Research Paper Ref. No 2576

تقييم طرق التعرف الموضوعي للنصوص العربية و الترابط بينها باستعمال مدونة مستخرجة من جريدة الوطن العمانية

Evaluation of Topic Identification Methods for Arabic Texts and their Combination by using a Corpus Extracted from the Omani Newspaper Alwatan

[1] مراد عباس، [2] كمال سماعيلي، [3] داود بركاني

Abbas, M., Smaili, K., Berkani, D.

[1] مركز البحث العلمي والتقني لترقية اللغة العربية، 1 شارع جمال الدين الأفغاني، 16011، الجزائر

[2] مخبر لوران للبحث في الإعلام الآلي وتطبيقاته (إنريا – لوريا)، ص.ب. 101، نانسي، 54602، فرنسا

[3] المدرسة العليا المتعددة التقنيات، 10 شارع حسن بادي، 16200، الجزائر

E-mail: m_abbas04@yahoo.fr, Kamel.smaili@loria.fr, dberkani@enp.edu.dz

المستخلص: يستعمل التعرف الموضوعي في تطبيقات عديدة، منها التعرف الآلي على الكلام، الترجمة الآلية ومحركات البحث. ويقصد بالتعرف الموضوعي، إيجاد الموضوع أو مجموعة المواضيع التي تعالج نصا معينا. يتبلور محور هذه الدراسة حول التعرف الموضوعي للنصوص العربية. للبدء في هذا العمل، قمنا بجمع عدد من النصوص من الموقع الإلكتروني للجريدة العمانية "الوطن"، والهدف من ذلك هو إنشاء مدونة عربية تمكننا من إجراء تجارب تقييم الطرق و الخوارزميات المستعملة. إن بعض الطرق التي تم عرضها في هذا المقال معروفة في ميدان تصنيف النصوص والتي استعملت لمعالجة اللغات اللاتينية مثل خوارزمية الجار الأقرب "خ.ج.أ" و (TF-IDF) (تردد اللفظة، عكس تردد الوثيقة)، وحديثا اقترحنا طريقة جديدة تدعى مصنف الزناد، تعتمد على حساب الزنادات أو المعلومة المتبادلة المتوسطة لكل زوج من الكلمات. للحصول على نتائج أفضل، قمنا بعملية الترابط بين مختلف الطرق المستعملة. وقد استعملنا لهذا الغرض ثلاث طرق هي على التوالي: تصويت الأغلبية، تصويت الأغلبية المحسن و الترابط الخطي.

كلمات مدخلية: *التعرف الموضوعي، "خوارزمية الجار الأقرب"، (TF-IDF)، مصنف الزناد، اللغة العربية، تصويت الأغلبية، تصويت الأغلبية المحسن، الترابط الخطي.*

Abstract: Topic identification is used in several applications, as adapting language models for speech recognition and machine translation, focusing on a specific use for search engines, etc. Topic identification consists to assign one or several topic labels to a flow of textual data. Labels are chosen from a set of topics fixed a priori. In this paper, we present a study about identifying topics of Arabic texts. For this, a considerable amount of data is needed. Thus, we started by collecting texts from the website of the Omani newspaper "Alwatan". The result is an Arabic corpus composed of more than 9000 articles corresponding to nearly 10 millions words. The considered topics in our experiments are: Culture, Religion, Economy, Local news, International news and sports. Some of the methods presented in this study, are well known in the text categorization community, as TFIDF classifier and kNN "k Nearest Neighbors". The objective to use these methods is to compare them to TR-classifier

"TRiggers-based classifier", a new method that we have proposed, which is based on computing triggers or the Average Mutual Information of each couple of words. In order to enhance performances, we have combined results of the three methods by using three approaches: Majority Vote, Enhanced Majority Vote and Linear Combination.

Keywords: *Arabic corpus, TR-classifier; kNN, TFIDF, Linear Combination, Enhanced Majority Vote.*

المقدمة

يعتبر التعرف الموضوعي لنص ما العملية الآلية التي تتيح إرفاق الموضوع الصحيح الذي يتسم به ذلك النص. أما التصنيف فهو عبارة عن تجميع النصوص التي تعالج موضوعا مشتركا، في فئة واحدة. ونظرا للتشابه الكبير بين هذين التعريفين، فإنه يمكن استعمال طرق التصنيف بهدف التعرف الموضوعي، والعكس صحيح. ومن بين هذه الطرق نذكر آلات الدعم الإتجاهي (Joachims, 1998)، آلات الدعم الإتجاهي متعددة الفئات (Abbas *et al*., 2009) (Lee *et al*., 2004), (TF-IDF) (Guermeur *et al*., 2004)، خوارزمية شجرة القرار (Fuhr) (Moulinier,) (Lewis and Ringuette, 1994) (*et al*., 1991)، شبكة العصبونات (Wiener *et al*., 1995, Ng *et al*., 1997) وخوارزمية الجار الأقرب (Creecy *et al*., 1992) (Yang, 1994).

