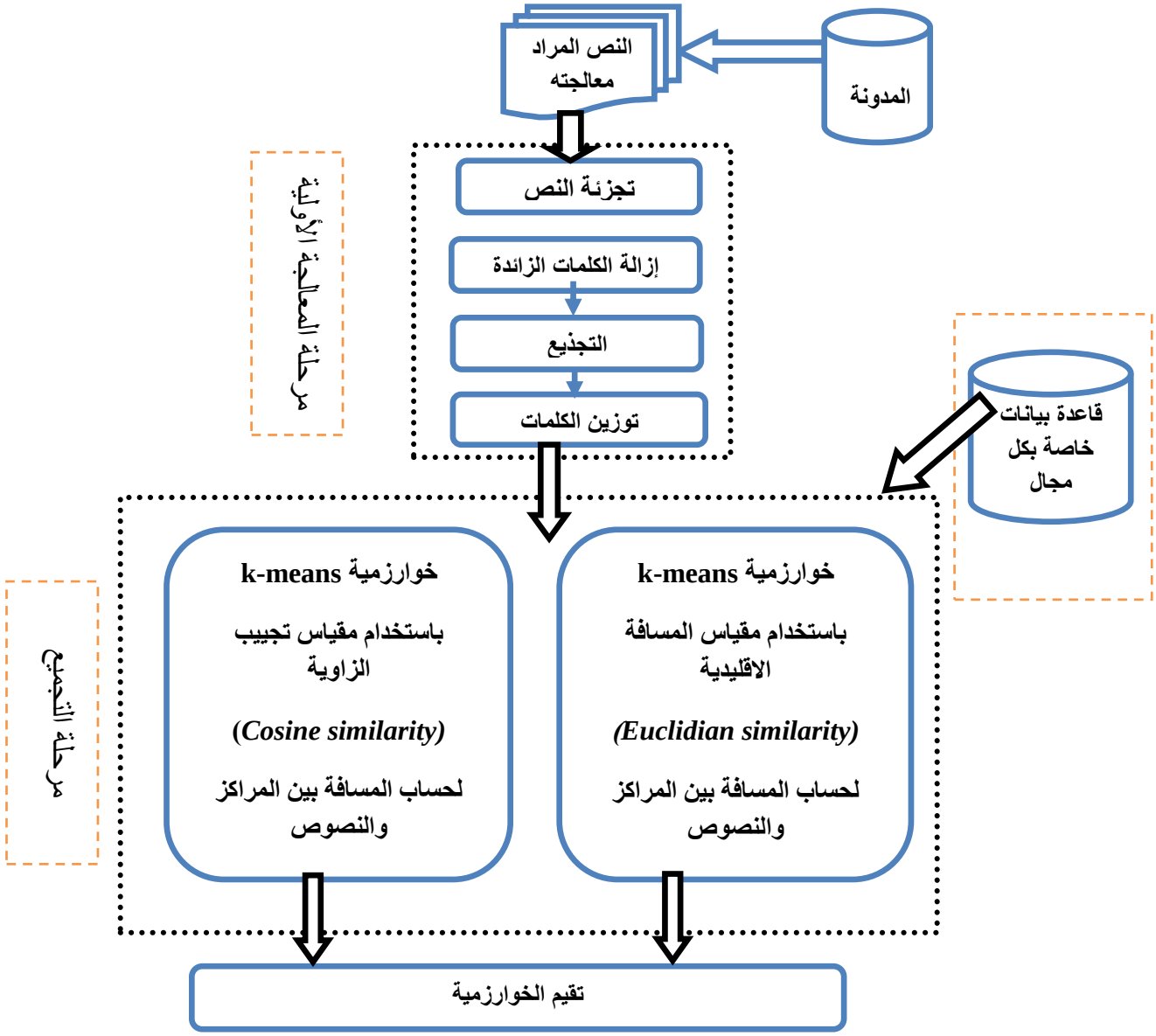
في عام 2000 قدم الباحثين (Michael Steinbach et.al)عملاً كمقارنة بين دراسة النتائج التجريبية لعدد من تقنيات العنقدة الشائعة حيث قدموا مقارنة بين طريقتين أساسيتين لعنقدة النصوص هي التقنية الهرمية وثلاثة أنواع لخوارزمية k-means وأثناء المقارنة وجد بالرغم من أن التقنية الهرمية تعطي نتائج أفضل إلا أنها تحدد بمضاعفة بوقت التعقيد على نقيض خوارزمية k-means بأنواعها إلا أن وقت التعقيد يكون خطي لعدد من الوثائق[10] . في عام 2007 قدم الباحث (Sameh) تقنية عنقدة للنصوص العربية طبقت باستخدام الدمج بين خوارزمية k-means مع التقنية الهرمية لتحسين الدقة في أنظمة استرجاع المعلومات من النص ، أيضا تم دراسة مدى تأثير عدد العنقدة في الدقة ، حيث وجد انه ليس من الضروري زيادة عدد العناقيد لتطوير انجازيه العنقدة للحصول على دقة أكثر[1]. في عام 2009 قدم الباحث(Haytham) طريقة لعنقدة المستندات العربية باستخدام التقنية الهرمية التي تستند على تكرار الكلمات (N-gram) فبعد إجراء سلسلة من المعالجات الأولية للنص اعتمد بشكل أساسي على معيار حساب تكرار الكلمات في النص وتكرارها في كل النصوص ثم بناء شجرة عنقدة للنصوص[11] . في عام 2011 قدم الباحث(Omaia) عملاً يهدف إلى تطبيق وتقييم خوارزمية k-means على عنقدة النصوص العربية وأيضا تخمين تأثير التجذيع على مثل هذه الخوارزمية ، وفي هذا العمل تم عرض نتائج لدقة الخوارزمية التي تتفاوت من مستوى بطيء إلى جيد جداً وأيضا تم عرض لنتائج بدون إجراء التجذيع على كلمات النصوص العربية مما أدى إلى تقليل دقة النظام لان للتجذيع دور مهم في استرجاع المعلومات من النص لان مهمته هو تجريد الكلمة من لواصقها وإعادتها إلى الشكل الأساسي [4]. في عام 2013 قدم الباحثان (Manjot & Navjot)طريقة لعنقدة مستندات الويب باستخدام خوارزمية k-means حيث اعتمد بشكل أساسي على اقل تشابه لإيجاد أفضل المراكز الأولية (centers) بدلا من اختيارها بشكل عشوائي وهو بذلك يزود طريقة كفوءة لتخصيص البيانات إلى العناقيد المناسبة مع تقليل وقت التعقيد، حيث صنفوا عملهم إلى ثلاث أصناف رئيسية الأول بالاعتماد على النص والثاني بالاعتماد على الرابط والثالث هجين مابين الصنفين[12] . في عام 2014 قدم الباحثين(Bashar Aubaidan et. al) دراسة تجريبية لتقنية عنقدة النصوص وكانت الدراسة كمقارنة بين خوارزمية k-means و++k-means وكانت الدراسة مخصصة بمجال واحد وهي نصوص الجريمة وبما إن إحدى التحديات في خوارزمية k-means هو اختيار مركز التجمع لذلك تم اقتراح خوارزمية k-means لإيجاد مركز التجمع الابتدائي وتم عرض نتائج الطريقتين [7].

