استخدام ذخائر النصوص لاستخلاص المصطلحات المتخصصة

د. عبدالمحسن بن عبيد الثبيتي
مركز المعلومات

الكلمات المفتاحية: معامل الغرابة، استخلاص المعلومات، ذخائر النصوص العربية، معالجة اللغات الطبيعية، التعامل الآلي مع النصوص.

ملخص: مع تزايد النتاج المعرفي في المجالات المتخصصة وتزايد النصوص العربية التي تغطي هذه المجالات، تتزايد الحاجة للتعامل الآلي مع هذه النصوص والمعارف التي تتضمنها. إحدى طرق التعامل والاستفادة من هذه المعارف هي استخلاص المصطلحات المتخصصة بطريقة آلية بحيث يمكن استخدامها كأداة مساعدة لاختيار مفردات المعاجم المتخصصة أو الفهرسة الآلية للنصوص المتخصصة في محركات البحث كما يمكن استخدامها أيضا كطريقة لاختيار المفردات الأنسب لعمليات تصنيف النصوص العربية المتخصصة. تطرقت هذه الورقة إلى توصيف طريقة إحصائية بسيطة لاستخلاص المصطلحات المتخصصة ذات المفردة الواحدة اعتمادا على ذخائر النصوص. حيث تعتمد الطريقة على حساب "معامل الغرابة" الذي يقيس مدى ابتعاد استخدام المفردة في مدونة المجالات المتخصصة عنه في مدونة المجالات العامة أو اللغة اليومية. وقد تم تقييم النتائج المستخلصة بواسطة مختصين في المجالات المتخصصة التي تمت دراستها وهي الذكاء الصناعي والفيزياء. وقد أظهرت الطريقة المقترحة نتائج مشجعة في المجالات المتخصصة التي تمت دراستها.

مقدمة

تشهد المعرفة الإنسانية نموا وتزايدا يتسم بالتسارع المطرد ونزعته إلى التخصص أكثر من ذي قبل. و يُعزى هذا التسارع والتوجه نحو التخصص إلى الاكتشافات العلمية بدور كبير والى تطور وتوافر وسائل تدوين المعرفة ونشرها. إن هذا النزوع إلى التخصص بشدة في النتاج المعرفي وتزايده يستدعي إيجاد طرق و وسائل فاعلة للتعامل معه خلال مراحل إنتاجه وتداوله وحفظه و استرجاعه. يمثل استخلاص المصطلحات المتخصصة آليا الخطوة الأولى نحو التعامل مع هذا النتاج كقاعدة وأساس لعمليات أخرى مهمة مثل بناء المعاجم والقواميس المتخصصة، المساعدة في عمليات الترجمة الآلية للنصوص العلمية والفنية ، استرجاع وفهرسة النصوص العلمية والفنية ، اختيار الخصائص اللغوية الملائمة لعمليات التصنيف الآلي للنصوص ، بناء شبكة العلاقات بين المفاهيم في المجالات المتخصصة.

تمثل اللغة المكتوبة الوسيلة المباشرة للتعامل مع هذا النتاج المعرفي . فاللغة تلعب دورا فاعلا في نقل المعرفة وفي صياغة وبناء الأفكار. فالمصطلح العلمي ما هو إلا إشارة إلى مفهوم أو فكرة أو ظاهره. و لهذا يسعى العلماء والتقنيون المنتجون للمعارف المتخصصة بصورة رئيسية إلى صياغة الأفكار والاكتشافات الجديدة أو الظواهر المشاهدة بطريقة تضمن فهمها بصورة واضحة وغير مشوشة وكذلك إلى عدم الخلط بين الاكتشافات الجديدة وما تم اكتشافه سابقا. ولتحقيق الوضوح وعدم اللبس ، يسعى العلماء إلى استخدام الكلمات والمصطلحات المتفق عليها بصورة متكررة. فالمصطلح العلمي لا يمكن أن يحمل أكثر من معنى وذلك تحقيقا لمبدأ الوضوح وعدم الخلط. وعندما تكون هناك اكتشافات أو ظواهر تعجز المفردات أو المصطلحات الحالية عن إيضاحها أو تمييزها عن غيرها فإن الأساليب تتعدد لصياغة مصطلحات جديدة مناسبة. تشمل هذه الأساليب الاستعارة من مجالات أو لغات أخرى، دمج مصطلحات مختلفة مع بعضها وتكوين مصطلحات من مفردتين أو أكثر و أخيرا اختراع كلمات جديدة تماما لم تكن معروفة (Ahmad and Gillam، ٢٠٠١) وهذا نادر الحصول.