معظم الدراسات السابقة كانت تعالج اللغات الأوروبية، كالإنجليزية والفرنسية والأسبانية وغيرها، وكذلك اللغات الآسيوية كاليابانية والصينية. أما اللغة العربية فلم تحظ بنفس القدر من الاهتمام، إذ أن عدد البحوث التي أجريت في هذا المجال كان قليلا، كتلك المنشورة في (Abbas and Smaili, 1995; El-Kourdi *et al*., 2004; Elhalees, 2007)

في هذا البحث سيتم عرض التجارب التي أجريت على بعض الطرق المعروفة مثل خوارزمية الجار الأقرب و (TF-IDF)، والطريقة الجديدة التي سميت بمصنف الزناد. لقد تم اختيار الطريقتين الأوليتين لإجراء مقارنة بينها وبين مصنف الزناد، و ذلك للترابط بين هذه الطرق باستعمال تصويت الأغلبية، تصويت الأغلبية المحسن و الترابط الخطي من أجل تحسين النتائج التي تم الحصول عليها عند استعمال أي طريقة على حدة.

ولإجراء هذه التجارب تم إنشاء مدونة باللغة العربية، وذلك بتحميل آلاف المقالات من الموقع الإلكتروني لجريدة "الوطن" العمانية، وذلك باختيار ستة مواضيع وهي: ثقافة، دين، اقتصاد، أخبار محلية، أخبار عالمية ورياضة.

المدونة وتمثيل الوثائق

التعرف الموضوعي يحتاج عددا كبيرا من النصوص، لذلك تم جمع عدد كبير من المقالات، مصدرها جريدة الوطن العمانية، وهي بحدود 9000 وثيقة، أي ما يعادل 10 ملايين كلمة. هذه النصوص موزعة حسب المواضيع الستة التي تم تحديدها، كما هو موضح في الجدول رقم 1.

جدول 1. عدد الكلمات موزع حسب المواضيع.

الموضوع	عدد الكلمات
ثقافة	1.359.210
دين	3.122.565
اقتصاد	855.945
أخبار محلية	1.460.462
أخبار عالمية	1.555.635
رياضة	1.423.549
المجموع	**9.813.366**

هناك الكثير من الكلمات غير مفيدة، بل تؤدي إلى تدني كفاءة المصنفات. فعلى سبيل المثال أدوات اللغة المتمثلة في حروف الجر، ظرف الزمان والمكان، أسماء الإشارة وغيرها هي غير مرغوب فيها ويجب إزالتها من المدونة (Frakes and Baeza-Yates, 1992)، وكذلك الكلمات التي لا يتجاوز تكرارها قيمة معينة، في أغلب الأحيان 5. جدول رقم 2 يوضح توزيع عدد الكلمات حسب المواضيع بعد حذف الأدوات الزائدة.

جدول 2. عدد الكلمات موزع حسب المواضيع بعد حذف أدوات اللغة.

الموضوع	عدد الكلمات
ثقافة	1.013.703
دين	2.133.577
اقتصاد	630.700
أخبار محلية	1.111.246
أخبار عالمية	1.182.299
رياضة	1.067.281

ولا يقف الأمر عند هذا الحد، فاستخراج جذور الكلمات وفصلها عن الضمائر المتصلة وأداة التعريف وغيرها من الزوائد يؤدي إلى تحسين النتائج، من خلال تزويد المصنف بالتكرارات الصحيحة للكلمات. وبدون هذه العملية فإن استعمال خوارزمية حساب التكرار لا يمكن أن يؤدي إلى إيجاد القيمة الحقيقية للتكرار الخاص بكلمة «قلم» إذا وجدت في النص بالأشكال

الثلاث «القلم»، «قلمهم»، «الأقلام». بعد إزالة أدوات اللغة من المدونة، تناقص حجمها بحوالي 27%.

يعتبر تمثيل الوثائق من الخطوات الأساسية، إذ تحتاج كل وثيقة إلى المعالجة ليتم تحويلها على شكل متجه ذي بعد يساوي عدد الكلمات المختلفة المكونة للوثيقة. وتمثل مركبات المتجه أوزان هذه الكلمات بحيث حصلنا على وزن كل كلمة بحساب تكرارها، وتكرار الوثيقة. وباختصار فإن تمثيل الوثائق الذي اعتمدناه هو نفسه الخاص بالمصنف (TF-IDF).

ولتكوين مجموعة المفردات يمكن استعمال طرق مختلفة منها تكرار الكلمات (Yang and Pedersen, 1997)، تكرار الوثيقة، المعلومة المتبادلة وتقنية النقطة الانتقالية (Pinto *et al*., 2006, Seymore *et al*. 1998). وقد تم في هذا البحث الاعتماد على الطريقة الأولى أي على حساب تكرارات هذه المفردات المستخرجة من مدونة التدريب. والدافع من وراء ذلك هو خلوها من التعقيدات الحسابية مع تقديمها لنتائج جيدة. ونشير إلى أن هناك بعض الطرق التي تستخدم مجموعة المفردات العامة كخوارزمية الجار الأقرب و (TF-IDF)، وبعضها الآخر يتطلب استعمال مجموعة مفردات خاصة بكل موضوع، مثل مصنف الزناد. وقد تم استعمال عدد صغير نسبي لمجموعة المفردات مقارنة بالأعداد المستعملة في (Abbas, 2008, Abbas and Smaili, 2005)، ولكن ما تم إضافته هو ترتيب مجموعة المفردات ترتيبا تنازليا حسب قيم تكرارها. حيث بلغ الحجم الأقصى لمجموعة المفردات العامة المستعملة 800 لفظة.