3. المدونة Corpus

في أي مجال من مجالات اللغات الطبيعية ولأجل تطبيق وفحص أداء تقنية العنقدة المستخدمة لابد أن يستند العمل على مدونة تحتوي على عدد من النصوص ، احتوت مدونتنا على نصوص بمجالات مختلفة تنوعت بين نصوص فن ، رياضة ، جرائم ، وأخيرا مقالات طبية وبذلك اشتملت المدونة على 100 نص عربي في المجالات التي ذكرناها أعلاه وتم تخزينها بملف بلغة دلفي لسهولة التعامل معها في مراحل المعالجة التي ستذكر لاحقا.

4. الطريقة المقترحة Proposed Method

سنتناول في هذا الجزء شرح لطريقة العنقدة التي استخدمناها للنصوص العربية بالاعتماد على خوارزمية k-means بعد إجراء سلسلة من المعالجات الأولية لتهيئة النصوص من اجل العمل عليها. الشكل رقم (1) يمثل الهيكل العام للطريقة المقترحة.



شكل رقم (1) الهيكل العام للطريقة المقترحة

تتألف الطريق المقترحة بصورة عامة من المراحل الآتية :-

أ- **المرحلة الأولى (مرحلة المعالجة الأولية)** :- في هذه الخطوة تتم المعالجة البدائية والتي سنحاول من خلالها تحليل النص العربي بتطبيق عدد من المهام لتحويل النص إلى قائمة من الكلمات يسهل التعامل مع مفرداتها وذلك بتنفيذ عدد من الخطوات المتسلسلة الآتية على النص المدخل :-

- **تجزئة النص (Tokenization)** :- وهي أول خطوه في تحليل النص وتتضمن مهمتين احدهما تبدأ بفصل العنوان عن محتوى النص من اجل الاستفادة منه في عمليات إحصائية في خوارزمية التجميع التي ستذكر لاحقا ، أما المهمة الثانية هي تجزئة النص إلى قائمة من الكلمات.

