نحو مقاربة شاملة لتحليل وتمثيل الوثائق العربية في الويب الاجتماعي الدلالي

إبراهيم بونحاس، يحيى سليماني
قسم الإعلامية، كلية العلوم بتونس،
جامعة تونس المنار، 1060، تونس

الخلاصة. إن التحول من الويب الحالي إلى الويب الاجتماعي الدلالي يتطلب تطوير منهجيات ومقاربات وبرمجيات تُمَكن من تمثيل الوثائق واستخراج المعارف التي تحتويها. بالنظر إلى خصائص اللغة العربية والأبحاث المنجزة في هذا المجال، يمكننا اكتشاف التحديات التي يواجهها أي مشروع يهدف إلى إدماج الوثائق العربية في الويب الاجتماعي الدلالي. نُقَدم في هذه الورقة نموذجا عاليا لتمثيل الوثائق العربية في شكل خريطة اجتماعية دلالية. كما نقترح بنية تُمَكّن من تحليل الوثائق شبه المنظمة لتنظيمها حسب هذا النموذج. إن البرمجيات المقترحة تُمكن من دراسة اعتمادية المعلومة التي ننظر إليها كمحور اجتماعي باعتباره يؤسس للثقة بين الجهات الفاعلة. من جانب آخر نقوم باستخراج هيكل الوثيقة ثم نُجَزِّئها لتقديمها بشكل يتناسب وممارسات المستخدمين. وبالتالي فإن الهيكل يلعب دورا مُهمّا في مراحل تحليل الوثائق. كما أننا نستغله لاستخراج المعارف بهدف بناء المكانز. في هذه المرحلة نعتبر أن الألفاظ الواردة في عناوين الأجزاء تُمثل أكثر من غيرها مصطلحات تعكس مدلول الوثيقة. كما أننا نستغل العلاقات المنطقية بين الأجزاء لاستخراج علاقات دلالية بين المصطلحات. إن إدماج هذه العناصر في بنية واحدة مَكّننا من بناء خريطة اجتماعية دلالية لموضوعي الزواج والأشربة من خلال تحليل كتب الحديث. النتائج التي تحصلنا عليها تُحَفّزنا على مزيد العمل من أجل توفير خدمات البحث والإبحار عبر الخريطة.

الكلمات الجوهرية. الويب الاجتماعي الدلالي، إعتمادية المعلومة، نوعية الاستخدام، بناء الكانز

1. المقدمة

إن صفحات الويب تمثل ثروة هائلة وفي نفس الوقت متباينة من المعارف. هذه الكتلة تتضخم باطراد كما يتزايد عدد المستخدمون الذين يرغبون في العثور على المعلومات بسهولة. إن كثرة المواضيع التي تتعرض لها صفحات الويب تجعل أن أي استعلام مكوّن من مجموعة من الكلمات المفاتيح يفرز في أغلب الأحيان عديد الصفحات في مجالات معرفية مختلفة. لاستخدام موارد الشبكة بشكل جيد وناجع لابد أن تتمكن محركات البحث من الوصول إلى موضوع كل صفحة ومعناها. علاوة على ذلك، فإن تنوع مصادر المعلومات على الشبكة العالمية (النصوص والصور، الخ) يدعو إلى معالجة المعلومة بطريقة مستقلة عن الشكل و كيفية التخزين، وهذا يعني، معالجتها عل مستوى المفاهيم. إن المعالجة الدلالية للموارد ستُمَكّن الآلات من القيام بعدة أعمال يقوم بها حاليا المستخدمون بطريقة يدوية. هذه الفكرة هي جوهر الويب الدلالي. حسب برنرس، فإن "رؤية الويب الدلالي تتمثل في توفير توصيف دلالي لمحتوى الويب يسمح للحاسوب بتأويله وبالتالي يمَكّنُه من القيام بالعديد من المهام التي يؤديها حاليا البشر" [برنرس-لي و من معه،2001]. في هذا المجال تلعب المكانز دورا أساسيا وهي في الحقيقة العنصر المركزي في الويب الدلالي. ولئن تعذر الاتفاق على تعريف مُعَين فإننا يمكن أن نستأنس بتعريف قروبر الذي يقول "إ ن المكنز هو توصيف علني لتصور" [قروبر، 1993]. وليكون المكنز ذا أهمية فإن هذا التصور لابد أن يكون مَحلّ اتفاق بين مجموعة من الأشخاص.

إن دور المكنز يتعدى تمثيل المعلومات فهو يمثل نموذجا يُسَهّل على المستخدم فهمَ المجال المعرفي و الإبحار فيه وهذا أمر أساسي خاصة في ظل وجود كمية كبيرة من المعلومات متعددة المصادر والأشكال و الأبعاد. بالإضافة إلى ذلك، فإن استعمال المكانز في أنظمة المعلومات يهدف إلى التقليل أو التخلص من ظاهرة خلط المفاهيم والالتباس الاصطلاحي والعمل من أجل التوصل إلى فهم مشترك لتحسين الاتصال والتشارك والعمل المشترك بما يمُكّن من إنتاج معارف يَسْهل إعادة استخدامها. وبالتالي فإن إعداد المكانز يسمح بدراسة جميع وجهات النظر وحصر الخلاف أو الاختلاف من أجل التوصل لرؤية مشتركة أو على الأقل واضحة.

