© *Communications of the Arab Computer Society, Vol. 4 No.2, December, 2011*
مجلة إتصالات الجمعية العربية للحاسبات، العدد الثاني من المجلد الرابع، ديسمبر 2011

نظم إسترجاع المعلومات الخاصة باللغة العربية والأمازيغية

أطلس الله فدوى
مركز الدراسات المعلومياتية وأنظمة الإعلام واالتواصل، المعهد الملكي للثقافة الأمازيغية،
شارس علال الفاسي، مدينة العرفان، الرباط-المعاهد، ص.ب. 2055، حي الرياض، الرباط، المملكة المغربية

البريد الالكتروني: ataaallah@ircam.ma

الخلاصة. لمواجهة ما يسمى بالاغراق المعلوماتي استخدمت أنظمة استرجلس المعلومات الاتوماتيكية ليتم من خلالها سد حاجيات العقل البشري عبر تنظيم وتصنيف المعرفة واسترجاعها بدقة متناهية في وقت قياسي عند الحاجة. ولزّيادة في تحسين جودة نظم اللغة العربية ندرج في هذا المقال مجموعة من الدراسات التي تخص بعض وسائل المعالجة الآلية الخاصة بهذه اللغة والمستعملة في نظم استرجلس المعلومات المكتوبة بها إضافة إلى مجموعة من المقترحات الجديدة والمتمثلة أولا في استعمال نموذج التحليل الدلالي الكامن الذي يعد من النماذج الجبرية المتقدمة التي تسمح باستنباط وإقامة روابط وعلاقات دلالية بين مجموعة المصطلاحات المستعملة للتكشيف، ثم ثانيا في إدراج بعض عمليات الترجيح والتي أثبتت جدواها خاصة منها أكابي. أما فيما يخص اللغة الأمازيغية فنظرا لافتقار هذه اللغة لنظام استرجلس قادر على تحليل النصوص المكتوبة بها خاصة تلك المتوفرة بخط تيفناغ فإننا نوضح في هذا المقال التصميم المقترح الخاص بهذه اللغة كما نوضح الأشواط التي قطعت في سبيل هذه الدراسة ونبرز الأشواط الأخرى التي هي بصدد الإنجاز.

الكلمات الجوهرية: إسترجلس المعلومات، معالجة اللغات الطبيعية، ذخائر النصوص، التقييم، اللغة العربية، اللغة الأمازيغية.

1 مقدمة

بعد الثورة الصناعية التي تخللت القرن التاسع عشر، تعيش البشرية حاليا ثورة جديدة عرفت بالثورة المعلوماتية أو ثورة المعلومات الناتجة عن ضخامة حجم المعلومات الإلكترونية المتوفرة وتنوس أشكالها وغنى مضمونها. وعلى غرار الثورات الأخرى فإن ثورة المعلومات صاحبتها أزمة تعي بالخصوص فئة العلماء والباحثين الذين تستدعي طبيعة عملهم البحث الدائم عن المكنوز العلمي الحديث منه والقديم، وتتمثل مظاهر هذه الأزمة في إيجاد المعلومات المطلوبة بدقة بين كم هائل من المعلومات المتوفرة في وقت معقول يتناسب مع الفترة المحددة للعمل المراد إنجازه.

من بين الحلول التي اقترحت لمواجهة هذه الأزمة نجد نظم إسترجلس المعلومات والتي تعتمد على علم البحث عن الوثائق وعن المعلومات داخل الوثائق وعن البيانات الوصفية المتعلقة بالوثائق بالإضافة إلى البحث في قواعد البيانات وشبكة الانترنت، والذي يقوم على عدة علوم من أهمها علوم الحاسوب والرياضيات وعلم المكتبات وعلم المعلومات واللغويات وعلم الاحصاء وعلم النفس الادراكي وعلوم أخرى.

فإذا كان قد اقتصر استعمال هذه النظم أول ما ظهرت بداية الخمسينات من القرن العشرين على المكتبات العامة ومكتبات المؤسسات الأكاديمية خاصة [1]، فإن هذه النظم عرفت تطورا سريعا وفعالا خصوصا خلال العشرينية الأخيرة بعد انتشارالانترنت وظهور مايعرف بالاغراق المعلوماتي. فانطلاقا من نظام أوتوماتيكي صمم خصيصا لمساعدة أمناء المكاتب في العثور على وثائق المكتبات في قواعد البيانات الببليوغرافية أصبحت النظم الحالية أكثر شعبية لا سيما من خلال محركات البحث وأكثر دقة وقوة بحيث لايستغنى عنها.

وبهذا يكون علم استرجلس المعلومات قد تطور من علم اقتصر فيه على البحث عن عدد محدود من الوثائق إلى علم يعتمد على مجموعة من المهام المتنوعة تسمح بالتخزين والتحليل والبحث واسترجلس كميات أوسع وأكبر من المعلومات بدقه متناهية في وقت قياسي على اختلاف لغاتها والأنولس المتوفرة منها عبر وسائل الإعلام من نص وصوت وصورة وفيديو. مما جعل هذا العلم يحضى باهتمام كبير من طرف مستخدمي متصفحي الويب خاصة، كما أدلت بذلك الإحصائيات واستطلاعات الرأي التي أجريت [2] [3]. فإلى جانب هذه الفائدة التي عمت المستفيدين من نظم استرجلس المعلومات والتي تمثلت في التوصل والإستفادة الواسعة والسريعة بالمعلومات، فإن هناك فائدة أخرى تهم بالخصوص اللغات الفقيرة معلوماتيا والتي تشمل المساهمة في إحيائها والنهوض بها وذلك عبر نشر ثقافاتها ومنتوجاتها العلمية والفنية ودعم تطويرها بجلب اهتمام المختصين والباحثين وتوسيع شبكة الدارسين لها.

إلا أنه رغم كل هذا التطور والصيت الذي ناله هذا العلم فإن فئة كبيرة من الباحثين لزالوا يهتمون وينجدبون بالعمل والبحث في هذا الميدان من أجل تحسين جودة نظم إسترجلس المعلومات والتقليص من مشاكل الغموض الذي قد تسببه الخصائص اللغوية الناتجة عن الألفاظ المشتركة أوالناتجة عن التركيبات المختلفة المستعملة داخل النصوص المسترجعة من جهة والمستعملة داخل طلب المستفيد من جهة أخرى والتي قد يكون لها نفس المفهوم الدلالي نظرا لاستعمال كلمات مترادفة أوشاملة أومشتقة أو متصرفة نحويا [4]. و بالفعل فإنه ضمن هذا الإطار تتجلى مساهمتي، حيث سأحاول من خلال هذا المقال توضيح بعض الدراسات التي أنجزت أو التي هي بصدد الإنجاز من أجل توفير أوتحسين جودة نظم استرجلس المعلومات المكتوبة باللغة العربية من جهة والمكتوبة باللغة الأمازيغية من جهة أخرى خاصة تلك المتوفرة بخط تيفناغ. حيث أنه قد تم تطوير نظام خاص باللغة العربية عبر دمج تقنية التحليل الدلالي الكامن التي تعتمد على تحليل العلاقات بين مجموعة من المستندات وبين الكلمات التي تحتوي عليها من خلال بناء "مفاهيم" تتعلق بالمستندات والمصطلحات، إضافة إلى استعمال عدد من عمليات الترجيح التي تسمح بتقييم أهمية الكلمات والمصطلحات المكونة للمستندات مقارنة مع طلبات المستفيدين. كما تم إنشاء نظام جديد خاص باللغة الأمازيغية الذي هو في طور التطوير من خلال استعمال بعض تقنيات المعالجة الآلية للّغة الأمازيغية.