❖ **إزالة الكلمات الزائدة (Stop Words Remove)** :- كل لغة تمتلك كلمات توقف وهذه الكلمات ليس لها معنى بحيث عند حذفها لا تؤثر على السياق لأنها بمثابة مفردات ربط ، وأفضل طريقة لإزالة هذه الكلمات هي مقارنة كل كلمة بقاعدة بيانات خاصة تم فيها تخزين اغلب كلمات التوقف المشهورة في اللغة العربية ، ومن أمثلة على تلك الكلمات حروف الجر وأسماء الإشارة ، مثلاً : هي ، التي وهذا والذين ....وغيرها.

❖ **التجذيع (Stemming)**:- هي أهم العوامل المؤثرة في استرجاع المعلومات من النصوص حيث لها فعالية في عامل التأثير على انجازيه نظام التجمع للنصوص ، يتم من خلالها عرض الكلمات التي حصلنا عليها من الخطوة السابقة على محلل صرفي كفوء يقوم بدوره بإزالة السوابق واللواحق من الكلمات بهدف تحويلها إلى الشكل الأساسي لها.

❖ **توزين الكلمات (Words Weighting)** :- في هذه الخطوة يتم إسناد وزن (قيمة) إلى كل كلمة بالنص وذلك من خلال استخدام معامل لتوزين الكلمات وهو مقياس مشهور يدعى tf×idf الذي هو اختصار ل (term frequency ×inverse document frequency) ويكون عمله إحصاء التكرارات وذلك بحساب نسبة تكرار الكلمة في النص إلى نسبة تكرارها في مجموعة النصوص الكلية. عدد من الصيغ تستخدم مع هذا المقياس لكن في هذا العمل استخدمنا الصيغة الآتية [7] :

$$wi = tf \times \log(N/ni)$$ .....معادلة رقم (1)

حيث :-

tf :- تكرار الكلمة في النص

N:- العدد الكلي للنصوص في االمدونة

ni :- رقم النص الذي يحتوي الكلمة (t)

ب- **المرحلة الثانية (بناء قاعدة بيانات)**:- يتم في هذه المرحلة بناء قاعدة بيانات تحتوي على الكلمات المفتاحية المشهورة في كل مجال ، وبما إننا اعتمدنا في عملنا على أربع مجالات كما ذكرناها سابقاً (فن ، رياضة ، جرائم ، طبية) لذا قمنا ببناء أربع قواعد بيانات ، كل قاعدة خاصة بمجال معين مع الأخذ بنظر الاعتبار عند إنشاء هذه القواعد قمنا بإعطاء كل منها اسم يدل على مجالها لتسهيل التعامل معها في الخطوة التالية. الهدف من هذه المعالجة هي اعتماد تلك القواعد للاستفادة منها في خوارزمية العنقدة كما سيتم شرحها في المرحلة اللاحقة والجدول رقم(1) يعرض نماذج قواعد البيانات المستخدمة.

الكلمات المفتاحية في مجال:-			
الرياضة	**الفن**	**الجرائم**	**الطبية**
هدف	**لوحة**	**قتل**	**مرض**
لاعب	**موسيقى**	**ضحية**	**دواء**
شوط	**رسم**	**اغتصاب**	**دم**
مدرب	**فنان**	**اعتدى**	**أصيب**
مباراة	**عزف**	**سرق**	**عالج**
كرة	**تصميم**	**مجزرة**	**التهاب**
قدم	**انسجام**	**اتهم**	**سرطان**
سلة	**معرض**	**ارتكب**	**حموضة**
ملاكمة	**نحت**	**قمع**	**إسهال**
بيسبول	**زخرفة**	**اقترف**	**صداع**