بيد أن كثرة المعلومات و تنوع مصادرها وتعدد المتدخلين في عملية إنتاج المعلومة و نقلها سبَّب قلقا حول اعتمادية المعلومة. إن التحقق من صحة المعلومات، الذي كان الكتاب والمحررون ومسئولو المكتبات يقومون به صار من الآن فصاعدا من واجبات المستخدم [فيقنيو، 2005]. هذا الأخير يجد نفسه في كثير من الحالات غير قادر على تحديد مصدر المعلومة أو الحكم على مصداقيتها خاصة إذا كان العديد من الأطراف يشاركون في

إنتاجها أو نقلها. هذا هو الحال في بعض المنتديات حيث يتم تبادل الكثير من المعلومات بدون أي وسيلة لتحديد المصادر الأصلية أو المسارات المُتَّبعة. وسبب هذه المشكلة هو عدم وجود " هيئة تنظيم" تراقب ما هو موجود على شبكة الانترنت. و من ثم فإن السؤال هو : كيف يمكن ضمان أو تقييم اعتمادية المعلومات؟

في السنوات القليلة الماضية، قَدَّم مانويل زاكلاد الويب الاجتماعي الدلالي الذي يهدف إلى دراسة التفاعلات الاجتماعية، و كيف تؤدي إلى إفراز تمثيلات معرفية علنية وغنية [زاكلاد، 2007]. بالنسبة لاعتمادية المعلومة، يؤكد زاكلاد أن تحديد هوية كاتب الوثيقة ضروري لفهمها وتفسيرها واستغلالها [زاكلاد، 2007]. لا يمكن للمستخدم أن يستفيد من الوثيقة من دون الشعور بالثقة تجاه كاتبها. وعلاوة على ذلك، يجب على المستخدم أن يَشعر بنفس الثقة تجاه الجهات الفاعلة التي تنقل المعلومات. وبالتالي، ليس موضوع الوثيقة هو المعيار الوحيد للحكم على أهميتها [كسو و شان، 2006]. وفقا لدا كوستا بيريرا وباسي، فإن أهمية الوثيقة مرتبط بمفهوم الاعتمادية [دا كوستا بيريرا و باسي، 2007]. لهذا السبب نقول إن تحديد هوية الفاعلين (منتجين وناقلين) ودراسة سيرتهم هو خطوة أساسية وجوهرية في الحكم على الوثيقة.

حدث تغيير آخر متصل بتعميم الويب ألا وهو تنوع احتياجات المستخدمين. من وجهة النظر الاجتماعية، فإن الظاهرة الرئيسية هي اختلاف طريقة قراءة الوثيقة من مستخدم إلى آخر أو ما يسمى ب "نوعية استخدام الوثيقة". هذا المفهوم قدّمه أوسنساك جيل و كوندامين الذيْن يقولان أنه ينبغي لنا أن نُمَوْذِج كُلاً من النص ونوعيات الاستخدام وأن نوعيات الاستخدام لا تفوق من حيث العدد المستخدمين [أوسنساك جيل و كوندامين، 2004]. يجب أن نضيف أن المستخدم قد لا يكون مهتما بالوثيقة كاملة بل إنه في الغالب ينظر إليها كأجزاء تتفاوت من حيث الأهمية. إن مفهوم "نوعية الاستخدام" يجعلنا ننظر إلى الوثائق من منظور اجتماعي، وهو ما يعني أن جماعة من المستخدمين تشترك في نفس طريقة القراءة. زاكلاد يؤكد هذه الحقيقة بالقول إن الويب الحالي تسبب في تنوع الممارسات الجماعية التي ترتكز على الوثائق [زاكلاد، 2007]. لهذا السبب، ننظر لمفهوم نوعية الاستخدام كظاهرة اجتماعية. في الواقع، احتياجات المستخدم واستعماله للوثائق (أجزاء أو الوثائق) لها علاقة بعضويته في جماعة ممارسة معينة. وفقا لفينغر، جماعة الممارسة هي "مجموعة من المهنيين مرتبطين بصورة غير رسمية من خلال مشاكل مشتركة وسعي مشترك لإيجاد حلول لها، وبالتالي يمثلون مجتمعين مخزنا للمعرفة" [فينغر، 1998].

إن التواجد في الويب الاجتماعي الدلالي يفرض علينا عدة أمور منها إعداد المكانز وتقييم اعتمادية المعلومة والأخذ بعين الاعتبار اختلاف نوعيات الاستخدام. إن الهدف الأول من الورقة هو دراسة خصائص الوثائق العربية ودراسة إمكانية إدماجها في مشروع الويب الاجتماعي الدلالي. ومن ثَم فإننا نتطرق في الجزء الثاني، لمفهوم نوعية الاستخدام وأثره على تحليل الوثائق وتجزئتها. تجدر الإشارة إلى أن اهتمامنا ينصب على الوثائق شبه المنظمة كالكتب والمقالات العلمية والموسوعات حيث المظهر والأنماط المستعملة يساعدان عل فهم هيكل الوثيقة وتمييز أجزاءها. إننا نعتقد أن الهيكل يمكن أن يلعب دورا مُهمًّا في تحليل الوثائق من الناحية الدلالية والاجتماعية. ولذلك فإننا ندرس إشكالية إعداد المكانز من النصوص العربية في الجزء الثالث مع التركيز على هذا الجانب. كما أن الحضارة العربية قدّمت نموذجا فريدا من نوعه لضمان اعتمادية المعلومة من خلال علوم الحديث. سيُمَثل الجزء الرابع ربطا لمقومات اعتمادية المعلومة كما ما هو متواجد في البحوث المعاصرة من ناحية وقواعد علوم الحديث من ناحية أخرى. في الجزء الخامس، نحاول دمج هذه العناصر لتقديم نموذج اجتماعي دلالي جديد لخريطة اجتماعية دلالية فائقة التداخل. كما أننا نقترح بُنْية متكاملة لتحليل الوثائق العربية شبه المنظمة تسمح بتنظيمها وفهرستها حسب هذا النموذج (أنظر الجزء السادس). في الجزء السابع، نقدم التجارب التي قمنا بها في إطار مشروع يهدف إلى بناء خريطة اجتماعية دلالية من كتب الحديث. الجزء الثامن يختم هذه الورقة ويقترح سبلاً لمزيد البحث في هذا المجال.

2. نوعية الاستخدام

حسب زاكلاد، تُعتبر الوثيقة إنتاجا سيميائيا يجب أن يكون له سِمَات تُسَهّل الممارسات المتصلة باستغلاله [زاكلاد، 2007]. تسمح هذه السِّمات بتداول الوثيقة عبر مختلف جماعات الممارسة. لتسهيل الاستخدام، ينبغي أن تتم تجزئة الوثيقة إلى أجزاء متماسكة. كما يجب ربط الأجزاء لتمكين المستخدم من الإبحار الدلالي عبر الوثائق. مفهوم "نوعية الاستخدام" يُقِيمُ رابطا بين المستخدمين وأجزاء الوثائق. هذا يعني أن أفراد جماعة مُمارسةٍ معينة يشتركون في نوعية استخدام محددة وبالتالي يهتمون بأجزاء من الوثيقة دون أخرى أو يرَتِّبون هذه الأجزاء ترتيبا مُعَينا حسب احتياجاتهم.