وصف طرق التعرف الموضوعي

خوارزمية الجار الأقرب

استعملت خوارزمية الجار الأقرب في تصنيف النصوص منذ بداية التسعينيات ((Yang, 1994) (Masand *et al*., 1992) (Iwayama and Tokunaga, 1995) (Yang and Liu, 1999)). ففي (Yang and Liu, 1999) تم عرض دراسة موضوعها مقارنة عدد من طرق التصنيف من بينها "خ.ج.أ" باستعمال مدونة بنشمارك رويترز (Reuters - 21450). وكانت النتيجة أن جاءت مرتبة "خ.ج.أ" بعد طريقة آلات الدعم الإتجاهي مباشرة. كما سجلت أبحاث أخرى نتائج جيدة لخوارزمية الجار الأقرب، وهذا باستعمال مدونات مختلفة (Yang and Liu, 1999) (Joachims, 1998) (Baoli *et al*., 2002).

وللتعرف على موضوع نص معين للاختبار «ن»، يقوم مبدأ الخوارزمية البسيط على البحث في مدونة التدريب عن متجهات النصوص المجاورة أو القريبة من«ن» وإخراجها، ليتم فيما بعد استعمال مواضيعها للتنبؤ بموضوع النص الإختباري.

ويتم إسناد الأوزان لمواضيع النصوص المجاورة بحساب التشابه الموجود بين هذه النصوص ونص الاختبار «ن». ولقياس هذا التشابه يمكن استعمال المسافة الإقليدية أو مسافة جيب التمام. وبعد ذلك تستعمل عتبة cutoff لإيجاد موضوع نص الاختبار (Yang and Liu, 1999).

خوارزمية (TF-IDF)

يقتضي التعرف الموضوعي إيجاد موضوع نص معين (فقرة أو مقالا). ويتم ذلك بالاعتماد على مدونات التدريب الخاصة بكل موضوع، والتي تمثل خصائصه. بحيث تتم مقارنة هذه الخصائص مع تلك المتعلقة بالنص.

إن فكرة خوارزمية (TF-IDF) تقوم أساسا على تمثيل كل وثيقة d بمتجه $D = (d1, d1, ...d_{|V|})$ في فضاء المتجهات، حيث يرمز $|V|$ إلى حجم مجموعة المفردات. ويتم حساب مركبات المتجه عن طريق ضرب تكرار اللفظة $TF(w, d)$، الذي هو عبارة عن عدد المرات التي تظهر فيها اللفظة w في الوثيقة d، بعكس تكرار الوثيقة IDF(w) (Salton, 1991) (Seymore and Rosenfeld, 1997 ???). ويمثل تكرار الوثيقة $DF(w)$ عن عدد الوثائق التي تظهر فيها اللفظة w مرة واحدة على الأقل. وتعرف القيمة di بوزن اللفظة wi في الوثيقة d وتعطى كالآتي:

$di = TF(w,d) * IDF(w)$ مع $IDF(w) = log(N/DF(w)$ و N هوعدد الوثائق.

ولحساب التشابه $sim(D_j, D_i)$ الموجود بين الوثيقة Di و الموضوع Dj تستعمل المعادلة رقم (1). وتنسب الوثيقة إلى الموضوع الذي يحصل على أكبر قيمة تشابه $sim(D_j, D_i)$.

$$sim(D_j, D_i) = \frac{\sum_{k=1}^{|V|} d_{jk} d_{ik}}{\sqrt{\sum_{k=1}^{|V|} (d_{jk})^2 \sum_{k=1}^{|V|} (d_{ik})^2}} \quad (1)$$

مصنف الزناد

تعرف مجموعة زنادات كلمة ما بأنها الكلمات التي لديها ترابط قوي بينها وبين تلك الكلمة (Haton *et al*., 2006) (Rosenfeld, 1994) (Abbas, 2008). ويتم حساب هذه الزنادات باستعمال المعلومة المتبادلة المتوسطة لكل زوج من الكلمات التي تنتمي إلى مجموعة المفردات V_i. والزنادات الأكثر أهمية هي التي تملك قيم كبرى للمعلومة المتبادلة المتوسطة الموافقة لكل زوج من الكلمات (GuoDong and KimTeng, 1999) (Tillman and Ney, 1996). وبالتالي يصبح كل موضوع ممثلا بعدد من الزنادات المستخرجة من مدونة التدريب الخاصة به. فعلى سبيل المثال فإن احتمال أن

تكون المعلومة المتبادلة المتوسطة لكلمتي "سباحة" و"رياضي" أكبر من تلك المتعلقة بكلمتي "سباحة" و"بترول". ويتم حساب المعلومة المتبادلة المتوسطة *AMI* لكلمتين *a* و*b* كالتالي:

$$AMI(a,b) = p(a,b)\log\frac{p(a,b)}{p(a)p(b)} + p(\bar{a},b)\log\frac{p(\bar{a},b)}{p(\bar{a})p(b)} + p(a,\bar{b})\log\frac{p(a,\bar{b})}{p(a)p(\bar{b})} + p(\bar{a},\bar{b})\log\frac{p(\bar{a},\bar{b})}{p(\bar{a})p(\bar{b})} \quad (2)$$

$p(a,b)$ عدد الوثائق التي وجد فيها *a* و*b* معا.
$p(a,\bar{b})$ عدد الوثائق التي وجد فيها a بدون *b*.
$p(\bar{a},\bar{b})$ عدد الوثائق التي لا يوجد فيها *a* و*b* معا.
$p(\bar{a})$ عدد الوثائق التي لا يوجد فيها *a*.
$p(a)$ عدد الوثائق التي تحتوي *a*.