2 خصائص اللغتان المدروستان

1.2 خصائص اللغة العربية

إن الخصائص اللغوية خاصة منها النحوية والصرفية جعلت من العربية لغة صعبة التحليل والمعالجة آليا [5] [6]، وذلك راجع لعدة أسباب نذكر على رأسها تغير شكل الحرف حسب تموضعه بأول أو وسط أوآخر الكلمة من جهة و كتابته منفصلا من جهة أخرى كما يوضحه الشكل 1، تغير معنى الكلمة حسب تغير الحركات التي تصاحب الحروف المكونة لها -الشكل 2-، ولكونها لغة لاصقة إذ يمكن لكلمة واحدة أن تشكل جملة تامة وكاملة المفهوم والمعنى - شكل 3-، إضافة إلى هذا فإن العربية تعد من اللغات الغنية فرغم أن مفرداتها تنقسم إلى ثلاثة أصناف فقط وهي الإسم والفعل والحرف إلا أننا نجد أن عددا كبيرا من الأسماء والأفعال قد يكون مشتقا من نفس الجدر الذي غالبا مايكون ثلاثيا وذلك بالإعتماد على أكثر من 150 وزنا مما يزيدها تعقيدا وصعوبة.

بداية الكلمة	وسط الكلمة	آخر الكلمة	منفصلة
غ	**ـغـ**	**ـغ**	**غ**

الشّكل1. مثال لتغيير شكل الحرف "غين"

الكلمة	المعنى الأول	المعنى الثاني	المعنى الثالث
كتب	كُتُبٌ	كُتِبَ	كَتَبَ
مدرسة	مُدَرَّسَةٌ	مُدَرِّسَةٌ	مَدْرَسَةٌ

الشّكل2. صعوبة الإستدلال الدلالي الذي قد يشكله غياب حركات الشكل

ف	**ي**	تذكر	ون	**ها**
لاصق أمامي	لاصق سابق	الأصل	لاصق لاحق	لاصق خلفي

يتكون اللاصق الأمامي غالبا من حروف العطف والجر، بينما يتكون اللاصق السابق و اللاحق من المزيدات الصرفية لتوضيح زمن الفعل أو حالة الإسم إضافة إلى نوس الجنس وعدده، أما اللاصق الخلفي فغالبا ما يكون ضميرا.

الشّكل3. البنية التركيبية لكلمة في اللغة العربية

2.2 خصائص اللغة الأمازيغية

اللغة الأمازيغية تعرف باسم البربرية وهي من أقدم لغات شمال إفريقيا، تمتد من البحر الأحمر شرقا إلى جزر الكناري غربا، ومن النيجر جنوبا إلى البحر الأبيض المتوسط شمالا. وتنقسم هذه اللغة، بسبب العوامل التاريخية والجغرافية والاجتماعية واللغوية، بالمغرب خصوصا إلى ثلاثة مناطق كبرى: منطقة تارفيت في الشمال، منطقة تامزيغت وسط المغرب وفي جنوبه الشرقي، ومنطقة تشلحيت في الجنوب الغربي وجبال الأطلس الكبير. فبالرغم من أن 50 ٪ من سكانة المغرب ناطقة باللغة الأمازيغية إلا أن هذه اللغة كانت ولمدة طويلة ينحصر تداولها في نطاقات عائلية وغير رسمية [8]. ولكن في العقد الماضي وبفضل الإلتفاتة الملكية أصبحت الأمازيغية لغة مؤسساتية تم دمجها في النظام التعليمي المغربي منذ بداية 2003.

تعد هذه اللغة فرعا من المجموعة الإفريقية الأسيوية (الحامية السامية) [9] [10]. احتفظت من قدم السينين بخطها الأصلي رغم أنها انتشرت وكتبت لفترات طويلة بالخط العربي واللاتيني. وللمحافظة على هذا الخط فقد قام باحثوا المعهد الملكي للثقافة الأمازيغية بتكييفه و تنميطه ومعيرته حيث يتكون هذا الخط والذي يحمل اسم تيفناغ - يركام 33 حرف، 27 منها صامت، حرفان شبها صامتين، وأربع صوائت، كما يوضحه الشكل 4، ويكتب هذا الخط من اليسار إلى اليمين.

الصوامت	ⵀ, ⵃ, ⵄ, ⵅ, ⵇ, ⵉ, ⵊ, ⵍ, ⵎ, ⵏ, ⵔ, ⵙ, ⵚ, ⵛ, ⵜ, ⵟ, ⵡ, ⵢ, ⵣ, ⵥ, ⵧ, ⵨
أشباه صوامت	ⵗ, ⵘ
الصوائت	ⴰ, ⴱ, ⴲ, ⴳ

الشّكل4. حروف تيفناغ- يركام

إلى جانب اللغة العربية فإن الغة الأمازيغية هي الأخرى تعد من اللغات الصعبة بالنسبة للمعالجة الآلية لكونها لغة اشتقاقية، حيث تنقسم مفرداتها إلى ثلاثة أصناف: الإسم والفعل والإشاريات. يتميز الإسم في اللغة الأمازيغية بجنسه المؤنث والمذكر وبعدده المفرد والجمع إضافة إلى حالتي الإرسال والإلحاق. و يكون الفعل في اللغة الأمازيغية في صورة بسيطة أو في صورة مشتقة، ويتصرف الفعل سواء في صورته البسيطة أو المشتقة إلى أربعة صيغ وهي: المجرد والغيرالتام والتام المثبت والتام المنفي، كما تلتصق به نفس العلامات الفعلية. بينما تتميز الإشاريات بنوعين: تلك التي تخصص الاسم ويطلق عليها المخصصات الإشارية وتلك التي تعوضه ويطلق عليها الضمائر الإشارية [11] [12].