إن تقسيم الوثيقة يعتمد على مستوى الحبوبية المختار. نعتقد أنه يمكن تحديد المستوى الأمثَل من خلال إجراء دراسة اجتماعية وذلك لتحديد ممارسات المستخدمين. وبالتالي، فإننا نفترض أن نفس المجموعة من الوثائق يمكن تحليلها وتجزئتها بطرق عديدة حسب نوعيات الاستخدام أو التنظيم الاجتماعي للمستخدمين [بونحاس و سليماني، 2009ب].

3. إعتمادية المعلومة

تُعَرَّف الاعتمادية على أنها "مدى ثقة المستخدم بالمعلومة" [نومان و رولكار، 2000]. في إطار الشبكة العالمية، حاول العديد من الباحثين تقديم أساليب ومقاييس وأدوات لتقييم الاعتمادية. من المُسَلّم به أن مفهوم "السلطة" هو البعد الأكثر أهمية في اعتمادية المعلومات. إلى جانب ذلك ، فإنه معيار مهم في تقييم الوثيقة [نومان و رولكار، 2000] [ريه، 2002]. يمكن للسلطة أن تُعَرّف بأنها مجموعة المؤشرات التي تُثبت (أو يمكن أن تُستَخدم لدراسة) مصداقية الجهات الفاعلة في إنتاج ونقل المعلومة. لدراسة السلطة في موقع على شبكة الإنترنت، يجب علينا التحقق من وجود معلومات من قبيل أسماء الكُتّاب ومعلومات الاتصال ونصوص حقوق النشر، وهلم جرا. كما ذكر زاكلاد، فإن تحديد هذه المؤشرات ضروري من أجل فهم الوثيقة وتفسيرها واستغلالها [زاكلاد، 2007]. إن دراسة اعتمادية الوثيقة من خلال مفهوم السلطة له علاقة ليس فقط بعناصر داخلية وإنما أيضا بأي معلومة خارجية يمكن أن تساعد في تقييم الوثيقة (مثل السير الذاتية للكُتّاب).

إن علم مصطلح الحديث يمَثل منهجية صلبة لضمان اعتمادية المعلومة. إن الحديث ينقل أحداثا تاريخية أو أقوالا تُنسب إلى شخص ما. عبر الأجيال قام أشخاص يُسَمّون الرُّواة بنقل الأحاديث. لأن هذه الروايات تنقل نصوصا وأحداثا تاريخية هامة، فإن العلماء العرب فرضوا قواعد صارمة لنقلها. إن كل راوٍ مُلْزَم بأن يعرض سلسلة الرواة الذين تلقى منهم الحديث. وبالتالي، نجد أن كل حديث مسبوق بسلسلة من الرواة تُسَمّى السَّنَد. أيضا، عندما ينقل الراوي (الشيخ) حديثا لأتباعه (التلاميذ) فإنه يستخدم أفعالا تُبين كيفية حصوله على الحديث من سلفه (شيخه) وهو ما يسمى صيغة الأداء.

إن دراسة الأسانيد هي خطوة أساسية يجب أن تَتِم قَبْل دراسة مضمون الحديث الذي يُسمّى المَتْن. ليكون الحديث مقبولا يجب أن ينقله رواة ثقات علما بأنه يتم تقييم الرواة من قِبَل علماء متخصصين على أساس سيرتهم. في النهاية يتم تصنيف الراوي حسب معيار المصداقية الذي يتألف من اثنا عشر درجة. بالإضافة إلى ذلك، ينبغي أن يكون السَّنَد مُتصلا مما يعني أن أي فجوة جغرافية أو زمنية بين راويين متتاليين تُعتبر مصدرا للشك في صحة الحديث. بالتالي، يجب أن تشمل دراستنا للحديث دراسة العلاقات التي تربط الرواة. بالإضافة للعلاقة بين الشيوخ والتلاميذ وعلاقات القرابة، يجب دراسة تواريخ ميلاد ووفاة الرواة وأماكن إقامتهم وأنسابهم. أخيرا، ينبغي أن يكون الحديث خاليا من التحيز. بعبارة أخرى ينبغي أن لا يكون للراوي أسباب سياسية أو مذهبية لتحريف الحديث. لنكون متأكدين من موضوعية النقل، يجب أن نجمع ونقارن روايات مختلفة لنفس الحديث لتحديد حالات الزيادة أو النقصان التي قد تحدث و التّعرف على الروايات الشاذة.

4. إعداد المكانز

إن البحوث في مجال بناء المكانز أو بصفة عامة الموارد المُصطلحية تُثبت في نفس الوقت فائدة وصعوبة هذا العمل [عطية و من معه ، 2008]. إن عدم وجود موارد لغوية خاصة على المستوى الدلالي أجبر بعض الباحثين على انتهاج طريقة يدوية (على سبيل المثال [فلبوم و من معه ، 2006]؛ [زيدي والعسكري ، 2005]؛ [عطية و من معه ، 2008]) ، أو شبه آلية [رودريغيز و من معه، 2008] أو استخدام موارد خارجية [الأنصاري و من معه، 2007].

إن هذه السطور لا تكفي لاستعراض مراحل بناء المكانز أو تفصيل كل مكوناتها. لكن المكنز في حدّه الأدنى لابد أن يشتمل على عنصرين هما : المصطلحات والعلاقات الدلالية التي تربطها. وبالتالي، فإن استخراج المعارف يتطلب تحديد المصطلحات ذات الصلة بالمجال الدلالي والعلاقات المعنوية بينها. العديد من البحوث تعتبر أن الأسماء هي العناصر التي تمثل موضوع الوثيقة (مثلا [بولكنادل، 2006]). بالإضافة إلى ذلك، فإننا نُمَيز بين نوعين من الأسماء : الأسماء البسيطة المتكونة من كلمة واحدة والأسماء المركبة من عدة كلمات أو التراكيب الاسمية. وهكذا، فان الخطوة الأولى هي التحليل الصرفي الذي يتيح تحديد الأسماء البسيطة. وثانيا، ينبغي إجراء تحليل نحوي لتشكيل مُركبات اسمية وفقا لقواعد النحو العربي. بعد ذلك يمكننا استخراج قائمة المصطلحات المرشحة لتمثيل المجال المعرفي. غير أنه ينبغي علينا ترتيبها وفقا لأهميتها بالنسبة إلى المجال. أخيرا يجب استخراج العلاقات الدلالية التي تُمَكّن من ربط المصطلحات وبالتالي تشكيل هيكل المكنز. في الجزأين الموالين ندرس تِباعا الظواهر اللغوية المرتبطة باستخراج المصطلحات من النصوص وكيفية ترتيبها وتنظيمها في شكل مكنز من خلال العلاقات الدلالية.