جدول رقم (1) نماذج من الكلمات المفتاحية الخاصة بكل مجال

ت-**المرحلة الثالثة (مرحلة العنقدة Clustering)**:- قمنا في هذه المرحلة بتطبيق إحدى تقنيات العنقدة المتمثلة بخوارزمية k-means (التحليل العنقودي على أساس الوسيط) لأجل تقسيم النصوص إلى مجموعة من العناقيد (clusters) اعتمادا على اشتراكها بالخواص المتشابهة ، تمتاز هذه الخوارزمية بكفاءة عالية بتوليد عدد من العناقيد وتجميع النصوص فيها من خلال قياس المسافات بين البيانات والعنقود الوسيط المركزي (centriod center) ، تعمل الخوارزمية عبر سلسلة من الخطوات المتتابعة الآتية :-

1- تحديد قيمة k: في هذه الخطوة يتم تحديد قيمة k ، حيث k هي عدد العناقيد المطلوبة والتي على أساسها نحصل على عدد العناقيد وهذه القيمة يمكن ان تحدد مسبقا من قبل المستخدم أو عشوائيا بشرط أن تكون قيمتها ضمن المدى (عدد النصوص $\geq K \geq 1$) ، في عملنا حاولنا اختيار اكثر من قيمة لل k كما سنعرض في النتائج.

2- اختيار مراكز التجميع (centeriod centers) :- تتم في هذه الخطوة عملية تحديد المراكز التي سيتم على أساسها التجميع وبناء العناقيد(التجمعات) ،فبدلاً من اعتماد الطريقة العشوائية لإيجاد المراكز اقترحنا طريقة جديدة لإيجادها بالاعتماد على قاعدة البيانات الخاصة بمجال النصوص من الجدول رقم (1) الذي ذكرناه في أعلاه والتي تم بنائها في المرحلة السابقة، وذلك بإسناد قيمة إلى كل نص بقياس التشابه بين عنوان كل نص مع قاعدة البيانات التي تم اختيارها لإيجاد الكلمات المتشابهة بين النص وقاعدة البيانات باستخدام المعيار الأتي:-

$Sim(X,Y)=\sum_{i=1}^{n}(xi,yj)$ $\quad i \geq j$ .....معادلة رقم (2)

حيث xi هي تكرار الكلمة في الجملة x

yj هي تكرار الكلمة في الجملة y

وبعد إسناد قيمة إلى كل النصوص، نقوم بترتيبها تنازليا ثم نختار اكبر قيمة بينها ونجعل هذا النص هو المركز للمجال المحدد وعلى أساسه تبنى العناقيد (التجمعات) . ويتم تغيير مراكز التجمع أربع مرات، كل مرة لمجال معين وفي كل مرة يتم حساب المركز بالاعتماد على مقياس التشابه أعلاه وقاعدة البيانات الخاصة بذلك المجال.

3- تحديد عدد المحاولات لفحص العناقيد :- عدد المحاولات المطلوبة يجب أن يكون اكبر أو يساوي واحد.

4- حساب المسافة بين النصوص ومراكز التجميع :- يتم إسناد كل نص إلى التجمع الأقرب له اعتماداً على المسافة بين المركز والنص التي تحسب باستخدام مقياس التشابه فيتم إضافة النص إلى التجمع اعتماداً على اقل المسافات بين المركز والنص. في عملنا هذا حاولنا تنفيذ الخوارزمية مرتين (كلاً على حدة) ، مرة باستخدام مقياس المسافة الاقليدية ومرة ثانية باستخدام مقياس تجييب الزاوية لقياس التشابهات ، ثم قمنا بعمل مقارنة بين المقياسين لإيجاد الأفضل بينهما كما سيتم عرضه في الجزء الخاص بالنتائج. والمعادلات الآتية توضح المقياسين [2]:-

معادلة رقم (3) .....

$$sim(X,Y) = EucDis(X,Y) = \sqrt{\sum_{i=1}^{n}|x_i - y_i|^2}$$

حيث xi تكرار الكلمة في المستند x

حيث yi تكرار الكلمة فى المستند y

معادلة رقم (4) .....

$$sim(X,Y) = Cos\Theta = \frac{\sum_{i=1}^{n} x_i . y_i}{\sqrt{(\sum_{i=1}^{n}(x_i)^2 * (\sum_{i=1}^{n}(y_i)^2)}}$$

حيث xi تكرار الكلمة في المستند x

حيث yi تكرار الكلمة فى المستند y

بعد الانتهاء من حساب المسافات باستخدام المقاييس أعلاه يتم تجميع النصوص في عناقيدها بالاعتماد على اقل مسافة ، حيث يجب أن تكون النصوص داخل التجمع الواحد مدى التشابه بينها كبير بينما تكون اقل تشابه مع النصوص في التجمعات الأخرى.