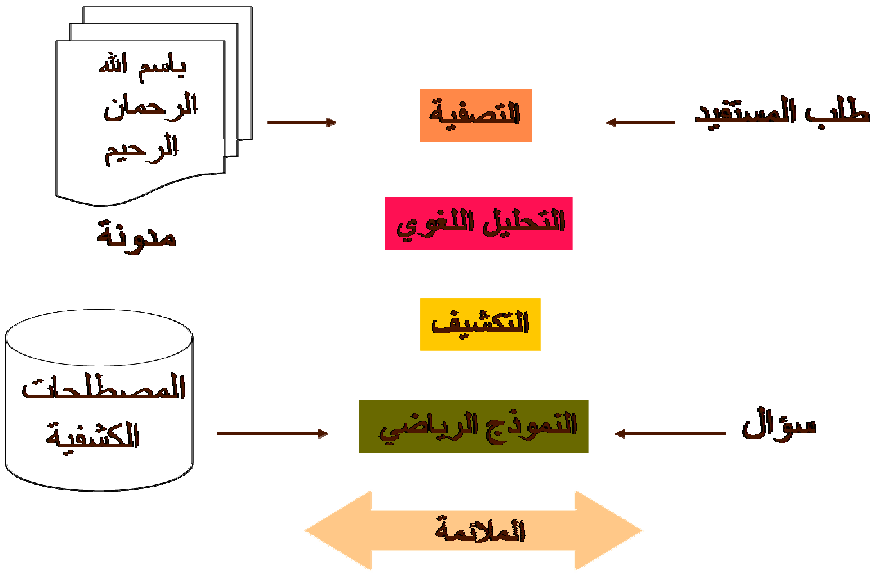
3 تصميم نظم إسترجاع المعلومات

نظرا لأهمية المعلومات في وقتنا المعاصر فان تطوير نظام استرجلس المعلومات وتفعليه وتحسينه ليعد من الظروريات حتى يتسنى لنا تقديم أفضل النتائج الممكنة بدقة متناهية وفي وقت قياسي مهما كانت المعلومات معقدة أو متداخلة.

فالبرغم لما لهذه النظم من تصميم عام ومتداول إلا أن كل مرحلة من مراحل هذه النظم و خاصة منها مرحلتي التصفية والتصنيف تختلف من لغة إلى أخرى وذلك بحسب الخصائص اللغوية وطرق المعالجة الآلية الملائمة لكل لغة.

1.3 تصميم عام

إن نظام إسترجلس المعلومات هو نظام قائم بحد ذاته حيث يعمل على استرجلس كل النتائج والمعلومات المتعلقة ببحث معين أو طلب بعينه وذلك حسب مجموعة من المراحل التي يوضحها الشكل 5.



الشّكل5. تصميم عام لنظم إسترجاع المعلومات

فبعد عملية خزن المعلومات التي تعتمد على جمع واختيار النتاج الفكري الذي يهم المستفيدين أو بعد تقديم المستفيد طلبه إلى نظام الاسترجلس عن طريق صياغة طلبه (سؤاله) بمجموعة من الكلمات تتم مرحلة التصفية والتحليل من أجل مرحلة التكشيف. فبالنسبة للوثائق يتم تحليلها وصفيا وموضوعيا من اجل تحديد موقعها ونقاط إتاحتها والوصول إليها وتحديد الموضوعات والمفاهيم الواردة فيها ثم التعبير عن هذه المفاهيم بمفردات قد تكون مما ورد في الوثيقة نفسها أو مشتقة منها

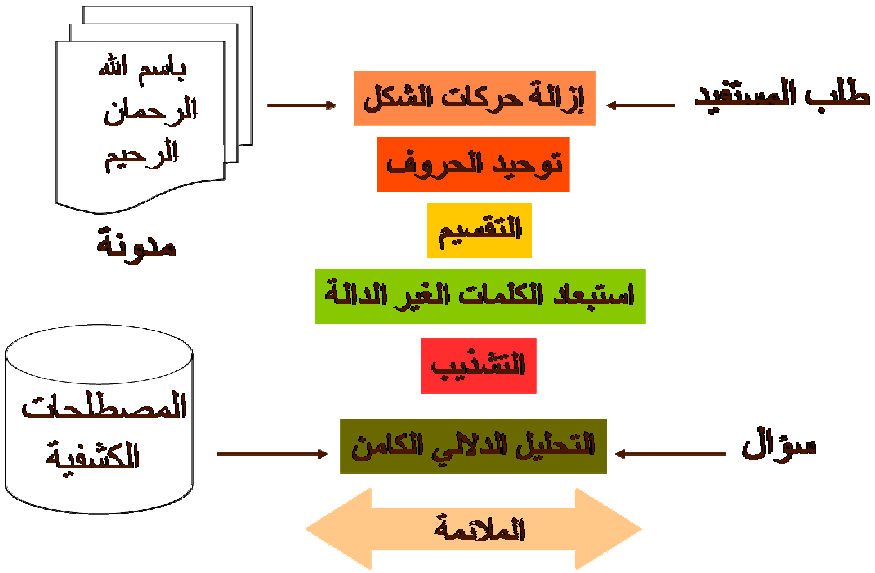
أو أن تكون مفردات مقننة مسبقا لغرض التحكم فيها والسيطرة عليها. أما بالنسبة لطلب المستفيد فبعد التصفية والتحليل تتم عملية الاسقاط على مجموعة المصطلحات الكشفية وذلك حسب النموذج الرياضي المتبع والذي يمكن تصنيفه حسب ثلاثة نماذج: نمودج بولياني يعتمد خصوصا على تطابق المصطلح البحثي للمستفيد مع المصطلحات الكشفية التي اعتمد عليها في إنشاء قاعدة البيانات [13]، نموذج فضاء المتجه (أو نموذج متجه المصطلحات) وهو نموذج جبري لتمثيل المستندات والوثائق النصية كمتجهات معرفة [14]، أونموذج احتمالي يمثل محتويات الوثائق اعتماد على احتمال وأهمية الوثيقة من جهة واحتمال أهمية المصطلحات الكشفية التي تحتوي عليها من جهة أخرى [15]. بعد هذه المرحلة تجرى مرحلة الملائمة حيث يتم قياس نسبة ارتباط طلب المستفيد وتقاربه من الوثائق والمستندات المخزونة وبالتالي يتم ترتيب وعرض النتائج على المستفيد.