هناك العديد من الأساليب المقترحة لاستخلاص المصطلحات ولكنها في عمومها تعتمد طريقتين. الطريقة الأولى تعتمد على اكتشاف التركيب النحوي للمصطلحات (Amigó et al، ٢٠٠٤) والتي تكون أغلبها أسماء أو جمل اسميه (Frantzi et al، ٢٠٠٠). فالمجالات المتخصصة لها تراكيب نحوية مقيدة كما هو الحال مثلا في وثائق براءات الاختراع الأمريكية (Ahmad and Al-Thubaityl، ٢٠٠٤). و استخدام مثل هذه الطريقة يتطلب استخدام أدوات خاصة بالإعراب الآلي للمفردات الموجودة في النص. وتعاني أدوات معالجة اللغة العربية آليا من شح وندرة في مثل هذا الدراسات وتوافر هذه الأدوات عدا بعض الدراسات التي بدأت حديثا (Khoja، ٢٠٠٣). أما الطريقة الثانية فتعتمد على الطرق الإحصائية لاستخلاص المصطلحات في النصوص المتخصصة (Frantzi and

الندوة الدولية الأولى عن الحاسب واللغة العربية
© مدينة الملك عبدالعزيز للعلوم والتقنية وجمعية الحاسبات السعودية

(Pante and Lin، ٢٠٠١) (Ananiadou، ١٩٩٩). وفي كلتا الطريقتين يبرز استخدام ذخائر النصوص (المدونات) كطريقة تثبت صلاحية الطرق المستخدمة لاستخلاص المصطلح العلمي آليا كون المدونات تعطي الدليل على التراكيب النحوية الأكثر استخداما التي يظهر فيها المصطلح أو حتى التوزيع الإحصائي لمفرداته.

تفترض هذه الدراسة اختلاف التوزيع الإحصائي للمفردات العربية في المجالات المتخصصة عنه في المجالات العامة أو اللغة العربية اليومية وذلك اعتمادا على الدراسات السابقة الخاصة بمدونات المجالات المتخصصة في اللغة الانجليزية (Ahmad et al ،١٩٩٤) ، فمفردات المصطلحات العلمية والتقنية تظهر بصورة أكثر في لغة النصوص المتخصصة أكثر من ظهورها في اللغة اليومية أو غير المتخصصة. للتحقق من هذه الفرضية تم إنشاء مدونة نصوص تمثل اللغة اليومية وسوف نعتبرها المدونة المرجع ومدونة أخرى تمثل مجالين متخصصين والتي سنقوم باستخلاص المصطلحات المتخصصة منها اعتمادا على اختلاف التوزيع الإحصائي للمفردات في المدونات المتخصصة عنها في المدونة الممثلة للغة اليومية.

توصيف المدونات

يمكن تعريف المدونات ببساطة بأنها مجموعة من النصوص يتم تجميعها لغرض معين. وتختلف الأغراض التي تبنى من اجلها المدونات مما يجعل تكوينها والفترات الزمنية التي تغطيها تختلف باختلاف أغراض بنائها. وهناك الكثير من الدراسات التي تتناول بناء المدونات والخصائص التي يجب أن تتوفر فيها (Biber et al، ٢٠٠٢) (Sinclair، ١٩٩٥) (Atkins et al، ١٩٩٢). ومن أهم العوامل التي تراعى عند بناء المدونات تمثيلها لواقع اللغة أو لمجال الدراسة إضافة إلى حجمها الذي يقاس بعدد الكلمات التي تحويها. وتحقيق هذين الشرطين منوط بعوامل كثيرة من أهمها توافر النصوص بصورة الكترونية. وللأسف فإنه لا يوجد حاليا دراسات متخصصة في المدونات العربية أو أدوات متخصصة للتعامل معها مثلما هو الحال في المدونة الوطنية البريطانية (BNC) يمكن اعتبارها كمرجع للدراسات اللغوية أو الحاسوبية المتعلقة باللغة. ولغرض إتمام هذه الدراسة تم تكوين مدونتين رئيسيتين. الأولى تمثل اللغة اليومية مكونة من حوالي ١,١ مليون كلمة و المدونة الأخرى تمثل المجالات المتخصصة يقارب حجمها ٥٨ ألف كلمة. لافتين نظر القارئ الكريم إلى صعوبة الحصول على نصوص الكترونية باللغة العربية في المجالات المتخصصة يمكن التعامل ومعالجتها آليا إضافة إلى الوقت الكبير الذي يتطلبه المسح الضوئي وتجهيز النصوص للتحليل الآلي.

مدونة المجالات المتخصصة (اللغة اليومية)

تم إنشاء مدونة تمثل اللغة اليومية مكونة من (١٠٨٧٣١٨) كلمه تحوي (١٠٣٥٩٩) مفردة مختلفة بهيكلة مشابهة لبنية مدونة (Collins-Birmingham) للنصوص الانجليزية. تم جمع نصوص المدونة من شبكة الانترنت من مواقع سعوديه ولكتاب سعوديين في الأغلب موزعة كالتالي: الصحف اليومية (٥٧٪) ، المجلات الأسبوعية (٢١٪) ، الكتب (١٥٪) ، المطويات التعريفية (٦٪) ، المواقع الشخصية (١٪).