5- تخزين النتائج في قاعدة بيانات ، حيث يتم خزن النصوص التي تم إدخالها عن طريق النظام في إحدى التجمعات للمجالات الأربعة (فن ، رياضة ، جرائم ، طب) بالاعتماد على نوع المجال التابعة له .

6- تكرار خطوات العمل :- يتم إعادة تنفيذ الخطوات (4 ، 5) حتى يتم الحصول على عدد المحاولات المطلوبة، الهدف من هذه الخطوة هي تحسين عمل الخوارزمية لأنها تعتمد على مراكز التجميع لذلك قمنا بتنفيذ الخوارزمية لمرات متكررة مع اختلاف مراكز التجميع في كل مرة عن المرات السابقة.

5. النتائج Results

سنتناول في هذا الجزء عرض لنتائج التقييم بعد عرض النصوص على خوارزمية k-means ،حيث لفحص انجازيه الطريقة المقترحة لابد من إجراء عملية التقييم وهي من الأمور المهمة في أنظمة استرجاع المعلومات للتأكد من صحة عمل خوارزمية k-means ولكن قبل إجراء هذه العملية لابد من تهيئة جميع النصوص في المدونة وذلك بخلق قاعدة بيانات يحتوي على جدول يتضمن فيه اسم كل نص مع حقل يشير إلى الصنف الذي يعود إليه والجدول رقم (2) يبين ذلك:-

اسم النص	صنفه
D_1	رياضة
D_2	فن
D_3	رياضة
:	
:	
D_n	طبية

جدول (2) مثال على تهيئة المدونة للتقييم

بعد تهيئة النصوص في المدونة ، نستخدم احد المقاييس لتقييم خوارزمية k-means حيث هنالك نوعين من المقاييس للتقييم ، النوع الأول هو مقياس الجودة الداخلي حيث يجري التقييم للتجمعات دون الإشارة إلى مصادر المعرفة الخارجية أما النوع الثاني فهو يستند على التقييم وذلك بعمل مقارنة بين التجمعات الناتجة عن الخوارزمية مع أصناف معرفة مسبقا وهذا يسمى مقياس الجودة الخارجي ومن الأمثلة على معايير الجودة الخارجي هي F-Measure ، Entropy ، التشابه الكلي (Overall Similarity) [10] .في عملنا هذا استخدمنا مقياس F-Measure حيث يدمج بين قياس الدقة والاسترجاع من النصوص. تم معاملة كل تجمع (عنقود) يقابل الاستفسار وكل صنف يقابل مجموعة

النصوص في المدونة التي تم تمثيلها في الجدول رقم (2) ، بعد ذلك لكل تجمع (cluster j) وصنف (class i) نحسب الدقة والاسترجاع من المعادلات الآتية رقم (5، 6 ،7) [10]:-

$$Recall(i, j) = n_{ij} / n_i$$ .....معادلة رقم (5)

$$Precision(i, j) = n_{ij} / n_j$$ .....معادلة رقم (6)

$$F(i, j) = (2 * Recall(i, j) * Precision(i, j)) / ((Precision(i, j) + Recall(i, j))$$ .....معادلة رقم (7)

حيث أن nij هي عدد النصوص التابعة للصنف i في التجمع j

nj هي عدد النصوص للتجمع j

ni هي عدد النصوص في الصنف i

ثم يتم احتساب القيمة الإجمالية لل F-$measure$ والذي يحسب من المعادلة الآتية :

$$F = \sum_i \frac{n_i}{n} \max\{F(i, j)\}$$ .....معادلة رقم (8)

حيث n هي العدد الكلي للنصوص في المدونة

ni هي عدد النصوص في الصنف i

F هو المجموع الكلي للتجمعات التي تحتوي على اكبر عدد من أصناف النصوص

وبما ان قيمة F محصورة بين $(1 \sim 0)$ اذن F في اعلاه تشير الى اعلى قيمة لدقة التجميع

بعد استخدام مقاييس التقييم سنعرض نبذة من النتائج نلخصها بالنقاط الآتية:-

1- يعرض الجدول رقم (3) وصف للمدونة التي اعتمدنا عليها والتي تألفت من 100 نص فالجدول الآتي يشمل وصف بسيط لعدد الأصناف التي استخدمناها ويقابلها عدد النصوص في كل صنف.