2.3 تصميم خاص باللغة العربية

تماشيا مع متطلبات وخصائص اللغة العربية فإن مرحلة التصفية والتحليل تستلزم مرحلة تخص إزالة كل حركات الشكل وأخرى لتوحيد شكل كتابة حرف الألف (أ، آ، إ، ا), والياء (ي، ى) والتاء (ه، ة) قبل تقسيم محتوى النصوص إلى كلمات يتوسطها الفراغ الأبيض. بعد ذلك تتم إزالة الكلمات الغير دالة واستبعادها من عملية التكشيف بما في ذلك أسماء الإشارة وحروف الجر وما شابه ذلك. ثم تجري مرحلة التحليل التي يستنبط من خلالها جذور الكلمات دون اعتبار اشتقاقاتها المختلفة (عملية التجدير)، أو استنباط أصل الكلمات حيث لا يحتاج هذا الأصل أن يكون مكافئ للجذر الصرفي للكلمة بل يكون كافيا لأن يجعل الكلمات ذات الصلة الدلالية لها نفس الاصل حتى لو كان هذا الاصل ليس في حد ذاته جذرا صالح (عملية التشذيب)، وذلك في محاولة للتغلب على تعدد الأشكال النحوية للكلمة الواحدة.

وقد أثبتت الدراسات التي نشرت بمؤتمراسترجلس المعلومات لسنة 2001 (تريك 2001) [16]، إضافة إلى الدراسات التي أجرتها كل من مجموعة الجلالي [6] ومصطفى [17] أن عملية التشذيب أكثر إفادة بالنسبة لنظم إسترجلس المعلومات المكتوبة باللغة العربية بالمقارنة مع العمليات الأخرى للمعالجة الآلية للّغة العربية والتي يعتمد عليها في نظم إسترجلس المعلومات.

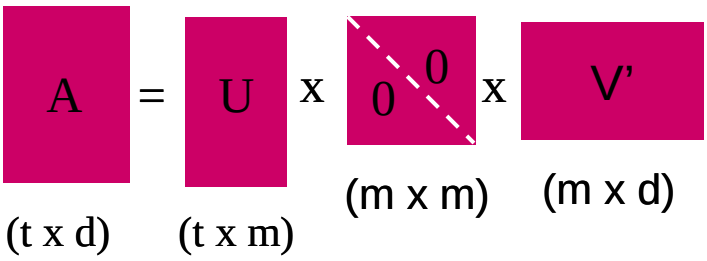
فاعتمادا على هذه الدراسات اقترحنا التصميم المبين على الشكل 6، والذي إضافة إلى هذا اقترحنا فيه استعمال نموذج الفضاء المتجهي (ن.ف.م) المتقدم الذي يحمل إسم التحليل الدلالي الكامن (ت.د.ك) والذي يعتمد على تحليل العلاقات بين مجموعة من الوثائق وبين الكلمات التي تحتوي عليها من خلال بناء مفاهيم تتعلق بالوثائق والمصطلحات [18] [19].



الشّكل6. تصميم نظام لإسترجاع المعلومات خاص باللغة العربية

1.2.3 نموذج التحليل الدلالي الكامن

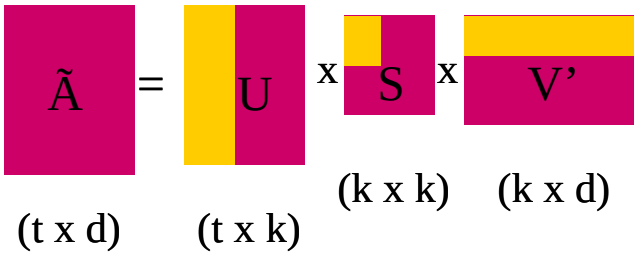
التحليل الدلالي الكامن هو نموذج لتمثيل متعدد الأبعاد لكلمات اللغة، ناتج عن التحليل الإحصائي للوثائق والمستندات المتوفرة باحدى المدونات، حيث يمثل معنى كل كلمة موجهة داخل فضاء متعدد الأبعاد. وبهذا يكون التحليل الإحصائي من شأنه بناء مصفوفة مكونة من عدد ترددات المصطلحات الكشفية التي سيعمل على تقليص عدد أبعادها من خلال تطبيق عملية التفكيك إلى قيم فردية وذلك من أجل إبراز العلاقات الدلالية الموجودة بين المصطلحات الكشفية ووثائق المدونة .[20]



$$A = U \times S \times V'$$

الشّكل7. تمثيل عملية التفكيك للمصفوفة A

يمثل الشّكل7 عملية التفكيك للمصفوفة $A=[a_{ij}]$ المكونة من t كلمة وd وثيقة، حيث تشكل a_{ij} عدد ترددات الكلمة i بالوثيقة j، وتمثل S المصفوفة القطرية التي تحتوي على القيم الفردية، U و V' مصفوفتان متعامدتان، أما m فهو رتبة المصفوفة A .



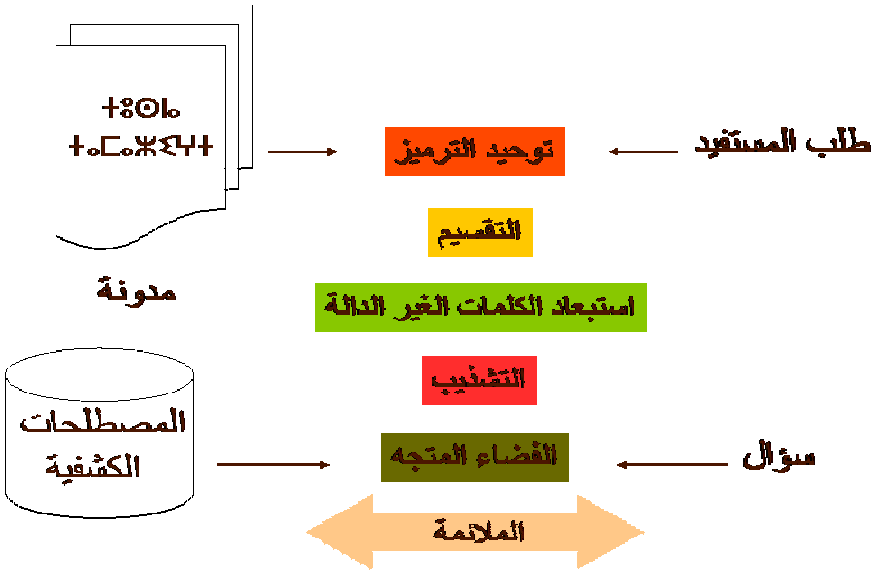
$$\tilde{A} = U_k \times S_k \times V_k'$$

الشّكل8. تقليص عدد أبعاد المصفوفة من خلال تطبيق عملية التفكيك

أما الشّكل 8 فهو يمثل المصفوفة الجديدة الناتجة عن التفكيك والتي تعبر كل متجهة فيها عن معنى دلالي موسع يتعدى تمثيل عدد ترددات الكلمات، حيث تمثل k عدد القيم الفردية التي اعتمد عليها في التحصيل على المصفوفة الجديدة.