لغرض عرض النتائج بصورة أوضح، يمكننا تقسيم الكلمات إلى مجموعتين كما هو الحال في اللغة الانجليزية (قد يكون هناك مثل هذا التقسيم في اللغة العربية لم يستطع الباحث الاطلاع عليه). الأولى المجموعة المغلقة من الكلمات وهي التي لا يمكن الإضافة إليها أو تغير معانيها وطرق استخدامها أو وظائفها النحوية أو مواقعها من الإعراب مثل أحرف العطف والجر و أسماء الإشارة و أسماء الصلة و ظروف الزمان والمكان وغيرها. والمجموعة الثانية هي المجموعة المفتوحة من الكلمات وهي الكلمات التي يمكن الإضافة أليها وتغير معانيها باختلاف السياق واختلاف مواقعها الإعرابية إضافة إلى اختلاف قبولها وسائغتها باختلاف الزمن وغالبا ما تحوي الأسماء ا والأفعال بأنواعها.

يبين الجدول (١) المائة كلمة الأكثر تكرارا في مدونة اللغة اليومية حيث تمثل ما نسبته ٣٠,٣٪ من حجم المدونة. وقد احتوت على ٧٩ كلمة تنتمي إلى المجموعة المغلقة من الكلمات حيث كانت الثلاثين كلمة الأكثر تكرار كلها تنتمي إلى هذه المجموعة ما عدا كلمة واحدة هي لفظ الجلالة (الله). كما احتوت المائة كلمة الأكثر تكرارا على٢١ كلمة تنتمي إلى المجموعة المفتوحة من الكلمات. وبالإمكان اعتبار الواحد والعشرين كلمه كتوقيع لفظي للمدونة نستطيع من خلالها التعرف على مواضيع واهتمامات المدونة. وبما أن الصحف اليومية السعودية قد شكلت النسبة الأكبر من حجم المدونة فليس من المستغرب ورود كلمات مثل العربية ، المملكة، السعودية، ورئيس ومجلس واسم العلم عبدالله. و بلا شك نستطيع تبين اهتمامات مواضيع المدونة بالمجتمع وبالناس لورود هاتين الكلمتين ضمن الكلمات الأكثر تكرار.

مدونة المجالات المتخصصة

تتكون مدونة المجالات المتخصصة من مدونتين فرعيتين. الأولى مدونة الذكاء الصناعي (٤٢٪) والمدونة الأخرى مدونة الفيزياء (٥٨٪). وتحوي المدونتين على (٥٧٧٩٧) كلمه تتضمن (١٣٢٨٠) مفردة مختلفة. وقد غطت مدونة الذكاء الصناعي مواضيع متعددة وخصوصا موضوعي الشبكات العصبية والأنظمة الخبيرة.كما غطت مدونة الفيزياء مواضيع فيزياء الكم والفيزياء الذرية والجزيئية.

يوضح الجدول (٢) الكلمات المائة الأكثر تكرارا في مدونة الذكاء الصناعي حيث تشكل ما نسبته ٣٧,٥٪ من مجموع كلمات المدونة. وقد احتوت على ٤٩ كلمة تنتمي إلى المجموعة المفتوحة من الكلمات بزيادة تقارب المرتين والنصف مقارنة بمدونة اللغة اليومية، و قد كانت كلها أسماء ما عدا كلمتين كانتا لفعل واحد بصيغتين مختلفتين (يقوم و تقوم). ويتضح لنا أيضا من الجدول (٢) خاصية الإنتاج الصرفي للكلمات الأكثر تكرارا في المدونات المتخصصة (Bauer، ٢٠٠١) مثل ورود بعض الكلمات في صيغة المفرد والجمع مثل(خبير وخبراء) أو بإضافة التعريف أو بدونه مثل(نظام و النظام). ونظرا لان مصطلحات هذا العلم مترجمه في اغلبها فإن بعض المصطلحات الدالة على نفس المعنى وردت بصيغ مختلفة مثل (العصبية والعصبونية). وكما هو الحال بالنسبة إلى مدونة المجالات العامة –بل وبصورة أكثر وضوحا– فإن هذه الكلمات تعتبر التوقيع اللفظي الذي يدل على هوية وسمة المدونة والمواضيع التي تحويها. ولا يختلف الحال في مدونة الفيزياء عنه في مدونة الذكاء الصناعي كما هو واضح في الجدول (٣)