1.4 إستخراج المصطلحات

إن ما يميز اللغة العربية هو كونها لغة اشتقاقية تعتمد على التصريف، يضاف إلى ذلك التصاق الحروف ببعضها وغياب الشكل في معظم النصوص المتاحة. ونتيجة لذلك، فان النصوص العربية ملتبسة على المستوى الصرفي والنحوي والدلالي. هذا الالتباس يؤثر على عدة مراحل في عملية بناء المكانز والبحث عن المعلومة ما يستدعي

منا تخصيص مرحلة خاصة تُعنَى بإزالة أو تقليل الالتباس. لاركي اقترح التشذيب الخفيف لإزالة البوادئ واللواحق [لاركي و من معه، 2002]. إن عدم وجود تحليل صرفي عميق يولّد قائمة كلمات ملتبسة قد تؤثر على أداء نظام البحث كما أنها لا تصلح لإعداد المكانز.
بصفة عامة يرتكز استخراج التراكيب الاسمية إمّا على طرق إحصائية أو على برامج تُنفّذ قواعد لغوية. غير أن كِلا الحلين له نقائص. أما بالنسبة للأول فإنه لا يُمكّن من استخراج التراكيب النادرة. وأما الثاني فإنه مرتبط كثيرا باللغة ولا يُمكنه التعامل مع الطبيعة المُعقدة للتراكيب. لذا فإن الحل الأمثل يَكْمُن في الطرق الهجينة. إن البحوث التي تعرضت لإشكالية استخراج التراكيب الاسمية من النصوص العربية تفتقر لمرحلة التحليل الصرفي أو للتدابير الإحصائية التي تمَكّن من غربلة الأسماء المستخرجة أو تتجاهل أنواعا معينة من التراكيب [بونحاس و سليماني، 2009أ]. لقد طورنا برمجية حاولنا من خلالها تجاوز هذه النقائص وقد تحصلنا على نتائج طيبة (راجع [بونحاس و سليماني، 2009أ]). في هذه البرمجية أدمجنا أداة العنونة النحوية التي طوّرها [ذياب و من معه، 2004] وهي تستخدم السياق داخل الجملة لإزالة الالتباس في الحلول التي يقترحها المحلل الصرفي أرامورف [هاجيك و من معه، 2005]. في الواقع، كان من الأفضل استخدام أداة مادا [حباش و من معه، 2009] التي تؤدي كل هذه المهام في مرحلة واحدة. بالإضافة إلى ذلك ، يجب تحسين طريقة إزالة الالتباس عن طريق النظر في أنواع أخرى من السياق وهو ما سنفصله لاحقا.

2.4 ترتيب المصطلحات وتنظيمها

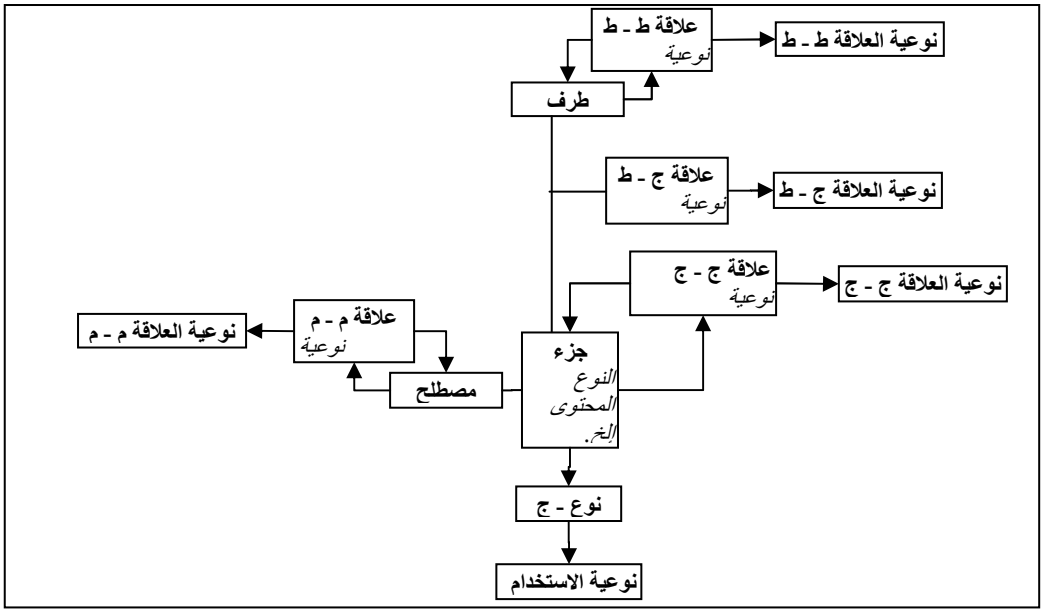
إن ترتيب المصطلحات - خاصة الأسماء البسيطة - حسب أهميتها بالنسبة للمجال يعتمد بالأساس على التردد. ومن أكثر القياسات المستعملة ت.ل.ع.ت.و(تردد اللفظة، عكس تردد الوثيقة). لقد قدّمنا طريقة جديدة تعتمد هذا القياس بشكل هرمي [بونحاس و سليماني، 2010] وهو ما سنُفصله في الجزء 6.
إن البحوث التي اهتمت بشكل أو بآخر بإشكالية البحث عن المعلومة في الوثائق العربية تعتمد أساسا على الطرق الإحصائية لتقييم درجة التقارب أو الترابط بين الألفاظ العربية [مقبل و من معه، 2001] [بولكنادل، 2006]؛ [برونزل و سبيليوبولو، 2006] [بنتو و من معه، 2007]؛ [يوسفي و من معه، 2008] [القباني و من معه، 2009]. الخطوة الأولى هي تحديد السياقات التي يتردد فيها كل مصطلح. الكلمات التي تتردد في سياقات متشابهة من المفترض أن تكون متقاربة من حيث المعنى. درجة التقارب تُحَدَّد عن طريق احتساب المسافة بين السياقات. وبالتالي، فإن السياق هو المعلمة الأهم. على الرغم من أننا يمكن أن نحدد السياق بطرق مختلفة ،فإن المسافة الأكثر استخداما هي عدد تساوق الكلمات (أي في الجملة ، الفقرة ، الجزء، وهلم جرا). في نفس السياق قدمنا طريقة إحصائية تتميز باستغلال هيكل الوثيقة لربط المصطلحات بعلاقات دلالية [بونحاس و سليماني، 2010] (أنظر الجزء 6)