3.3 تصميم خاص باللغة الأمازيغية

على خلاف اللغة العربية فإن المعالجة المسبقة للّغة الأمازيغية تستلزم مرحلة توحيد الترميز وذلك بسبب تعدد أنولى الترميز والحروف المستعملة لكتابة اللغة الأمازيغية، بعد ذلك نقترح إستعمال مرحلة إزالة الكلمات الغير الدالة فمرحلة التشذيب ثم التكشيف فالملائمة حسب نموذج الفضاء المتجه [21] [22]، كما يفصله التصميم الممثل على الشكل 9. إلا أن هذا التصميم ليس نهائيا، إذ أن هذه الدراسة لا زالت في طور الإنجاز كما ستوضحه الفقرة 2.4.

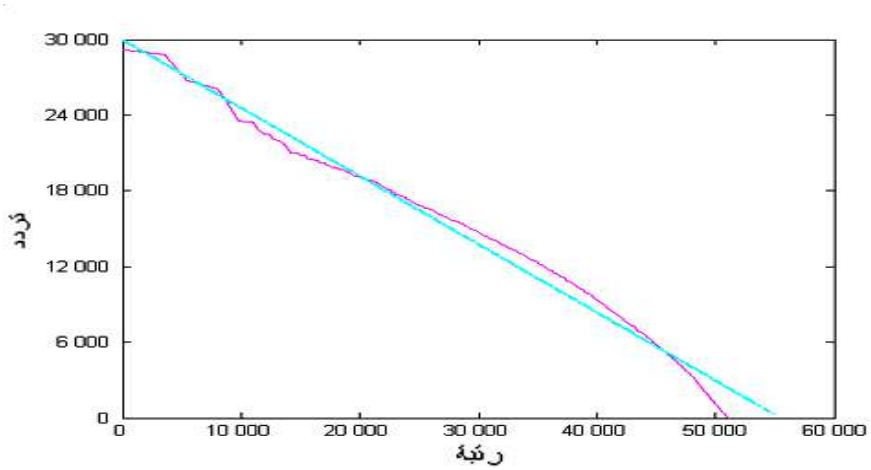


الشّكل9. تصميم نظام لإسترجاع المعلومات خاص باللغة الأمازيغية

4 الدراسات المقترحة، تحليلها ونتائجها

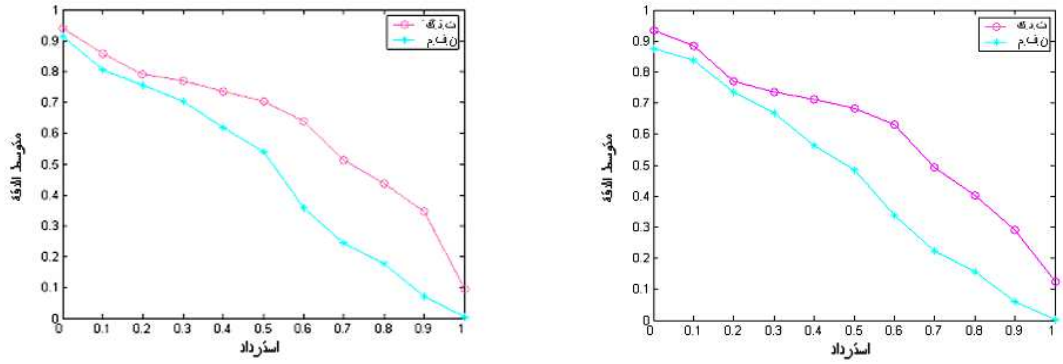
1.4 بالنسبة للّغة العربية

لدراسة جدوى المقترحات الجديدة كان من الضروري الحصول على مدونة إختبار مكونة من مجموعة من الوثائق وبالفعل فقد عملنا على جمع عدد من المستندات وصفحات الويب الخاصة بعالم البيئة، وكل ما يتعلق بمشاكل التصحر والضوضاء والكوارث الطبيعية والتلوث بجميع أنواعه سواء الهوائي أو المائي أو الترابي، ثم تنظيمها وترتيبها بحيث أصبحنا نتوفر على مدونة مكونة من 1060 وثيقة، و30 طلب مصاغ بشكل مختصر وآخر موسع. ويصل عدد مفردات هذه المدونة إلى 475148 كلمة منها 54705 كلمة مختلفة. وحسب قانون زيبف، الذي يهتم بتمثيل عدد ترددات المصطلحات المكونة للمدونة ويبرز أهميتها بقدر تقارب منحناها مع المنحنى المنحدر ذو الرتبة 1- والمرسوم باللون الأزرق على الشكل 10، وبعض المعايير الأخرى المشار إليها في [18] و[23] يتبين أن هذه المدونة تعتبر غنية وجديرة بأن تستعمل في اختبار جودة التصميم المقترح الجديد.



الشّكل10. تمثيل المدونة حسب قانون زيبف

من بين التجارب الاولى التي أجريناها، تجربة مقارنة النموذج التحليل الدلالي الكامن بنموذج الفضاء المتجه فتبين، كما يوضحه الشكل 11، أن المنحى ذي اللون الوردي والممثل لنموذج التحليل الدلالي الكامن يعطي نتائج جيدة مقارنة مع المنحى الأزرق الممثل لنموذج الفضاء المتجه وذلك بالنسبة لكل من الطلبات الموسعة - رسم أ - و الطلبات المختصرة - رسم ب.

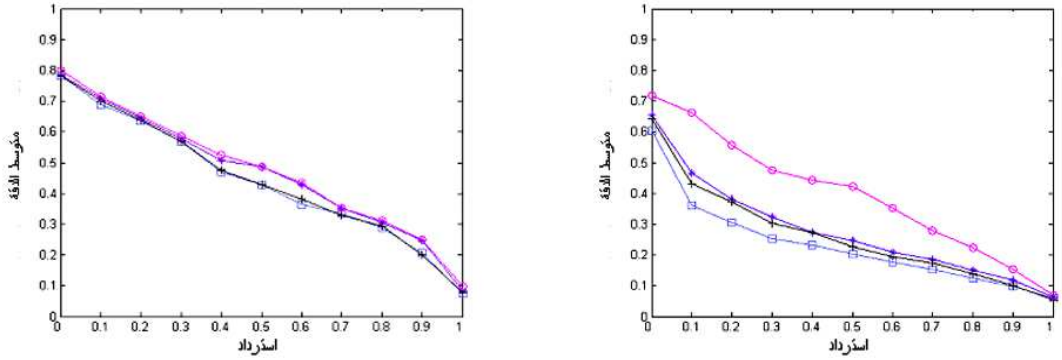


- ب - - أ -

الشّكل11. مقارنة النموذج التحليل الدلالي الكامن بنموذج الفضاء المتجه

إضافة إلى هذه التجربة فقد قمنا بدراسة تأثير كل من استبعاد الكلمات الغير الدالة - المنحنى الأسود ذو العلامة + - والتشديب - المنحنى البنفسجي ذو العلامة * - من جهة وتأثير استعمالهما معا - المنحنى الوردي ذو العلامة o - بالمقارنة مع استعمال كل الكلمات التي تتوفر عليها المدونة كمصطلحات للتكشيف دون أي معالجة آلية - المنحى الأزرق ذو العلامة ▫ -، فتوضح من خلال

هذه التجربة، كما يبين ذلك الشكل 12، أن إستبعاد الكلمات الغيردالة مؤثرة بالنسبة للطلبات الموسعة - رسم أ- على خلاف حالة الطلبات المختصرة - رسم ب- بينما أن عملية التشديب فهي مهمة و تحسن من جودة النظام في كلتا الحالتين.