جدول ١ : الكلمات المائة الأكثر تكرارا في مدونة اللغة اليومية

	المفردة	النسبة
1	و، في، من، على، أن، إلى، التي، لا، ما، عن	19.1
2	هذا، الله، الذي، هذه، أو، مع، ان، ذلك، هو، كان	3.4
3	قد، لم، كل، بين، بن، بعد، هي، إن، كما، الى	1.9
4	خلال، العربية، حتى، عبد، له، قال، حيث، المملكة، كانت، غير	1.3
5	لكن، بعض، محمد، عليه، السعودية، إلا، قبل، به، بها ، فيها	1.1
6	يكون، ثم، هناك، لها، أي، تلك، عام، العمل، فيه، إذا	0.9
7	مجلس، أنه، العام، ليس، الدكتور، منها، العالم، رئيس، بل، الذين	0.8
8	عند، مثل، أكثر، يمكن، اليوم، الرياض، تكون، فقد، الملك، الناس	0.7
9	العزيز، عدد، أما، عدم، السعودي، التعليم، عبدالله، عليها، تم، بما	0.6
10	عبدالعزيز، أخرى، جميع، عندما، فإن، دون، المجتمع، بأن، مما، يجب	0.5
	المجموع	30.3

جدول ٢ : الكلمات المائة الأكثر تكرارا في مدونة الذكاء الصناعي

	المفردة	النسبة
1	و، في، من، على، التي، أن، إلى، أو، هذه، **الذكاء**	19.5
2	عن، هي، هذا، هو، ما، لا، مع، **الشبكة**، كل، الذي	4.8
3	**الصناعي**، ذلك، **الإنسان**، حيث، **النظام**، قد، **الاصطناعي**، **الخبير**، **المعرفة**، بين	2.8
4	أي، يمكن، **أنظمة**، لكن، **عملية**، كما، يتم، **الشبكات**، بها، كان	2.2
5	بعد، غير، تكون، **نظام**، **التعرف**، ثم، **المعلومات**، يكون، **الأنظمة**، مثل	1.8
6	**طريق**، **البحث**، **طبقة**، **مجموعة**، **الخبيرة**، **علم**، **قاعدة**، إذا، **تقوم**، لها	1.5
7	**الأوزان**، **الحاسوب**، فيها، **الخبراء**، بعض، **معالجة**، **الطريقة**، **العصبية**، **مجال**، **يقوم**	1.4
8	فإن، يجب، أنه، حتى، **الصوتية**، **العصبونية**، **الكلام**، **المشاكل**، ذات، لم	1.2
9	**التعليم**، **الحاسب**، **العمليات**، **المشكلة**، منها، أكثر، إن، **بواسطة**، **حل**، **مرحلة**	1.2
10	هناك، **المعالجة**، كانت، إلا، **الآلية**، **البشري**، **الكمبيوتر**، **المجال**، **المنطق**، **النظم**	1.1
	المجموع	37.5

جدول ٣ : الكلمات المائة الأكثر تكرارا في مدونة الفيزياء

	المفردة	النسبة
1	و، في،من، على، أن، هذه، إلى، هذا، التي، عن	16.3
2	ان، لا، **النظرية**، يمكن، مع، هو، **الضوء**، هي، **النسبية**، **نظرية**	4.5
3	ما، الذي، **الكم**، بين، **سرعة**، كل، **بالنسبة**، أو، **الزمن**، حيث	3.0
4	ذلك، كان، يكون، **الطاقة**، لكن، **الكون**، كما، بعد، أي، **الكمومية**	2.4
5	الى، فإن، **عام**، **النواة**، حول، تكون، **الإلكترون**، **طاقة**، غير، **الجسيمات**	1.9
6	لم، عند، **الجسيم**، **جسيمات**، **حالة**، كانت، **جسيم**، خلال، سوف، قد	1.6
7	**الأرض**، **المراقب**، **الذرة**، **ذرة**، **بسرعة**، مثل، **العالم**، **الفيزياء**، عندما، **الهيدروجين**	1.4
8	**نموذج**، **العلماء**، **القوة**، **الموجة**، **الفا**، **الخاصة**، **العامة**، هنا، أنه، **تفسير**	1.2
9	**الموجية**، **حركة**، **مبدأ**، **معادلة**، نفس، **الحالة**، حتى، هناك، **المادة**، **بشكل**	1.1
10	ذات، **ميكانيكا**، احد، **الثانية**، **الجسم**، **الطبيعة**، **كتلة**، **الأبعاد**، **الذرية**، **الفضاء**	1.0
	المجموع	34.4

استخلاص المصطلحات من النصوص

لغرض استخلاص المصطلحات العلمية سنقوم بتعريف "معامل الغرابة" والذي يقيس مدى ابتعاد معدل استخدام المفردة في المجال المتخصص عن استخدامها في المجالات العامة أو اللغة اليومية. وقد تم استخدام هذا المعامل في اختيار الخصائص اللغوية المناسبة لعمليات التصنيف الآلي للنصوص العلمية (Ahmad et al، ٢٠٠١) وكذلك ضمن دراسة لتطور المعرفة في المجالات المتخصصة (Al-Thubaity، ٢٠٠٤).