الأصناف	عدد النصوص
رياضة	25
فن	20
جرائم	30
طبية	25

جدول رقم (3) أصناف النصوص مع أعدادها

2- الجدول رقم (4) يعرض عينة من النتائج المطلوبة لمعالجة عنقدة النصوص العربية والتي تمثلت بقيم مقاييس الدقة (Precision) والاسترجاع (Recall) وF-measure، حيث كما نلاحظ الجدول يعرض نتائج بمثابة مقارنة بين قيم مقاييس التقييم الثلاث التي حصلنا عليها من خوارزمية k-means مرة باستخدام مقياس المسافة الاقليدية لحساب المسافات بين المركز والتجمعات ومرة أخرى لقياس نفس المسافة باستخدام مقياس اخر هو مقياس تجييب الزاوية ، فمن تقييم النتائج لاحظنا بان انجازيه الخوارزمية باستخدام مقياس تجييب الزاوية في حساب المسافة كان يعطي نتائج أفضل من مقياس المسافة الاقليدية للأصناف الأربعة (رياضة ، فن ، جرائم ، طب) .

اسم الصنف	**خوارزمية k-means باستخدام مقياس المسافة الاقليدية (Euclidian similarity)**			**خوارزمية k-means باستخدام مقياس تجييب الزاوية (Cosine similarity)**		
	P	*R*	*F-Measure*	*P*	*R*	*F-Measure*
رياضة	0,65	0,430	0,516	0,832	0,4	0,540
فن	0,693	0,297	0,415	0,564	0,393	0,463
جرائم	0,391	0,217	0,279	0,698	0,314	0,433
طب	0,362	0,125	0,185	0,261	0,28	0,270

جدول رقم (4) نتائج مقاييس التقييم للخوارزمية مع المقياسين (المسافة الاقليدية، تجييب الزاوية) للأصناف

3- الجدول رقم (5) يعرض نتائج لخوارزمية k-means في الدقة والاسترجاع بعد أن اعتمدنا على عدد التجمعات وليس على الأصناف ،وأيضا بتطبيق الخوارزمية على مقياسين للتشابه (Cosine, Euclidian) كما ذكرنا بالجدول السابق رقم (4)، فمن متابعة الجدول يمكن أن نلاحظ بأنه أيضا عندما فحصنا انجازيه الخوارزمية المقترحة على عدد العناقيد (التجمعات) وجدنا أيضا بان مقياس تجييب الزاوية يعطي نتائج أفضل من مقياس المسافة الاقليدية.

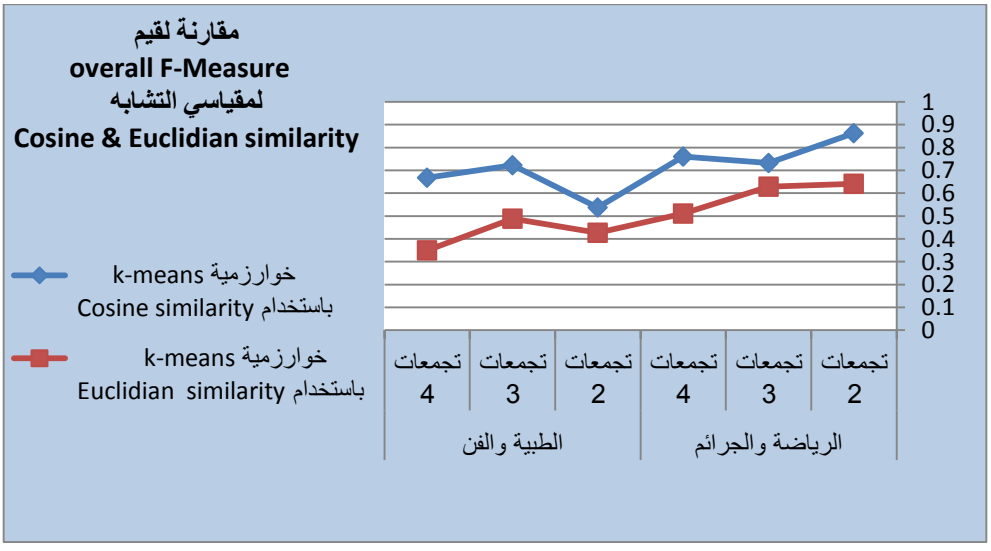
عدد التجمعات	**خوارزمية k-means باستخدام مقياس المسافة الاقليدية (Euclidian similarity)**			**خوارزمية k-means باستخدام مقياس تجييب الزاوية (Cosine similarity)**		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
2	0,776	0,428	0,551	0,931	0,620	0,730
3	0,763	0,588	0,664	0,831	0,65	0,744
4	0,259	0,15	0,189	0,329	0,222	0,266