- أ - - ب -

الشّكل12. تأثير وسائل المعالجة الآلية على جودة النظام

من جهة أخرى اهتممنا بدراسة أثر عمليات الترجيح على دقة النظام وذلك باستعمال ما يعادل 25 عملية للترجيح، والتي من دورها تقليص أهمية اختلاف طول أو قصر الوثائق المستعملة مقارنة مع عدد تكرر نفس الكلمة في هذه الوثائق، إذ يمكن أن تتكرر الكلمة لعدة مرات في وثيقة طويلة ويقتصر ذكرها على مرة واحدة في وثيقة قصيرة فتعطى الاهمية خلال البحث للوثيقة الطويلة على حساب القصيرة رغم أهمية هذه الأخيرة أحيانا بالنسبة للمستفيد.

و بالفعل فقد أثبتت هذه الدراسات أن هناك بعض العمليات التي تؤثر سلبا على مردودية النظام بينما تؤثر أخرى إيجابا [18] ، ومن بين هذه الأخيرة اخترنا أفضل خمسة عمليات وهي موضحة على الشكل 13 الذي يبرز أن أحسن عملية للترجيح من بين كل العمليات التي درسناها هي عملية أكابي ب م- 25 والتي تكتب معادلتها الرياضية على الشكل التالي:

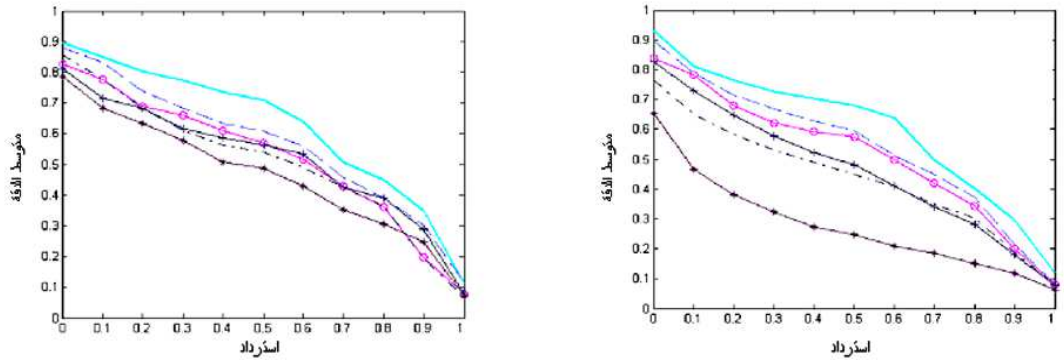
$$\frac{3*\left(\log N-\log df_i\right)*\mathrm{f}_{\mathrm{ij}}}{2*\left(0.25+\left(0.75*\frac{N*dl_j}{\Sigma_{k=1}^{N}dl_k}\right)\right)+\mathrm{f}_{\mathrm{ij}}}$$

حيث أن dl يساوي طول الوثيقة،

f_{ij} يساوي عدد تردد الكلمة i بالوثيقة j،

df_i تمثل عدد الوثائق التي تظهر بها الكلمة i،

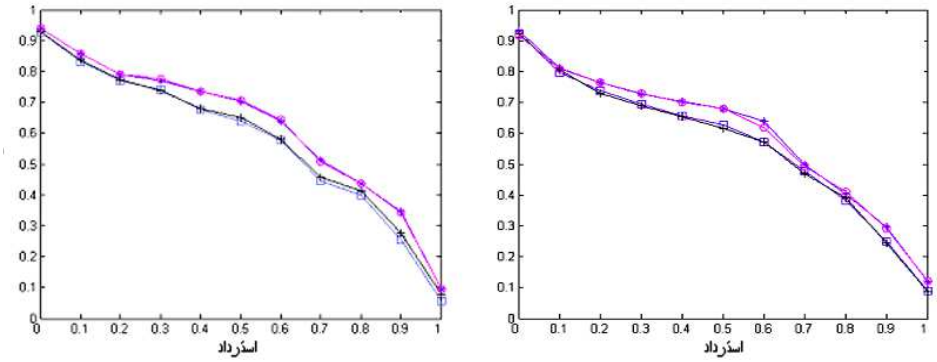
وN تمثل عدد الوثائق المتوفرة بالمدونة.



- أ - - ب -

الشّكل13. تأثير عمليات الترجيح على جودة النظام

وفي النهاية حاولنا دراسة تأثير وسائل المعالجة الآلية وعملية الترجيح أوكابي ب م -25 معا على جودة ودقة النظام فلاحظنا من خلال النتائج الموضحة على الشكل 14 أن عملية استبعاد الكلمات الغير الدالة لم يعد لها أي تأثير حتى في حالة الطلبات الموسعة ـ رسم أ- وذلك راجع لما للعملية أوكابي ب م -25 من تأثيرعلى التقليص من أهمية هذه الكلمات والتي غالبا ما تكون جد منتشرة ومتكررة لعدة مرات، أما فيما يخص عملية التشذيب فهو واضح مدى جدوى هذه العملية من خلال النتائج.