سوف نعرف مدونة اللغة اليومية CG المكونة من العدد n من النصوص ومدونة المجال المتخصص CS المكونة من العدد m من النصوص كالتالي

$$C_G = \{d_{G1}, d_{G2}, d_{G3}, ..., d_{Gn}\} \qquad (1)$$

$$C_G = \{d_{S1}, d_{S2}, d_{S3}, ..., d_{Sm}\} \qquad (2)$$

حيث يعبر الرمز d_{Gy} عن النص y الموجود داخل مدونة اللغة اليومية والرمز d_{Sz} عن النص z الموجود داخل مدونة المجال المتخصص. بالإمكان أيضا التعبير عن أي من المدونتين بدلالة المفردات w وتكرار هذه المفردات f_w كما يلي.

$$C_G = \{(w_{G1}, f_{w_{G1}}), (w_{G2}, f_{w_{G2}}), (w_{G3}, f_{w_{G3}}), ..., (w_{Gi}, f_{w_{Gi}})\} \qquad (3)$$

$$C_S = \{(w_{S1}, f_{w_{S1}}), (w_{S2}, f_{w_{S2}}), (w_{S3}, f_{w_{S3}}), ..., (w_{Sj}, f_{w_{Sj}})\} \qquad (4)$$

تقاس احتمالية ظهور المفردة (*معدل استخدام المفردة*) w في أي من نصوص مجالات اللغة اليومية أو نصوص المجالات المتخصصة بناء على ملاحظة ظهورها في أي من المدونتين كالتالي

$$p(w_n) = \frac{f_{w_n}}{\sum f_w} \qquad 0 \le p(w_n) \le 1 \qquad (4)$$

وبناء على ما تقدم يمكننا تعريف معامل الغرابة W رياضياً كالتالي

$$W = \frac{p_S(w)}{p_G(w)} = \frac{\frac{f_w}{\sum f_S}}{\frac{f_w}{\sum f_G}} \qquad (5)$$

وبتعريف أكثر بساطه لغير المهتمين بالتمثيل الرياضي يمكننا أن نعبر عن معامل الغرابة كالتالي:

$$\textit{معامل الغرابة} = \frac{\textit{معدل استخدام المفردة في المجال المتخصص}}{\textit{معدل استخدام المفردة في المجالات العامة}}$$

وبطبيعة الحال فإن قيمة معامل الغرابة تزداد كلما ازداد تكرار الكلمة في المجال المتخصص عنه في المجالات العامة والعكس بالعكس فتصبح قيمته مالا نهاية إذا لم يظهر في مدونة المجالات العامة وتصبح قيمته صفرا إذا لم يظهر في مدونة المجالات المتخصصة. وهنا يجب توضيح أن استخدام معامل الغرابة لاستخلاص المصطلحات المتخصصة من المدونات العربية المتخصصة يخضع للاعتبارات التالية:

- أن الكلمات التي يطبق عليها معامل الغرابة تنتمي للمجموعة المفتوحة من الكلمات.

- أن تكرار الكلمات مهم جدا في اعتبار الكلمة مصطلحا مقبولا أو متعارفا عليه في المجال المتخصص فهناك علاقة مباشره بين التكرار وقبول الكلمة في المجال (Quirk، ١٩٨٥).

- أن تكون قيمة معامل الغرابة عالية للكلمات التي تراعي الفقرتين السابقتين أعلاه.

وفي الحقيقة فان اختيار الحد الأدنى الأمثل من التكرار للكلمات واختيار القيمة الدنيا المثلى لمعامل الغرابة أمر يختلف اختياره من شخص لأخر و من مدونة إلى أخرى و لا توجد طريقه لمعرفة ذلك غير التجربة.وقد تم بداية اختيار القيمة الدنيا للتكرار النسبي ب(٠,٠٠٠٢) وقيمة معامل الغرابة بمائه لاستخلاص المصطلحات وقد نتج عن هذا الاختيار استخلاص ٦٣ مفردة من مدونة الذكاء الصناعي و١٤٤ مفردة من مدونة الفيزياء.الجدولين (٤) و(٥) يبينان الكلمات العشرين ذات معامل الغرابة الأعلى في مدونتي الذكاء الصناعي والفيزياء على التوالي.