5. النموذج الاجتماعي الدلالي

لقد صمّمنا هذا النموذج ليستجيب لمتطلبات الويب الاجتماعي الدلالي و هو يتضمن العناصر التي ناقشناها سابقا. إنه متكون من أربعة أبعاد. ففي جزءه العمودي نجد المحورين الاجتماعيين. من ناحية نأخذ بعين الاعتبار مفهوم السلطة من خلال ربط الوثائق أو أجزاء الوثائق من جهة والجهات الفاعلة المنتجة أو الناقلة للمعلومة من جهة أخرى. هذه الجهات يمثلها أفراد من الفئة "طرف". الفئة "علاقة ط – ط" تسمح بتمثيل أنواع مختلفة من العلاقات بين هذه الأطراف.من ناحية أخرى كل جزء من الوثيقة يتم ربطه حسب نوعه (نوع – ج) بنوعيات الاستخدام الممكنة. المحور الأفقي يُعنى بربط أجزاء الوثيقة بروابط هيكلية غير أننا لا نفرض أنواعا معينة من الروابط التي يجب تحديدها حسب المجال ونوعية الوثائق وحاجيات المستخدمين. أخيرا، نربط كل جزء بالمصطلحات المنَظمة عبر علاقات دلالية. لكل رابط وزن يعكس درجة الارتباط بين الجزء والمصطلح أو بين المصطلحين. النموذج الاجتماعي الدلالي المقترح يرد في الرسم الأول.

6. البنية المقترحة

نعتمد وجهة نظر الويب الاجتماعي الدلالي التي تقول إن الوثيقة هي نتيجة لمعاملات كُلّية تتألف من عدة معاملات جزئية بين العديد من الأطراف الفاعلة. وفقا لزاكلاد، فإن الوثائق تزداد تفتتا كلما ازداد عدد الأطراف الفاعلة المشاركة في إنتاجها [زاكلاد، 2007]. إلى جانب ذلك فإن المعاملات الجزئية ترتبط داخل الوثيقة بعناصر تتعلق بهيكل الوثيقة عناوين الأجزاء وأرقامها وسمات أخرى تحدد حالة الأجزاء أو صلاتها. يمكننا أن نستنتج أن الهيكل المنطقي للوثيقة يناظر العملية الاجتماعية التي أدت إلى إنتاجها. هذا هو السبب الذي يجعلنا نعتمد على الهيكل المنطقي للوثيقة وهو يساعدنا على استخراج المصطلحات والعلاقات بينها كما أننا يمكن أن نستعمله في إزالة الالتباس اللغوي كما سنبينه لاحقا.



الرسم الأول: النموذج الاجتماعي الدلالي

إننا نقترح مقاربة اجتماعي دلالية تتكون من المراحل الواردة في الجدول الأول. الجدول يحدد لكل مرحلة الفئات المستهدفة من النموذج الاجتماعي الدلالي.

1- الدراسة الاجتماعية:	الفئات المُستهدفة من النموذج
1.1 - دراسة مراحل إنتاج الوثائق ما يُمَكّن من تحديد طبيعة الجهات الفاعلة وأدوارها	نوعية العلاقة ج – ط
2.1 - دراسة التنظيم الاجتماعي للجهات الفاعلة ونوعية العلاقات الاجتماعية	علاقة ط – ط نوعية العلاقة ط - ط
3.1 - تحديد فئات المستخدمين وممارساتهم	نوعية الاستخدام
4.1 - تحديد أنواع الأجزاء وكيفية ترابطها	نوع – ج نوعية العلاقة ج - ج
5.1 - ربط أنواع الأجزاء بممارسات المستخدمين	الربط بين نوع – ج و نوعية الاستخدام
2- استخراج هيكل الوثيقة	جزء علاقة ج – ج
3 - تقييم الاعتمادية	طرف نوعية العلاقة ط - ط
4 - التحليل اللغوي	مصطلح
5- فهرسة الوثائق	الربط بين جزء و مصطلح
6 - استخراج العلاقات الدلالية	علاقة م – م نوعية العلاقة م - م