جدول رقم (5) يعرض نتائج مقاييس التقييم للخوارزمية مع المقياسين (المسافة الاقليدية، تجييب الزاوية) للتجمعات

4- الجدول رقم (6) يعرض مقارنة لقيم Overall F-Measure باستخدام المعادلة رقم (8) لخوارزمية k-means باستخدام مقياسين للتشابه، حيث قمنا بتطبيق المعادلة وفقا للنصوص المتوفرة في المدونة والتي يبلغ عددها 100 نص عربي لأربعة أصناف (رياضة ، فن ، جرائم ، طبية) ، عرضنا في الجدول الأتي القيم الكلية لمقياس F-Measure لحالتين، الأولى(الرياضة والجرائم) والثانية هما (الطبية والفن) ولتجمعات مختلفة وأيضا بتطبيق الخوارزمية على مقياسين للتشابه (Cosine , Euclidian) وأيضا كان مقياس تجييب الزاوية يعطي نتائج أفضل من المقياس الآخر وهذا ما يتم ملاحظته بالشكل رقم (2) الذي يوضح بان مقياس تجييب الزاوية افضل من مقياس المسافة الاقليدية.

عدد الأصناف	التجمعات	خوارزمية k-means باستخدام (Euclidian similarity)	خوارزمية k-means باستخدام (Cosine similarity)
		Overall F-Measure	Overall F-Measure
عدد الأصناف 2 (الرياضة والجرائم)	2	0,641	0,863
	3	0,629	0,732
	4	0,511	0,76
عدد الأصناف 2 (الطبية والفن)	2	0,427	0,538
	3	0,488	0,722
	4	0,35	0,668

جدول رقم (6) يعرض النتائج الاجمالية لمقياس F-Measure



شكل رقم (2) يوضح مقارنة بين انجازية خوارزمية k-means باستخدام مقياسي التشابه (المسافة الاقليدية ، تجييب الزاوية)

5- الشكل رقم (2) يعرض الواجهة الرئيسية التي تم استخدامها لتطبيق الطريقة المقترحة على النصوص العربية باستخدام برنامج دلفي :

شكل رقم (3) الواجهة الرئيسية للطريقة المقترحة

6- الاستنتاجات Conclusion

لكثرة النصوص الالكترونية المتوفرة على الانترنت ظهرت الحاجة إلى توفير آلية لتنظيم تلك النصوص وبالتالي يسهل استرجاع المعلومات منها, ومن تلك الآليات هي آلية عنقدة النصوص التي تمثل إحدى تقنيات تنقيب النصوص.وعلى الرغم من وجود العديد من الأبحاث التي طورت وطبقت على النصوص في اللغات المختلفة لكننا وجدنا هناك قلة في الأبحاث التي طبقت على النصوص العربية ، لذا قدمنا طريقة جديدة لعنقدة النصوص العربية باستخدام خوارزمية k-means. وبما إن إحدى التحديات في خوارزمية k-means هو اختيار قيمة مركز التجميع ,حيث ان من نقاط ضعف الخوارزمية هي اختيار مركز التجميع عشوائياً ، لذلك قمنا بتعديل خوارزمية k-means من اجل تقويتها بجعل اختيار المراكز تعتمد على قاعدة بيانات للكلمات المفتاحية لكل صنف , وعند تقييم الخوارزمية المقترحة حصلنا على نتائج جيدة . بالإضافة الى ذلك قيمنا الخوارزمية باستخدام مقياسين للتشابه (حيث طبقنا خوارزمية k-means مرتين, المرة الأولى باستخدام مقياس المسافة الاقليدية لحساب المسافات بين المراكز والتجمعات والمرة الثانية استبدلنا مقياس المسافة الاقليدية بمقياس التجييب لحساب المسافات) ومن النتائج التي عرضناها في جزء النتائج استنتجنا بان مقياس التجييب يعطي نتائج أفضل من مقياس المسافة الاقليدية وذلك لان قيمة مقياس F-measure تتزايد في الحالة الثانية وهذا دليل على تحسين اداء الخوارزمية ، وأيضا قمنا بعمل مقارنة للنتائج على أصناف وتجمعات مختلفة. من الأعمال المستقبلية لتطوير هذا العمل هو استخدام خوارزمية k-means في تلخيص عدد من النصوص وايضا ممكن استخدامه مع أجراء بعض التعديلات البسيطة على الطريقة لعنقدة النصوص على الويب.