ويمكن تلخيص المراحل التي يجب إتباعها للتعرف على موضوع نص معين باستعمال مصنف الزناد كما يلي:

في مرحلة التدريب

1. إيجاد زنادات كل كلمة w_k تنتمي إلى مجموعة المفردات V_i الخاصة بالموضوع T_i.
2. إنتقاء عدد محدود من الزنادات الأكثر أهمية التي تمثل الموضوع T_i.

في مرحلة التقييم

1. يتم اسخراج زنادات كل كلمة w_k تنتمي إلى النص التجريبي.
2. حساب القيم Q_i باستعمال مسافة الزناد المعطاة في العبارة (3):

$$Q_i = \frac{\sum_{i,k} AMI(w_k, w_k^i)}{\sum_{l=0}^{n-1}(n-l)} \quad (3)$$

ويشير المتغير i الذي يأخذ قيمه من 1 إلى 6 إلى عدد المواضيع الست السابق ذكرها. بينما يمثل المقام في العبارة (3) تطبيعا لحساب قيم المعلومة المتبادلة المتوسطة. كما تمثل w^i_k الزنادات المميزة للموضوع T_i الموجودة في النص التجريبي.

3. يتم إرفاق الموضوع T_i إلى النص التجريبي بناء على اختيار القيمة القصوى لـ Q_i مع العلم أن $i=1,...,6$.

طرق الترابط بين المصنفات

في بداية الستينيات، أظهرت بعض الأبحاث أنه يمكن اعتماد الترابط بين النماذج كبديل لما يسمى بانتقاء النماذج، عندما يتعلق الأمر بكثير من المسائل الإحصائية (Bates and Granger, 1969) (Dickinson, 1975) (Jacobs, 1995). ثم أصبح الترابط بين النماذج شائعا ويطبق في كثير من البحوث. ومن بين طرق الترابط نذكر الترابط الخطي و شبكة العصبونات.

ويكمن هدف الترابط بين النماذج بصفة عامة في استغلال خصوصية كل نموذج. فالاختلاف الذي يميز كل واحد من هذه النماذج والطريقة التي يستخدمها كل نموذج لمعالجة المعلومة مختلفة. فمثلا، مصنف الزناد يعتمد على حساب المعلومة المتبادلة المتوسطة وعلى حساب مسافة الزناد المعطاة في العبارة (3)، كما يستخدم مجموعة مفردات خاصة بكل موضوع. بينما يقوم مصنف (TF-IDF) باستعمال مجموعة مفردات عامة، كما تستخدم مسافة جيب التمام المعبر عنها في المعادلة (1) لمعرفة موضوع نص الاختبار. أدناه طرق الترابط الثلاث: تصويت الأغلبية، تصويت الأغلبية المحسن و الترابط الخطي.

تصويت الأغلبية

يعتبر تصويت الأغلبية من طرق الترابط البسيطة، بحيث يعتمد على طريقة التصويت. لاستخراج موضوع نص معين w_1^N يحتوي على N كلمة، يتم فرز نتائج طرق التعرف الموضوعي المستعملة بحيث ترفق كل طريقة موضوعا من بين المواضيع. ويعتبر الموضوع الذي يسجل أكبر عدد من الأصوات هو الذي يميز النص w_1^N.

تصويت الأغلبية المحسن

يعتبر تصويت الأغلبية المحسن طريقة تجريبية بحتة، فهي لا تختلف عن الطريقة التقليدية (تصويت الأغلبية) كثيرا، إذ يكمن الفرق في إضافة أوزان للطرق المؤلفة. أما وظيفة الأوزان فهي ترجيح القرار الصائب، حيث أن هناك بعض الحالات التي تؤدي فيها كل من الطرق الثلاث إلى إرفاق موضوع مختلف للنص التجريبي. فمثلا لو أن مصنف الزناد أرفق الموضوع «إقتصاد» و خوارزمية الجار الأقرب أدت إلى انتقاء الموضوع «رياضة» ومصنف (TF-IDF) إلى الموضوع "ثقافة" فإنه لا يمكن لطريقة تصويت الأغلبية التقليدية أن ترفق أيا من المواضيع الثلاثة.

ويتم حساب الأوزان على ضوء النتائج المحققة

باستعمال "خ.ج.أ"، (TF-IDF) ومصنف الزناد. حيث يؤخذ بعين الاعتبار قيمة التذكير المتحصل عليها باستعمال كل طريقة بالنسبة لكل موضوع. وبالتالي كلما كانت قيمة التذكير كبيرة كان الوزن أكبر والعكس صحيح. من أجل ذلك تم تخصيص 5 مدونات للنصوص التجريبية، يمثل حجم كل منها 10 % من مدونة التدريب، تم تسجيل قيم التذكير الناتجة عن الطرق الثلاث والخاصة بكل موضوع، ليتم استخدامها فيما بعد لحساب الأوزان. إن استخدام خمس مدونات تجريبية هو للحصول على نتائج تتميز بمصداقية أكبر.