جدول ٤ : الكلمات المائة الأكثر تكرارا في مدونة اللغة اليومية

الذكاء الصناعي			
الكلمة	معامل الغرابة	الكلمة	معامل الغرابة
العصبونية/العصبية	∞	درجة	1038
الاصطناعي	∞	الاوزان	507
الضبابي	∞	محاكاة	415
المدخلات	∞	الأذرع	311
النظم	∞	البرمجة	294
استنتاجات	∞	مسموعة	277
الانحيازات	∞	شبكه	277
الخوارزميات	∞	صورة	216
مصفوفة	∞	طبقة	206
الخبيرة	1626	الأنماط	206

جدول ٥ : الكلمات العشرين ذات معامل الغرابة الأعلى في مدونة الفيزياء

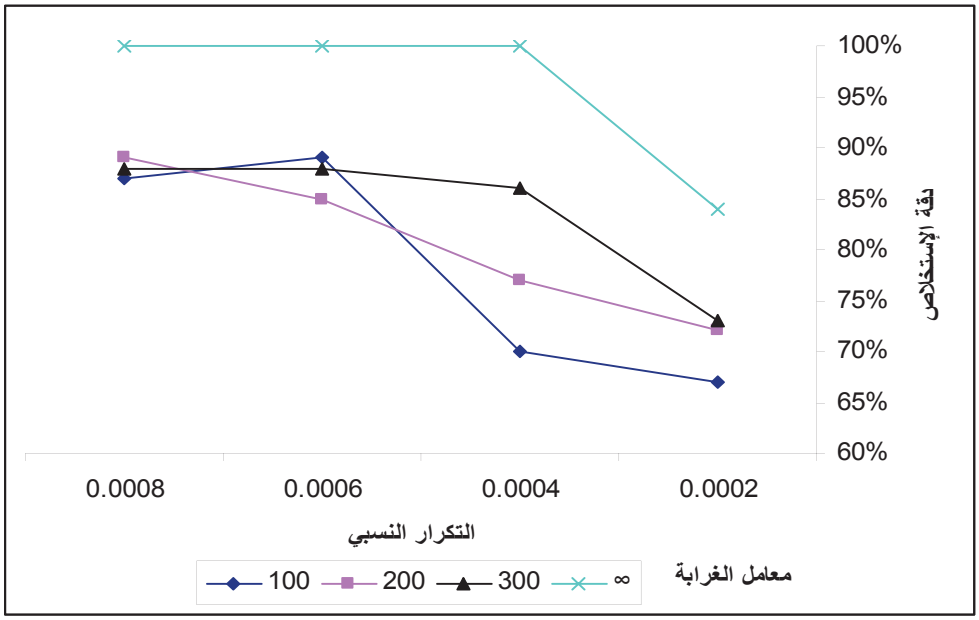
مدونة الفيزياء			
الكلمة	معامل الغرابة	الكلمة	معامل الغرابة
الكمومية	∞	بوهر	∞
الإلكترون	∞	ديراك	∞
جسيمات	∞	اينشتاين	∞
الموجية	∞	الزاوي	∞
ميكانيكا	∞	الكواركات	∞
الفوتونات	∞	شرودنغر	∞
الذرات	∞	المدار	∞
الأوتار	∞	الارتياب	∞
الزمكان	∞	الجزيئات	∞
بلانك	∞	الحيود	∞

وبدراسة الكلمات التي تم استخلاصها كمصطلحات علمية لوحظ التالي:
- بعض الكلمات التي تدل على نفس المعنى وردت بصيغ مختلفة كصيغة المفرد والجمع والمعرفة والنكرة مثل (العصبونية ، العصبونات، العصبون، عصبون، عصبونية، عصبونات) في مدونة الذكاء الصناعي ومثل (الكمومية والتكميم) و (الكوارك و الكواركات) في مدونة الفيزياء وهذا بلا شك يعود لخاصية الإنتاج الصرفي للكلمات الأكثر تكرارا في المدونات المتخصصة .

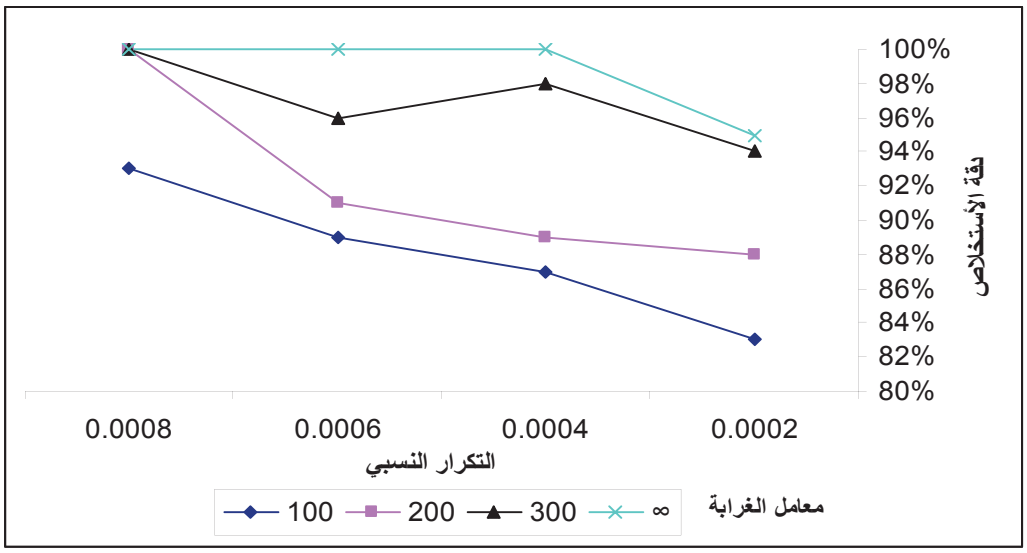
– ورود بعض الكلمات التي لها علاقة بالتخصص ولكن من المفترض أن يكون تكرارها النسبي عاليا في المجالات العامة مثل (اذرع، صورة، الحبال) في مدونة الذكاء الصناعي ومثل (قياسات ، أثير، الأطوال) في مدونة الفيزياء ويعزى ظهور مثل هذه الكلمات إلى ثلاثة أسباب رئيسيه. السبب الأول هو أن حجم مدونة المجالات العامة لم يكن بالقدر الكافي. والسبب الثاني أن تكون الكلمة قد ظهرت بصيغ أخرى في مدونة المجالات العامة.والسبب الثالث يعود إلى إهمال استخدام الهمزة أو التاء المربوطة في بعض نصوص المدونة.