- أ - - ب -

الشّكل14. تأثير كل من وسائل المعالجة الآلية وعمليات الترجيح على جودة النظام

2.4 بالنسبة للّغة الأمازيغية

نظرا لعدم توفر اللغة الامازيغية على نظام خاص بها يأخذ فيه بعين الإعتبار كل خصائصها اللغوية والثقافية فكان من اللازم بداية الأمر أن نحاول تصميم نظام جديد. وبالفعل فقد أنشانا محرك للبحث خاص بهذه اللغة يعتمد على التصميم الموضح في الشكل 9 دون الإستناد إلى عملية التشذيب، وأدمجناه في الموقع الإلكتروني للمعهد الملكي للثقافة الأمازيغية. إلا بما أن محرك البحث اعتمد فيه على تصميم اختصر فيه بالنسبة لوسائل المعالجة الآلية على عملية استبعاد الكلمات الغير الدالة فقط، فإننا لا نزال نتابع هذه الدراسة محاولين تحسين جودة نظام استرجاع المعلومات المكتوبة باللغة الأمازيغية خاصة بخط تيفناغ، و بالتالي تحسين جودة محرك البحث الخاص بهذه اللغة. من أجل هذه الغاية فقد ارتأينا بادئ الأمر أن ندرس جدوى عملية التشذيب معتمدين في ذلك على التجارب والدراسات التي أجريت على اللغة العربية من جهة وعلى التشابه الحاصل بين الخصائص الصرفية لهتين اللغتين من جهة أخرى. فبعد تصميم وإنشاء وسيلة التشذيب [22]، ودمجها في نظام استرجاع المعلومات الأمازيغية، نحن حاليا بصدد جمع وتصنيف مدونة خاصة باللغة الأمازيغية، حتى يتسنى لنا اختبار مفعول هذه العملية على النظام. وقد راعينا في ذلك مجموعة من الأسس والقواعد المتبعة عالميا في إنشاء مثل هذه المدونات خاصة تلك المنصوص بها عند مجموعة تريك [24] [25]، حيث نحاول في مرحلة الجمع أن تكون هذه المدونة شاملة لمجمل المصطلحات والأمثلة اللغوية، غنية بمختلف المواضيع الأدبية، العلمية، الفنية والإخبارية.

بعد ذلك نتابع مرحلة التنقية والتي تعتمد على إزالة حقائب لغة النصوص التشعيبية ثم توحيد الترميز. ثم نأتي إلى مرحلة التصنيف حيث أننا نجمع نصوص كل الوثائق في ملف واحد ونصنفها على النحو الموضح من خلال الشكل 15، بحيث نحدد لكل وثيقة رقمها الدلالي، تاريخ صدورها، نوعيتها، عنوانها ومضمونها.

<doc>
<docid> 10 </docid>
<date>2008 </date>
<fld> story </fld>
<text>
<title> ⴰⵏⵎⵓⵔ ⵏ ⵜⵉⵎⵣⵉ </title>
ⵖⵓⵏ ⵜⵓⵙⵙ, ⴰⵀⵀⵢ ⵀⵛⵛⴰ ⴰⵏⵎⵓⵔ, ⵙⵅ ⵜⴰⵏⵏⴰⵙⵜ ⵏⵙ, ⴰⵏⵏⴰ ⵢⵙ ⵜⴰⵅⵓⵜ ⴰⵏ ⴰⵙⵙⴰⵙ.
ⵖⵉⵀⴰ ⵅ ⵜⴰⵅⵓⵜ ⵀⵛⵛⴰ ⵉⵜⵜⵉ.
ⴰⵏⴰ ⴰⵙ ⵖⴰⵏⵎⵓⵔ: ... .
</text>
</doc>

الشّكل15. مثال لوثيقة مدونة

كما نقترح مجموعة من الطلبات التي تصنف على النحو الموضح بالشكل 16. فنضع لكل طلب رقمه الدلالي، مضمونه بصفة مختصرة، وصف لمحتواه ثم مضمونه الموسع.

<top>
<num> Number: 2 </num>
<title> ⵜⴰⵎⴰⵡⴰⵙⵜ ⵏ ⵜⵎⵣⵓⵖⵜ </title>
<desc> Description :
ⵓⵔⵔ ⵅⵍ ⵔⵎⵙⵎⵓ ⵙⴰⴰⵎⵎⵓⵔⵏ ⵅⵍ ⵜⴰⵎⴰⵡⴰⵙⵜ ⵏ ⵜⵎⵣⵓⵖⵜ
</desc>
<narr> Narrative:
ⵙⵎⵙⵙⵓⵏ ⵙⴰⴰⵎⵎⵓⵔⵏ ⵅⵍ ⵔⵣⴰⴰⴰ ⵉⴰⵓ ⵜⴰⵎⴰⵡⴰⵙⵜ ⵏ ⵜⵎⵣⵓⵖⵜ ⴷ ⵜⴰⵎⴰⵡⴰⵙⵜ ⵏ ⵔⵅⵍⴰ, ⵜⴰⵔⵓⵙ ⵏ ⵜⵎⵣⵓⵖⵜ, ⵜⵔⴷⵓⵜ ⵜⵔⵜⵜⵅⵢ ⴷ ⴰⵅ ⵔⵃⴷⴷⵙⴰ ⵏ ⵜⵎⵣⵓⵖⵜ.
</narr>
</top>

الشّكل16. مثال لطلب مدون

بعد ذلك نقوم بتحديد الوثائق التي تعد كإجابات صحيحة لكل طلب حتى يتسنى لنا بعد ذلك إستعمال هذه المدونة لإختبار جودة النظام المقترح.

5 الخاتمة

إن الدراسة التي أجريت على اللغة العربية أثبتت أن نموذج التحليل الدلالي الكامن قادر على تحسين وتجويد مردودية النظام المقترح من أجل استرجاع المعلومات المكتوبة باللغة العربية وأن استعمال طريقة الترجيح أكابي هي أيضا كان لها دور فعال في زيادة دقة هذا النظام. ومن جهة اخرى فإن الدراسة القائمة على خصائص اللغة الأمازيغية فهي في طور الإنجاز ونتمنى أن نحصل قريبا على نتائج مشجعة لنساهم بدورنا في الإرتقاء بهذه اللغة إلى مستوى العالمية ونساهم في نشر ثقافتها ومنتوجاتها عبر المساهمة في توفير نظم خاصة بها لاسترجاع المعلومات.

6 شكر وتقدير

شكر خاص للزميلة بولقنادل سهام التي ساهمت إلى جانبي في إنجاز عدد كبير من الدراسات والتجارب المقترحة في هذا المقال.

موجز من السيرة الذاتية

فدوى أطاع الله حاصلة على إجازة في الرياضيات وعلى دبلوم الدراسات العليا المعمقة في علوم الحاسب الآلي والاتصالات من كلية العلوم التابعة لجامعة محمد الخامس أكدال بالرباط، المغرب . بالإضافة إلى شهادة الدكتوراه في علوم الهندسة تخصص علوم الحاسب الآلي والاتصالات من نفس الكلية بالتعاون مع جامعة ديربورن، ميشيغان بالولايات المتحدة الأمريكية.

حاليا تعمل كباحثة متخصصة في مجال المعالجة الآلية للّغات الطبيعية بالمعهد الملكي للثقافة الأمازيغية، بمركز الدراسات المعلومياتية وأنظمة الإعلام واالتواصل. وترتكز اهتماماتها البحثية بالأساس في مجال المعالجة الآلية للّغات الطبيعية، استرجاع المعلومات، إنشاء ذخائر النصوص وتصنيفها، وقواعد البيانات. لها عدة مقالات علمية منشورة في مجلات علمية عالمية ومحلية محكمة.