الترابط الخطي

إن الترابط الخطي من أهم الطرق التقليدية المعروفة والمطبقة في ميادين عديدة. وترتكز عملية الترابط على موازنة متوسط النتائج المرفقة إلى المواضيع من طرف كل طريقة. ليكن $\mathrm{Ci}(w_1^N)$ النتيجة المرفقة باستعمال الترابط الخطي للموضوع T_i بالنسبة للنص w_1^N. يمكن الحصول على هذه القيمة باستعمال المعادلة (3):

حيث يعبر K عن عدد طرق التعرف الموضوعي الخاضعة لعملية الترابط. بينما يعبر r_{ij} عن النتيجة المرفقة للموضوع T_i باستعمال الطريقة رقم K.

التجارب والتعليق على النتائج

من خلال التجارب التي أجريت في هذا البحث تم تقييم كفاءة كل طريقة على حدة باستعمال بعض القياسات المعروفة مثل التذكير، الدقة وF1، وبعد ذلك تم دراسة أثر طرق الترابط بين المصنفات على النتائج. ويمكن تعريف قياس التذكير بالنسبة لنصوص أو وثائق تجريبية تعالج الموضوع T_i بأنه حاصل قسمة عدد الوثائق المعلمة بعلامة صحيحة T_i بطريقة آلية على العدد الكلي للوثائق ذات العلامة الفعلية T_i. أما الدقة فتعرف بأنها حاصل قسمة عدد الوثائق المعلمة بعلامة صحيحة T_i بطريقة آلية على عدد الوثائق المعلمة بعلامة T_i بطريقة آلية. ويجمع القياس F1 بين القياسين السابقين من أجل حساب كفاءة الطرق المستعملة ويؤدي إلى معرفة عدد الوثائق المعلمة بالموضوع الصحيح بدقة. ويمكن اعتبار طريقة ما للتعرف الموضوعي بأنها جيدة عندما تكون قيمتا التذكير والدقة متقاربتين.

كفاءة المصنفات

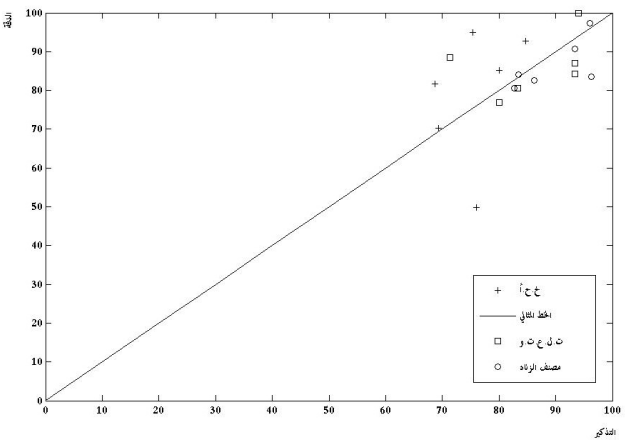
تتفاوت كفاءة المصنفات لأن كلا منها يعتمد على طريقة معينة. فمثلا مصنف (TF-IDF) يقوم بتمثيل الوثائق بالطريقة المشار إليها في الجزء "خوارزمية (TF-IDF)"، بعد ذلك يتم حساب المسافة بين المتجه الممثل للنص التجريبي و بين المتجهات الممثلة لكل موضوع، بينما يعتمد مصنف الزناد على الترابط الكائن بين كل زوج من الكلمات وهذا باستعمال المعلومة المتبادلة المتوسطة.

بلغ متوسط قيمة التذكير بالنسبة لخوارزمية الجار الأقرب % 75.66جدول (3). وتعتبر هذه النتيجة مقبولة إذا ما تم الأخذ بعين الاعتبار حجم مجموعة المفردات المتواضع والذي بلغ 800 مفردة، وإذا ما تم اعتبار النتائج المعروضة في (Yang and Liu, 1999) كمرجع. والملاحظ أن النتائج تختلف من موضوع لآخر حسب الطريقة المستعملة، و كذلك لأن بعض المواضيع متشعبة وتحتاج إلى تجزئتها إلى مواضيع ثانوية.

جدول 3. خوارزمية الجار الأقرب، (حجم مجموعة المفردات 800).