– ورود بعض أسماء الأعلام الذين كان لهم انجازات في مجالات الذكاء الصناعي و الفيزياء أو ممن ارتبط اسمه بطريقه أو نظرية في المجال مثل (تورنج و راذرفورد و بور).

وقد تم اختيار قيم مختلفة للحد الأدنى من التكرار النسبي للكلمات و القيمة الدنيا لمعامل الغرابة والتحقق من دقة النتائج المستخلصة بواسطة الباحث فيما يختص بمدونة الذكاء الصناعي بحكم معرفته بالمجال وبمختص أخر في مجال الفيزياء. العرض البياني (١) و العرض البياني (٢) يوضحان دقة استخلاص المصطلحات لقيم مختلفة من التكرار النسبي ومعامل الغرابة لمدونة الذكاء الصناعي ومدونة الفيزياء على التوالي.



شكل ١ : دقة استخلاص المصطلحات لقيم مختلفة من التكرار النسبي ومعامل الغرابة لمدونة الذكاء الصناعي



شكل ٢ : دقة استخلاص المصطلحات لقيم مختلفة من التكرار النسبي ومعامل الغرابة

وبدراسة النتائج الموضحة في العرضين ١ و٢ يمكن استنتاج التالي

– إن دقة الاستخلاص تتناسب طرديا مع معامل الغرابة ومع التكرار النسبي ولكن ازدياد قيم معامل الغرابة والتكرار النسبي يؤدي إلى قلة الكلمات التي ممن الممكن تحديدها كمصطلحات.

– إن حجم المدونة يؤثر على دقة الاستخلاص فكلما ازداد الحجم ازدادت الدقة. فالنتائج الخاصة بمدونة الفيزياء كانت أكثر دقه في عمومها من مدونة الذكاء الصناعي.

– إن حجم مدونة المجالات العامة ومدى عموميته لهذه المجالات يؤثر على قيمة معامل الغرابة ، فقد كانت أكثر الكلمات التي لم تصنف كمصطلحات كانت بسبب عدم ورودها بصورة كافيه في مدونة المجالات العامة أو بسبب إهمال الهمزة أو التاء المربوطة.

الخاتمة

تطرقت هذه الورقة إلى توصيف طريقة إحصائية بسيطة لاستخلاص المصطلحات المتخصصة ذات المفردة الواحدة اعتمادا على مقارنة ذخائر نصوص المجالات العامة بذخائر نصوص المجالات المتخصصة. وقد أظهرت الطريقة نتائج مشجعة. وكختام لهذا البحث سوف نوجز الخطوات اللازمة لإنشاء نظام حاسوبي بسيط لاستخلاص المصطلحات المتخصصة.

١. بناء مدونة للمجالات العامة ومدونة للمجالات المتخصصة مع ملاحظة انه كلما ازداد حجم المدونتين كانت النتائج أفضل.

٢. معالجة نصوص المدونات مثل إزالة الكشيدة وفصل واو العطف عن الكلمات و إزالة اللواصق مثل حرف الجر اللام والفاء وحرف التشبيه الكاف غير ذلك.

٣. استخلاص المفردات التي تنتمي للمجموعة المفتوحة من الكلمات وتكرارها النسبي في كلتا المدونتين.

٤. حساب قيمة معامل الغرابة لكل مفرده في الخطوة رقم ٢.

٥. اختيار القيم الدنيا للتكرار النسبي ولمعامل الغرابة التي سيتم الاستخلاص بناء عليها.