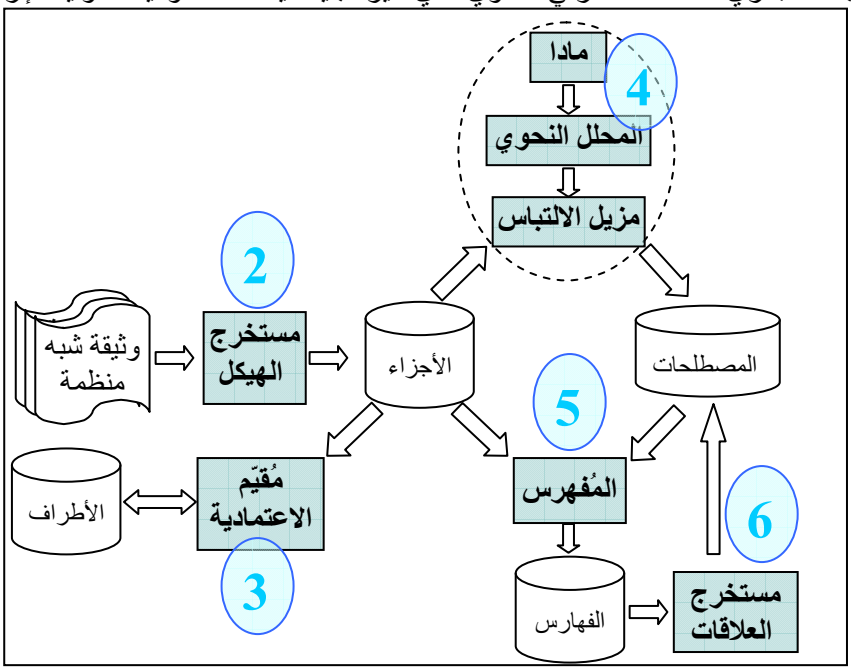
الجدول الأول: مراحل تحليل الوثائق

لقد فَصَّلنا مراحل الدراسة الاجتماعية (راجع [بونحاس و سليماني، 2009ب]) وهي تعتمد على المراوحة بين التحليل اليدوي السريع للوثائق والمقابلات مع الخبراء. بعد هذه الدراسة، نقوم باستخراج هيكل الوثيقة على ثلاثة مراحل. نبدأ باستخراج الهيكل المادي لكل وثيقة. ثم نستخرج الهيكل المنطقي من خلال تحديد عناوين الأجزاء الكبرى كالفصول أو الأبواب. أخيرا، علينا التعرف على المكونات الجزئية لكل جزء كُلّي ما يسمى التحليل الجزئي للهيكل المنطقي. الفكرة من وراء الفصل بين المستويين هي تطوير برمجية تحليل منطقية كُلية قابلة لإعادة الاستخدام وعدة محللات منطقية جزئية كل واحدة منها متخصصة في تحليل نوع معين من الأجزاء الكلية. إن مَهمة المحلل المنطقي الجزئي تكمن في تحليل نصوص شبه منظمة وهو يستعمل قاعدة لغات الحرة يتم تَعلّمها بطريقة شبه آلية من مجموعة من الأمثلة. ومن أهم الأدوار التي تقوم بها هذه البرمجيات تحديد أسماء الجهات الفاعلة وأدوارها والمعلومات الأخرى المتعلقة بها كالعناوين وأسماء المنظمات التي ينتمي إليها الأشخاص. هذه المعلومات يتم استغلالها في من طرف مُكَوّن متخصص يقيم اعتمادية الوثيقة بطريقة آلية (أنظر مُقيّم الاعتمادية

في الرسم الثاني). غير أن مثل هذا المُكون يحتاج إلى معلومات خارجية ولذلك نُدمج قاعدة بيانات معلومات بيانية تحتوي على معطيات إضافية تُمَكّن من تقييم مصداقية الجهات الفاعلة [بونحاس و سليماني، 2009ب].
فيما يتعلق بالجانب اللغوي قمنا بتطوير نسخة جديدة من المُحلل الذي جربناه سابقا في مجال البيئة [بونحاس و سليماني، 2009أ]. التغييرات تشمل تعويض أداتي العنونة النحوية [ذياب و من معه، 2004] والمحلل الصرفي أراموورف [هاجيك و من معه، 2005] بأداة مادا [حباش و من معه، 2009] ما سرّع وبسّط عملية التحليل. بالإضافة إلى ذلك نقترح الاستفادة من هيكل الوثيقة لإزالة الالتباس. على سبيل المثال إذا وجدنا كلمة "مسكر" في فقرة عنوانها "الخمور" علمنا أن المقصود هو "مُسْكِر" وليس "مُسَكَّر". و بالتالي فإننا لا نقتصر على السياق الصرفي أو النحوي داخل الجملة كما هو الحال في النسخة القديمة بل نتجاوزه للنظر إلى السياق داخل الوثيقة من أجل إزالة الالتباس. بالنتيجة نحصل على قائمة المصطلحات البسيطة والمُرَكبة لكل جزء من الوثيقة.
لفهرسة الوثائق، نستعمل النموذج الذي تمت تجربته على مدونة وثائق تتحدث عن الحيوانات ومكن من استخراج شجرة الحيوانات [بونحاس و سليماني، 2010]. نبدأ بتحديد المصطلحات التي تُمَيز كل وثيقة من خلال القياس ت.ل.ع.ت.و(تردد اللفظة، عكس تردد الوثيقة). في هذه الخطوة، يتم حساب تردد كل مصطلح حسب موقعه في الهيكل المنطقي من خلال المعادلة التالية:

$$\text{تردد(م،د)} = \sum \text{(تردد(م،ج)*وزن(ج))}$$

لو فرضنا أن **ن** هو عمق الوثيقة أي عدد مستويات عناوين الأجزاء فإن العناوين ذات المستوى الأعلى تأخذ الوزن **ن**. عناوين المستوى الثاني تأخذ الوزن **ن**-1 وهكذا. حسب المعادلة السابقة فإن الألفاظ التي تتردد في العناوين الأعلى ستحصل على وزن أعلى لأنها الأفضل من حيث تمثيل موضوع الوثيقة وهذا ما نسميه الفهرسة النوعية. ثم إنه يتم توريث فهرس الوثيقة إلى أجزاءها بشكل نازل عبر استعمال القياس ت.ل.ع.ت.و بشكل هرمي. المصطلحات التي يتجاوز وزنها بالنسبة للوثيقة 0.5[1] يتم توريثها الأجزاء التي تحمل الوزن **ن**. بعد ذلك نستعمل القياس ت.ل.ع.ت.و بين هذه الأجزاء. عمليات القياس والتوريث المتلاحقة تُمكن من فهرسة الأجزاء في مختلف المستويات. [بونحاس و سليماني، 2010].
إن هذا النموذج يسمح أيضا باستخراج علاقات دلالية بين المصطلحات. من ناحية نربط الألفاظ التي تظهر في عنوان جزء معين بالألفاظ الموجودة في فهارس أجزاءه الفرعية و هو ما يُمَكّن من ربط المصطلحات بشكل عمودي. ومن ناحية أخرى يتم ربط الألفاظ التي تتردد في نفس الجزء بعلاقات أفقية. كذلك اقترحنا قياس درجة التقارب بين المصطلحات عبر القياس ل.ل.ر [دونينق، 1994]. الرسم الثاني يُشَكّل البنية المقترحة حيث نُبين أرقام المراحل في دوائر مرقمة. إنها بينة قائمة على إعادة الاستعمال حيث نُدمج عدة برمجيات منها ما طوره باحثون آخرون كمادا ومنها ما طورناه بأنفسنا وأدمجناه بدون أي تغيير كالمفهرس ومنها ما أدخلنا عليه تغييرات حَسّنت أداءه بشكل جذري كالمحلل الصرفي النحوي الذي غيرنا بنيته ليعتمد مادا وهيكل الوثيقة لإزالة الالتباس.