المواضيع / الكفاءة(%)	التذكير	الدقة	F1
ثقافة	76	49.78	60.15
دين	75.33	94.95	84.00
اقتصاد	68.66	81.74	74.63
أخبار محلية	69.33	70.27	69.79
أخبار عالمية	80	85.11	82.47
رياضة	84.66	92.70	88.49

أما طريقة "ت.ل.ع.ت.و" فقد كانت أكفأ من "خ.ج.أ"، حيث بلغ متوسط قيم التذكير 85.88%، أي بفارق يعادل حوالي 10 %جدول (4). وكان الموضوع «ثقافة» قد سجل أقل نسبة للتذكير 71.33 % وأكبرها بالنسبة لموضوع «الرياضة» 94%. أما مصنف الزناد فقد أدى إلى تحسن في النتائج مقارنة بالطريقتين الأخريين، إذ بلغت القيمة المتوسطة للتذكير 89.67% جدول (5). و سبب فعالية مصنف الزناد هو استغلال الترابط الموجود بين الكلمات المكونة لمدونات التدريب، ومن ثم استخراج الزنادات المميزة لكل موضوع. وتعتبر المصنفات جيدة إذا تقاربت قيمتا كل من التذكير والدقة، ويمكن من خلال الشكل 1 استنتاج أن مصنف الزناد هو الأحسن ثم يليه مصنف (TF-IDF) وأخيرا "خ.ج.أ". إذ أن اقتراب النقاط من الخط المثالي (المتصل) دليل على جودة المصنف.



شكل 1. الدقة بدلالة التذكير بالنسبة للمصنفات الثلاث.

جدول 4. مصنف (TF-IDF) (حجم مجموعة المفردات 800).

المواضيع / الكفاءة(%)	التذكير	الدقة	F1
ثقافة	71.33	88.43	78.96
دين	93.33	86.95	90.02
اقتصاد	83.33	80.64	81.96
أخبار محلية	80	76.92	78.42
أخبار عالمية	93.33	84.33	88.60
رياضة	94	100	96.90

جدول 5. مصنف الزناد، حجم مجموعة المفردات 300 لكل موضوع.

المواضيع / الكفاءة(%)	التذكير	الدقة	F1
ثقافة	82.66	80.55	81.59
دين	96.33	83.56	89.49
اقتصاد	83.50	84.05	83.77
أخبار محلية	86.25	82.53	84.34
أخبار عالمية	93.33	90.66	91.97
رياضة	96	97.33	96.66

أثر طرق الترابط على النتائج

إن الدافع من وراء الترابط هو محاولة إيجاد نوع من التكامل بين مجموعة الطرق المؤلفة. وسنعرض في فيما يلي النتائج المحققة، باستعمال طرق الترابط الثلاث، وهي: تصويت الأغلبية، تصويت الأغلبية المحسن و الترابط الخطي. ففي الجدول 6 قمنا بعرض قيم التذكير، الدقة و F1 المتعلقة باستعمال تصويت الأغلبية فكانت النتيجة أحسن من تلك المتعلقة بخوارزمية الجار الأقرب، لكنها لم تؤدي إلى تجاوز كفاءة مصنفي الزناد و (TF-IDF).

جدول 6. النتائج باستعمال تصويت الأغلبية.

المواضيع / الكفاءة(%)	التذكير	الدقة	F1
ثقافة	69.5	95	80.27
دين	90	88.50	89.24
اقتصاد	80.66	90	85.07
أخبار محلية	78.33	98.66	87.32
أخبار عالمية	92.33	89.65	90.97
رياضة	93.33	99.90	96.50

أما تصويت الأغلبية المحسن فقد تجاوزت كفاءته قيمة التذكير 90.76 %. وكما أشرنا في الجزء 4.2، يتم استنباط الأوزان من خلال حساب متوسط قيم التذكير الناتجة عن استعمال مدونات التجريب الخمس. وهكذا تكون كل طريقة مزودة بالوزن الملائم بالنسبة لكل موضوع. فيكون الوزن كبيرا لطريقة ما إذا كان متوسط قيم التذكير كبيرا، والعكس صحيح.

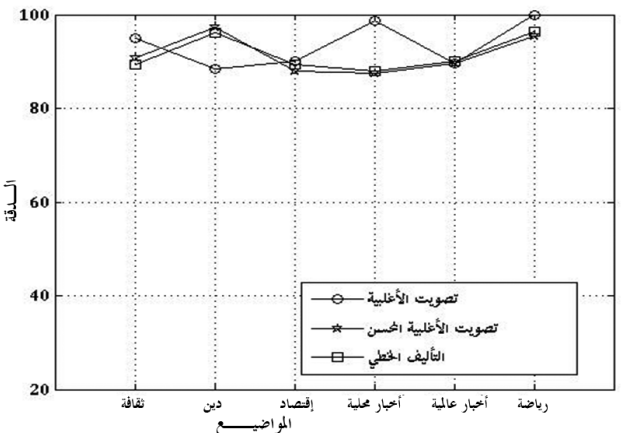
جدول 7. النتائج باستعمال تصويت الأغلبية المحسن.