References المصادر

[1] Sameh H. Ghwanmeh ," ***Applying Clustering of Hierarchical K-means-like Algorithm on Arabic Language***" , i*n Proc. of the* International Journal of Information Technology Volume 3 Number 3, 2007, pp. 168-172.

[2] Mohammad & Abdulatif Al-Abdo," ***Experiments in Mining Arabic Texts***" , i*n Proc. of the* J. of Commun. & Comput. Eng. Volume 2, Issue 1, 2012, Pages 14:18.

[3] Milos Radovanovic & Imirjana Ivanavic ," ***Text Mining: Approaches and Applications*** " , i*n Proc. of the* novisad J.Math ,Volume 38, Number 3, 2008, pp. 227-234.

[4] Omaia M. Al-Omari ," ***Evaluating The Effect Of Stemming In Clustering Of Arabic Documents***" , i*n Proc. of the* Academic Research International, Volume 1 ,Number 1, 2011, pp. 284-291.

[5] Xiaohui Cui & Thomas E. Potok," ***Document Clustering Analysis Based on Hybrid PSO+K-means Algorithm***" , i*n Proc. of the* Journal of Computer Sciences (Special Issue) ,2005,pp.27-33.

[6] K.Sathiyakumari et. al ," ***A Survey on Various Approaches in Document Clustering*** " , i*n Proc. of the* G Manimekalai et al, Int. J. Comp. Tech. Appl., Volume 2 , Number 5, 2011,pp.1534-1539.

[7] Bashar Aubaidan et. al," ***Comparative Study Of K-MEANS And K-MEANS++ Clustering Algorithms On Crime Domain***" , i*n Proc. of the* Journal of Computer Science ,Volume 10 Number 7, 2014, pp. 1197-1206.

[8] Osama Abu Abbas ," ***Comparison Between Data Clustering Algorithms***" , i*n Proc. of the* International conferences on Arab Journal of Information Technology ,Volume 5 Number 3, July 2008, pp. 320-325.

[9] Mahmud S. Alkoffash," ***Comparing between Arabic Text Clustering using K - Means and K Mediods*** " , in Proc. of the International Journal of Computer Applications (0975 – 8887) ,Volume 51, Number 2, August 2012, pp. 5-8.

[10] Michael Steinbach et.al ," ***A Comparison of Document Clustering Techniques*** " , i*n Proc. of the* KDD workshop on text mining, 2000, pp. 109-111.

[11] Haytham S. Al-sarrayrih ," ***Clustering Arabic Documents Using Frequent Itemset-based Hierarchical Clustering with an N-Grams***" , in Proc. of the International conferences on Arab Journal of Information Technology, 2009.

[12] Manjot Kaur & Navjot Kaur," ***Web Document Clustering Approaches Using K-Means Algorithm*** " , i*n Proc. of the* International Journal of Advanced Research in Computer Science and Software Engineering Volume 3 Number 5, May 2013, pp. 861-864.

Comparative Study the Effect of Similarity Measures on K-Means Algorithm in Clustering Arabic Texts based on Keywords

Suhad muhajer kareem

Basra University \ Science Collage \ Computer Science Dept.

Abstract

Texts clustering is one of important and effective tasks in texts mining, it aims to divide a large sets of texts into subsets called clusters, these clusters contain objects have high similar among themselves but are dissimilar to objects in the other clusters. In this work, we proposed method is used to cluster Arabic texts using one of the famous techniques called K-Means algorithm. The proposed method include analysis of text as a primary step to prepare it to clustering algorithm which applied to 100 Arabic texts in four different groups included (sport, art , crime , health). Our method developed by using database of keywords for each field to select cluster centers rather than selected it randomly , then two similarity measures(Euclidian similarity, Cosine similarity) are used to calculate the distances between the centers and the texts for building clusters. In addition , we evaluate the impact of the two similarity (Euclidian similarity, Cosine similarity) on the results of k-means by using F-Measures and the results were as a compared between Euclidian similarity and cosine similarity based on the number of factors such as number of clusters and number of groups. Finally, we found that the performance of k-means algorithm using cosine similarity work better than k-means algorithm using Euclidian similarity.

Keywords

Text mining , Arabic text clustering , K-means, Euclidian similarity , Cosine similarity