الرسم الثاني: البنية المقترحة

[1] لقد تم اختيار عتبة التوريث 0.5 بعد عدة تجارب [بونحاس و سليماني، 2010]

إن المقاربة المقترحة تُمَثّل منهجية عامة لتحليل الوثائق في الويب الاجتماعي الدلالي كما أن النموذج الذي في الرسم الأول هو نموذج عالي يمكن تخصيصه ليتكيف مع عدة مجالات معرفية. وقد حرصنا على تطوير البرمجيات بطريقة تُسهل إدماجها في أي مشروع في الويب الاجتماعي الدلالي. رغم ذلك فإننا بحاجة لمثال يجسد أفكارنا ويُمَكّننا من تقييم هذه البرمجيات وهو ما سنهتم به في الجزء التالي.

7. التجارب والنتائج

لقد قمنا بعدة تجارب في مجالات مختلفة كالحديث [بونحاس و سليماني، 2009ب] والبيئة [بونحاس و سليماني، 2009أ] والحيوانات [بونحاس و سليماني، 2010]. و قد شملت هذه التجارب الجانبين الاجتماعي والدلالي كل على حدة. غير أننا نريد أن نقدم مثال مشروع يأخذ بعين الاعتبار كِلا الجانبين نُدمج فيه كل البرمجيات التي طورناها سابقا مع التحسينات التي تحدثنا عنها في الجزء السادس.
إن كتب الحديث تمثل مجالا تطبيقيا مِثاليا لهكذا مشروع. إنها وثائق ثرية من الناحية الدلالية إذ تتعرض بالإضافة إلى المسائل الدينية، لمواضيع اجتماعية (كالزواج) وأمور الحياة اليومية (كالمأكل والمشرب) وقواعد الطب والصحة وغيرها. من الناحية الاجتماعية يتميز هذا الميدان بتدخل عدة أطراف كالرواة والشخصيات التي تساهم في القصة بالنسبة للأحاديث التي تسرد أحداثا والعلماء الذين قَيَّموا الرواة من حيث الموثوقية أو الأحاديث من حيث الاعتمادية بالإضافة إلى الشُرّاح والفقهاء. وهو بالنتيجة مجال للاختلاف وتبادل وجهات النظر. وكما أسلفنا فإن علوم الحديث قدمت نموذجا فريدا لاعتمادية المعلومة. وفي الجملة فإن هذه النصوص وما يتعلق بها من علوم تُمثل جانبا بالغ الأهمية من الحضارة العربية الإسلامية من الناحية التاريخية والعلمية يصعب حصرها وهي في النهاية من أهم مميزات هذه الحضارة. وكنتيجة لهذا الثراء ونظرا لأن عدة أحاديث ما تزال إلى يومنا هذا تحتاج لأن تُقَيَّم من حيث الاعتمادية وتُدرس من حيث الدلالة، فإننا نجد أنفسنا تجاه نفس المشاكل التي ناقشناها في المقدمة من صعوبة الحصول على المعلومة وضرورة تمثيل المعلومة بشكل دلالي وتقديمها حسب ممارسات المستخدمين.
إن الحل الذي نُقدمه هو خريطة اجتماعية دلالية متعددة الأبعاد تُجَسّد النموذج الذي اقترحناه من خلال ربط الأحاديث والمصطلحات والأطراف الفاعلة ونوعيات الاستخدام. في الوقت الحاضر نهتم بكتب الحديث الستة الأكثر اعتمادا وهي: صحيح البخاري وصحيح مسلم و سنن أبي داود وسنن النسائي وسنن الترمذي وسنن ابن ماجة. لقد قمنا بدراسة اجتماعية مَكّنت من تحديد نوعيات الاستخدام وربطها بأنواع الأجزاء. ثم قمنا باستخراج أجزاء الوثائق كالأحاديث وأسماء الرواة والعبارات التي تُذيل الأحاديث لتقدم وجهة نظر العَالِم حول صحة الحديث أو تأويله أو لتحيل إلى روايات أخرى لنفس الحديث. تحصلنا على قاعدة لغات حرة تسمح باستخراج أسماء العلم العربية من الوثائق وبتحليل مُكونات كل اسم (كاللقب والكنية والنسب والنسبة) [بونحاس و سليماني، 2009ب]. إن هذه القاعدة صُمّمت بشكل يساعد على إعادة استخدامها وقد حققت نتائج طيبة في التعرف على أسماء الرواة. في نسختها الحالية تعمل هذه القاعدة عل النص مباشرة دون أي تحليل لغوي ما يجعلها ترتكب بعض الأخطاء. و بالتالي فإن إدماجها مع المحلل اللغوي سيساعد على تحسين أداءها وهو ما ننوي القيام به في المستقبل القريب.
ثم إننا درسنا نموذج اعتمادية المعلومة وقمنا بخطوة أولى مَكّنت من التعرف على هوية كل راو وبالتالي تَمَكّنّا من ربط الرواة بقاعدة بيانات تحتوي على معلومات ضافية حول كل رواة الحديث في مختلف الفترات التاريخية [بونحاس و سليماني، 2009ب]. إننا بصدد تطوير مُقَيّم الاعتمادية الذي يستغل هذه المعلومات غير أنه لا يمكننا تقديم تفاصيل حوله نظرا لأنه يجب علينا مقارنة قراراته بآراء العلماء الواردة في الوثائق. لكن يمكننا القول إن مهمّته هي تصنيف كل حديث ضمن الفئات المعروفة في علم الحديث: "ضعيف" و "حسن" و "صحيح".
بالنسبة للمحور الدلالي قمنا بتطبيق الفهرسة النوعية للوثائق. كما قمنا بتجارب من أجل استخراج المصطلحات والعلاقات الدلالية بينها في موضوعَي الزواج والأشربة. في هذه المرحلة تمكنا من رفع دقة استخراج المصطلحات ففي النسخة الأولى كانت دقة استخراج التراكيب الاسمية تساوي 65,50% بينما بلغت نسبة المصطلحات الصحيحة المستخرجة باستعمال النسخة الثانية 87,14 % وهذا يثبت نجاعة أداة مادا ومزيل الالتباس في نسخته الجديدة التي تعتمد على هيكل الوثيقة. بالنسبة للعلاقات الدلالية، تحصلنا على مجموعات من المصطلحات متجانسة كما تمكنا من تحديد المصطلح العام لكل مجموعة. الجدولان الثاني والثالث يقدمان أمثلة من المجموعات في موضوعَي الأشربة والزواج تباعا. نلاحظ في باب الأشربة أن المجموعة الثانية تحتوي على أسماء المواد التي تصنع منها الخمر ولذلك تم اختيار هذه الكلمة كمصطلح عام. بصفة عامة تمكّنا من تجميع المصطلحات بشكل جيد ولكننا سجلنا بعض التذبذب خلال استخراج العلاقات العمودية مقارنة بالتجارب التي أجريناها في مجال الحيوانات. هذا راجع إلى أن عناوين الأجزاء في كتب الحديث تمثل مواضيع لا مصطلحات ولذا نجد مثلا أن كل المصطلحات في باب الزواج تتصل بشكل عمودي بكلمة "نكاح". إن حالات الخطأ راجعة