المواضيع / الكفاءة(%)	التذكير	الدقة	F1
ثقافة	84.50	90.66	87.47
دين	94.66	97.33	95.97
اقتصاد	89.45	88	88.71
أخبار محلية	88.33	87.50	87.91
أخبار عالمية	93.33	89.66	91.45
رياضة	94.33	95.55	94.93

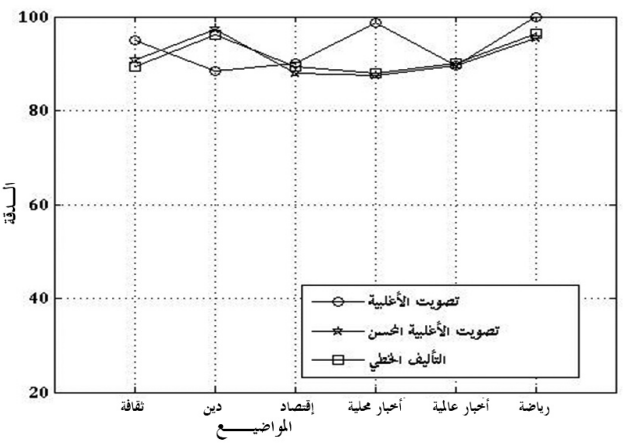
بينما أدى الترابط الخطي إلى نتائج أفضل بحيث بلغ متوسط التذكير 92.83 %. ويمثل الشكلان 2 و 3 منحنيات للمقارنة بين كفاءة كل من تصويت الأغلبية المحسن، تصويت الأغلبية والترابط الخطي. فمن خلال الشكل 2، نستطيع بوضوح ملاحظة منحنى التذكير وتفوق الترابط الخطي بشكل طفيف على تصويت الأغلبية المحسن، ويأتي في الأخير تصويت الأغلبية.

جدول 8. النتائج باستعمال الترابط الخطي.

المواضيع / الكفاءة(%)	التذكير	الدقة	F1
ثقافة	88.66	89.40	89.02
دين	95	96.20	95.59
اقتصاد	92.33	89.33	90.80
أخبار محلية	89.66	88	88.82
أخبار عالمية	94.66	90	92.27
رياضة	96.66	96.33	96.49



شكل 2. قيم التذكير الناتجة باستعمال طرق الترابط الثلاث بالنسبة لكل موضوع.



شكل 3. الدقة الناتجة باستعمال طرق الترابط الثلاث بالنسبة لكل موضوع.

كما يستنتج من خلال الشكل 3، تساوي الترابط الخطي و تصويت الأغلبية المحسن من حيث الدقة، بينما سجل تصويت الأغلبية نتائج أفضل للدقة بالنسبة لكل المواضيع ماعدا الموضوع "دين".

الخاتمة

بينت النتائج التي تم الحصول عليها فعالية بعض الطرق مقارنة بالبعض الآخر وذلك باستعمال أحجام صغيرة لمجموعة المفردات. فقد تجاوزت كفاءة مصنف الزناد كلا من (TF-IDF) و"خ.ج.أ"، حيث بلغ متوسط قيم التذكير لديه 89.70%.

ولتحسين هذه النتائج، تم الترابط بين المصنفات الثلاث باستعمال ثلاث طرق هي: تصويت الأغلبية، تصويت الأغلبية المحسن والترابط الخطي. تصويت الأغلبية لم يؤد إلى تحسين الكفاءة حيث بلغت القيمة المتوسطة للتذكير 84%، وهذا مقارنة مع كفاءة أفضل مصنف وهو مصنف الزناد. أما تصويت الأغلبية المحسن فقد أفضى إلى نتائج أفضل من تلك المتعلقة بمصنف الزناد بنسبة تفوق 1%. بينما أدى الترابط الخطي حصل على قيمة متوسطة للتذكير بلغت 92.83%. وتجدر الإشارة إلى أن الترابط الخطي يعتمد على موازنة النتائج المحصل عليها من طرق التعرف الموضوعي، بينما يأخذ تصويت الأغلبية المحسن بعين الإعتبار كفاءة كل طريقة تجاه كل موضوع. ويمكن لبعض العوامل، والتي لم يتم التطرق إليها في هذا البحث أن تجعل النتائج أفضل، مثل استخراج جذور الكلمات لإظهار تلك المشتقة من مصدر واحد، واعتبارها بالتالي كلمة واحدة مما يتيح حساب تكرارات الكلمات بدقة. وكذلك فإن الزيادة في حجمي المدونة ومجموعة المفردات تؤدي إلى تمثيل أفضل لكل موضوع.

النتائج التي تم الحصول عليها في هذا المجال وهو التعرف الموضوعي للنصوص العربية يمكن توظيفها وتجريبها في إحدى التطبيقات، كالترجمة الآلية أو التعرف الآلي على الكلام، لمعرفة إلى أي مدى يمكن الاعتماد علي التعرف الموضوعي.

المراجع

Abbas, M (2008) *Topic Identification for Automatic Speech Recognition*. Ph.D. thesis, National Polytechnic School of Algiers.

Abbas, M, Smaili, K, and **Berkani, D** (2009) Multi-Category Support Vector Machines for Identifying Arabic Topics. *Journal of Research in Computing Science* **41:** 217-226.

Combining Protein Secondary Structure prediction Models with Ensemble Methods of Optimal Complexity. *Neurocomputing* **56:** 305-327.

GuoDong,Z, and **KimTeng,L** (1999) Interpolation of n-gram and mutual information based trigger pair language models for Mandarin speech recognition. *Computer Speech and Language* **13:** 125-141.

Haton, JP, Cerisara, C, Fohr, D, Laprie, Y, and **Smaili, K** (2006) Reconnaissance automatique de la parole: du signal à son interprétation, Paris, France.

Iwayama, M, and **Tokunaga, T** (1995) Cluster-based text categorization: a comparison of category search strategies. *Proceedings of the 18th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95)*, Seattle, Washington, USA, pp. 273-281.