٦. المفردات التي تكرارها النسبي و معامل الغرابة لها اكبر أو يساوي من القيم المختارة في الخطوة ٥ يتم تحديدها كمصطلحات متخصصة ويتم التعامل معها بعد ذلك حسب الغرض من الاستخلاص.

٧. وتزداد قيمة هذا العمل إذا استمر تزويد المدونات بنصوص جديدة في كلتا المدونتين بحيث يكون بالإمكان متابعة التطورات الحاصلة في المجال وكذلك دقة استخلاص المصطلحات.

وكمتابعة لهذا البحث في المستقبل بحول الله فإنه يمكن استخدام المصطلحات ذات المفردة الواحدة التي تم استخلاصها في عمليات أخرى لاستخلاص المصطلحات المكونة من كلمتين أو أكثر باستخدام طرق عده منها معامل الترابط أو بفحص الكلمات لمحيطة بها (Sinclair، ١٩٩١).

الحواشي والتعليقات

لا يسعى هذا البحث إلى إنشاء قاموس أو قاعدة بيانات للمصطلحات العلمية المتخصصة مثل البنك الآلي السعودي للمصطلحات (باسم) بل يوضح إمكانية استخدم المقارنة بين ذخائر النصوص كأداة يمكن استخدامها لعدة أغراض من ضمنها عرض المفردات التي يمكن أن تكون مصطلحات علميه يمكن أضافتها إلى المعاجم المتخصصة بعد إجراء العمليات اللازمة لذلك من قبل المختصين أو الأنظمة الحاسوبية المساعدة في مجال إنشاء القواميس المتخصصة.

الأستاذ الدكتور فائز بن حماد الغريبي وكيل كلية العلوم التطبيقية وأستاذ الفيزياء بجامعة أم القرى بمكة المكرمة.

المراجع

Ahmad. K. & A. Davies & H. Fulford & M. Rogers. 1994. What is a term? The semi-automatic extraction of terms from text. in: Snell-Hornby. M. & F. Pöchhacker & K. Kaindl (eds.). Translation Studies: An Interdiscipline. Amsterdam: John Benjamins Publishing Company.

Ahmad. K. and Al-Thubaity. A. 2003. Can Text Analyses Tell us Something about Technology Progress?. ACL 2003 Workshop on Patent Corpus Processing. 12July 2003. Sapporo. Japan.

Ahmad. K. and Gillam. L. 2001. Scientific Texts and the Evolution of Knowledge - 'Tunnelling Electrons and Resonating Circuits'. The 13th European Symposium on Language for Special Purposes. Vaasa. Finland. 20-24 August.2001.

Ahmad، K.، Vrusias، B.، and Ledford، Anthony. 2001. Choosing Feature Sets for Training and Testing Self-Organising Maps: A Case Study. Neural Computing & Applications. Volume 10، pp 56-66.

Al-Thubaity، A O. 2004. Knowledge evolution and consolidation in specialist domains : a corpus-based approach. - University of Surrey، 2004. (Ph.D.)

Amigó، E.، Gonzalo، J.، Peinado، V.، Peñas، A.، and verdejo، F. 2004. Using Syntactic Information to Extract Relevant Terms for Multi-document Summarization. Proceedings of the 36th Annual Conference on Computational Linguistics.

Atkins، S.، Clear، J.، and Ostler، N. 1992. Corpus Design Criteria. Literary and Linguistic Computing، 7(1)، 1-16.

Bauer، L. 2001. Morphological productivity. Cambridge Studies in Linguistics، 95. Cambridge: Cambridge University Press.

Biber، D.، Conrad، S.، and Reppen، R. 2002. Corpus Linguistics: investigating language structure and use. Cambridge: Cambridge University Press.

Frantzi، K. ،Ananiado، S. and Mima H.، 2000. Automatic recognition of multi-word terms: the cvalue /nc-value method. International Journal on Digital Libraries، 3:115--130.

Frantzi، K.T. and Ananiadou، S. 1999. The C-Value/NC-Value domain independent method for multi-word term extraction. Natural Language Processing، 6(3):145-179

Khoja، S. 2003 APT: An Automatic Arabic Part-of-Speech Tagger. Ph.D. thesis، Lancaster University.

Pante، P and Lin، D. 2001. AStatistical Corpus-Based Term Extractor. In:Stroulia، E. and Matwin، S. (Eds.) AI 2001، Lecture Notes in Artificial Intelligence. pp. 36-46. Springer-Verlag.

Quirk، R.، Greenbaum، S.، Leech، G. and Svartvik، J. 1985. A Comprehensive Grammar of the English Language. London and New York: Longman.

Sinclair، J. 1991. Corpus، concordance، collocation. Oxford University Press.

Sinclair، J. 1995. Corpus typology – a framework for classification. In Melchers، G. and Warren، B. (eds.) Studies in Anglistics. Almqvist & Wiksell، Stockholm، pp. 17 – 33.