المراجع

[1] C.N. Mooers, “Application of Random Codes to the Gathering of Statistical Information”, Master Thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, 1948.

[2] http://www.nielsen-online.com/pr/pr_040223_us.pdf

[3] http://en-us.nielsen.com/rankings/insights/rankings/internet

[4] F. Moreau, and P. Sébillot, “Contributions des techniques du traitement automatique des langues à la recherche d'information”, Research Repport n° 1690, IRISA, Rennes, France, 2005.

[5] L. S. Larkey, L. Ballesteros, and M. Connell, “Improving Stemming for Arabic Information Retrieval : Light Stemming and Cooccurrence Analysis”, In Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002), Tampere, Finland, August 2002, p. 275-282

[6] M. Aljlayl, and O. Frieder, “On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach”, In 11 the International Conference on Information and Knowledge Management (CIKM), November 2002, Virginia (USA), p. 340-347.

[7] S. Baloul, M.Alissali, M.Baudry, and P.Boula de Mareüil, “Interface syntaxe-prosodie dans un système de synthèse de la parole à partir du texte en arabe”, 24ème Journées d’Étude sur la Parole, June 24-27 2002, Nancy, p. 329-332.

[8] A. Boukous, “ Société, langues et cultures au Maroc : Enjeux symboliques”. Najah El Jadida. Casablanca, Maroc, 1995.

[9] J. Greenberg, “The Languages of Africa”. The Hague, 1966.

[10] O. Ouakrim, “Fonética y fonología del Bereber”. Survey at the University of Autònoma de Barcelona, 1995.

[11] M. Ameur, A. Bouhjar, F. Boukhris, A. Boukous, A. Boumalk, M. Elmedlaoui, E.

M. Iazzi, and H. Souifi. “Initiation à la langue amazighe”. The Royal Institute of Amazigh Culture, 2004.
[12] F. Boukhris, A. Boumalk, E. H. El Moujahid, and H. Souifi. “La nouvelle grammaire de l'Amazighe”. The Royal Institute of Amazigh Culture, 2008.
[13] W. G. Waller, and D. H. Kraft, “A mathematical model of a weighted Boolean retrieval system”, Journal of Information Processing & Management, Vol. 15, No. 5, pp. 235-245, 1979.
[14] G.Salton, A. Wong, and C. S Yang, “A vector space model for automatic indexing”, Commun of the ACM, November 1975, Vol. 18 Issue 11, p. 613-620.
[15] S. E. Robertson, and K. S. Jones, “Relevance weighting of search terms”, Journal of the American Society for Information Science, vol. 27, no 3, p. 129-146, 1976.
[16] F. C. Gey, D. W. Oard, “The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic Using English, French or Arabic Queries”, Proceedings of the 2001 Text Retrieval Conference (TREC-2001), National Institute of Standards and Technology, 13-16 November 2001, p. 16-26.
[17] S. H. Mustafa, and Q. A.Al-Radaideh, “Using N-grams for Arabic text searching”, JASIST, September 2004, Vol 55 n.11, p.1002-1007.
[18] F. Ataa Allah, “Information retrieval: applications to english and arabic documents”, Ph.D. Dissertation, Mohamed V -Agdal University, Rabat, Morocco, 2008.
[19] F. Ataa Allah, S. Boulaknadel, A. El Qadi, and D. Aboutajdine, “Evaluation de l’Analyse Sémantique Latente et du Modèle Vectoriel Standard Appliqués à la Langue Arabe”, Revue of Technique et Science Informatiques, vol. 27, n° 7, p. 851-877, 2008.
[20] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R.Hrashman, “Indexing by latent semantic analysis”, Journal of th american society for information science, Vol 41(6), p. 391-407, 1990.
[21] F. Ataa Allah, and S. Boulaknadel, “Amazigh Search Engine: Tifinaghe Character Based Approach”, Proceeding of the International Conference on Information and Knowledge Engineering (IKE 2010), Las Vegas, Nevada, USA, 12-15 juillet 2010, pp. 255-259.
[22] F. Ataa Allah, and S. Boulaknadel, “Pseudo-racinisation de la langue amazighe”, Proceeding of the Traitement Automatiques des Langues Naturelles (TALN 2010), Montréal, Canada, 19-23 juillet 2010.
[23] S. Boulaknadel, Recherche d'Iinformation en Langue Arabe, Ph.D. Dissertation, Mohamed V -Agdal University, Rabat, Morocco, 2008.
[24] N.Craswell, D.Hawking, R.Wilkinson, and M.Wu, “Overview of the TREC 2004 web track”, TREC 2004 Proceeding (13rd Conférence), 2005.
[25] Voorhees E.M. (2005). “Overview of TREC 2004”, TREC 2004 Proceeding (13rd Conférence), 2005.

الملخص باللغة الانجليزية

Title: Information retrieval systems for Arabic and Amazigh languages

Abstract. To face what is called the Information dump, information retrieval systems have been used, in order to help people to organize and classify knowledge, and to retrieve pertinent information rapidly when it is needed. In this aim, we have tried to improve the performance of the existing Arabic information retrieval systems by using, in addition to some natural language processing, the latent semantic analysis method. This later consists in analyzing relationships between the terms that are contained in a

set of documents and producing concepts related to the documents and terms. Moreover, we have included to this system and tried some weighing schemes that have proved their performance, specially the Okapi BM-25 scheme. Regarding the Amazigh language, in the aim to compensate the lack of an Amazigh information retrieval system, especially for the tifinagh script, we have proposed a new information retrieval system structure that is in the process of evaluation.

Keywords: Information retrieval, Natural language processing, Corpora, Evaluation, Arabic language, Amazigh language.

المصطلحات

التشذيب	Stemming
التشكل اللغوي	Linguistic morphology
محركات البحث	Search Engines
ذخائر النصوص	Corpora
تكشيف الكلمات	Word indexing
قائمة الاستبعاد	Stop list
بيانات وصفية	Metadata
استرجاع المعلومات النصية	Text information retrieval
اسم شامل	Hyperonymy
لفظ مشترك	Polysemy
لغة فقيرة معلومياتيا	Less-resourced language
متصفح ويب	Web browser
نموذج فضاء المتجه	Vector space model (VSM)
تحليل دلالي كامن	Latent semantic analysis (LSA)
ترجيح	Weighting
لغة لاصقة	Agglutinative language
أصل الكلمات	Stem
التفكيك إلى قيم فردية	Singular value decomposition
أوكابي ب م -25	Okapi BM-25
قانون زيبف	Zipf’s law
لغة النصوص التشعبية	Hypertext markup language