Jacobs, RA, (1995) Methods for Combining Experts' Probability Assessments. *Neural Computation* **7:** 867-888.

Joachims, T (1998) Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *In* **Nédellec, C** and **Rouveirol, C (eds.)** *Proceedings of the European Conference on Machine Learning*. Berlin, pp. 137-142.

Lee, Y, Lin, Y, and **Wahba, G** (2004) Multicategory Support Vector Machines: Theory and Application to the Classification of Microarray Data and Satellite Radiance Data, *Journal of the American Statistical Association* **99(465):** 67-81.

Lewis, DD, and **Ringuette, M** (1994) A Comparison of two Learning Algorithms for Text Categorization. *In the Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94)*, Las Vegas, USA, pp. 81-93.

Masand, B, Lino, G, and **Waltz, D** (1992) Classifying news stories using memory based reasoning. *In the Proccedings of the 15th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval*

Abbas, M, and **Smaili, K** (2005) Comparison of Topic Identification Methods for Arabic Language. *Proceedings of the International conference on Recent Advances in Natural Language Processing (RANLP05)*, Borovets, Bulgary, pp. 14-17

Baoli, L, Yuzhong, C, and **Shiwen, Y** (2002) A Comparative Study on Automatic Categorization Methods for Chinese Search Engine. *Proceedings of the 8th Joint International Computer Conference*. Hangzhou, Zhejiang University Press, China, pp. 117-120.

Bates, JM, and Granger, CWJ (1969) The Combination of forecasts. *Operational Research Quarterly* **20:** 451-468.

Creecy, RH, Masand, BM, Smith, SJ, and **Waltz,DL** (1992) Trading MIPS and Memory for Knowledge Engineering: Classifying Census Returns on the Connection Machine. *Communication of the ACM*, **35:** 48-63.

Dickinson, JP (1975) Some comments on the combination of forecasts. *Operational Research Quarterly* **26:** 205-210.

El-Halees, A (2007) Arabic Text Classification Using Maximum Entropy. *The Islamic University Journal (Series of Natural Studies and Engineering)* **15(1):** 157-167.

El-Kourdi,M,Bensaid,A, and **Rachidi,T** (2004) Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm. *In the Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, Geneva, Switzerland, pp. 51-58.

Frakes, WB, and **Baeza-Yates, R** (1992) Information Retrieval: Data Structures and Algorithms. Prentice Hall, Englwood Cliffs, NJ, pp. 1-12.

Fuhr, N, Hartman, S, Lustig, G, Schwantner, M, and **Tzeras, K** (1991) A rule-based Multistage Indexing Systems for Large Subject fields. *In the Proceedings of RIAO'91*, Barcelona, Spain, pp. 606-623.

Guermeur, Y, Pollastri, G, Elisseeff, A, Zelus, D, Paugam-Moisy, H, and **Baldi, P** (2004)

Symposium on Document Analysis and Information Retrieval (SDAIR'95), Las Vegas, pp. 317-332.

Yang, Y (1994) Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval. *Proceedings of the 17th Ann. Int. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*, Dublin, Ireland, pp. 13-22.

Yang, Y and **Liu, X** (1999) A re-examination of text categorization methods. *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'99*, Berkeley, CA, USA, pp. 42-49.

Yang, Y, and **Pedersen, JO** (1997). A comparative study on feature selection in text categorization. *Proceedings of the 14th International Conference on Machine Learning*, San Francisco, USA, pp. 412-420.

(SIGIR'92), Copenhagen, Denmark, pp. 59-64.

Moulinier, I, (1997) Is Learning Bias an issue on Text Categorization Problem? Technical Report, LAFORIA-LIP6, University Paris VI.

Ng, HT, Goh, WB, and **Low, KL** (1997) Feature selection perceptron learning, and a usability case study for text categorization. *Proceedings of the 20th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*, Philadelphia, USA, pp. 67-73.

Pinto D, Jiménez H, and **Rosso P** (2006) Clustering abstracts of scientific texts using the Transition Point technique. *In the Proceedings of the 7th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2006, Springer-Verlag, LNCS(3878)*, pp. 536-546.

Rosenfeld, R (1994). *Adaptive Statistical Language Modeling: A Maximum Entropy Approach*. Ph.D. thesis, Carnegie Mellon University.

Salton, G (1991) Developments in Automatic Text Retrieval. *Science* 253**:** 974-979.

Seymore, K, Chen, S, and **Rosenfeld, R** (1998). Nonlinear interpolation of topic models for language model adaptation. *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, pp. 2503-2506.

Seymore, K, and **Rosenfeld, R** (1997) Using Story Topics for Language Model Adaptation. *Proceedings of the European Conference on Speech Communication and Technology, Rhodes*, Greece, pp. 1987-1990.

Tillman, C, and **Ney, H** (1996) Selection criteria for word trigger pairs in language modeling. *In:* **Miclet, L** and **de la Higuera, C** (**eds**.) Grammatical inference: Learning syntax from sentences. Lecture Notes in Artificial Intelligence N. 1147, pp. 95-106.

Wiener, E, Pedersen, JO, and **Weigend, AS** (1995) A neural network approach to topic spotting. *Proceedings of the Fourth Annual*