أيضا إلى كوننا اقتصرنا على موضوعين فقط. إننا نعتقد أن هذه النتائج ستتحسن بشكل جذري إذا قمنا بتحليل كل المواضيع.

المجموعة	المصطلح العام
دباء ؛ نَقِير ؛ جَرّة ؛ مزفت ؛ قِدْح ؛ سِقاء ؛ آنِية ؛ وِعاء ؛ مقير ؛ قِرْبَة ؛ الحنتمة و الدباء و النَّقِير ؛ حنتمة ؛ ظَرْف	وعاء
خَمْر ؛ عِنَب ؛ تَمْر ؛ عَسَل ؛ حِنْطَة ؛ شَعِير ؛ بُسْر ؛ بتع ؛ نَبيذ ؛ زَبيب ؛ عَصِير ؛ زَهْو	خَمْر
خَمْر ؛ عَسَل ؛ شَرَاب؛ نَبيذ ؛ عَصِير ؛ ماء ؛ شَراب ؛ نَبيذ الجَرِّ ؛ نَبيذ ؛ لَبَن	شُرْب أو شَرَاب

الجدول الثاني: مجموعات المصطلحات في موضوع الأشربة

المجموعة	المصطلح العام
إمْرَأة ؛ نِساء ؛ جارِية ؛ مَرْأة ؛ بِكْر ؛ ثَيِّب ؛ نِساء ؛ فَتاة ؛ اِمْرَأةٌ غَنِيَّةٌ ؛ اِمْرَأة بِكْر	مَرْأة أو إمْرَأة
نِكاح ؛ شغار ؛ تَزْوِيج ؛ مُتْعَة ؛ تَزَوُّج	نِكاح
أُمّ ؛ أُخْت ؛ عَمَّة ؛ خالَة ؛ بِنْت	نِساء

الجدول الثالث: مجموعات المصطلحات في موضوع الزّواج

بعد أن استكملنا مُعظم مراحل بناء الخريطة الاجتماعية الدلالية، يجدر بنا إيجاد أو تطوير البرمجيات التي تمَكّن المستخدم من استغلالها. إن المرحلة القادمة تتمثل في عرض الخريطة بشكل تفاعلي وتجهيزها بوسائلَ لبحث وتصفية وتصنيف المعلومات. في هذا الصدد يمكننا استعمال تقنيات البحث التقليدية المعتمدة على الاستعلام أو اقتراح مقاربة إبحارية كما هو الحال في المشاريع كارينا وسِبيل [كرمبس و رانواز، 2000]. كما أن الخريطة يمكن أن تُستغل لتكوين كتب افتراضية جديدة على طريقة فالكات [فالكات و من معه، 2004]. مثلاً يمكننا تجميع كل الأحاديث التي نقلها راو معين أو روايات الحديث الواحد من كتب مختلفة.

8. الخاتمة

لقد مثلت هذه الورقة فرصة لدراسة التحديات التي تواجه الوثائق العربية لتندمج في الويب الدلالي الاجتماعي. كما مثلت مناسبة لعرض موروث حضاري عربي نعتقد أنه يمكن أن يساهم في حل مشكلة اعتمادية المعلومة. إننا نعتقد أن البرمجيات التي طورناها على أسس علم مصطلح الحديث لن تُمَكّن من دراسة صحة الأحاديث فحسب بل إنها تُمَثل نواة صلبة لتقييم اعتمادية الوثائق في الشبكة العالمية. إن النموذج والمقاربة المقترحان يستغلان هيكل الوثيقة من أجل تمثيل المعارف التي تتضمنها وتسهيل عملية البحث عن المعلومة. إن الهيكل يمثل من ناحية مراحل وأدوار الجهات الفاعلة في عملية إنتاج المعلومة. ومن ناحية أخرى يُمْكِننا من خلاله تكييف وسائل البحث والإبحار حسب ممارسات المستخدمين وهو ما جسدناه عبر مفهوم "نوعية الاستخدام". إن هذا المفهوم يقضي بتقديم أجزاء الوثيقة بشكل يتناسب مع رغبات المستخدمين وطبيعة عملهم. كما أننا أثبتنا أن الهيكل يساعد في معالجة ظاهرة الالتباس اللغوي التي تمثل حسب تجربتنا أوكد التحديات نظرا لصعوبة التعامل مع النصوص التي تفتقر للشكل. كما أكدنا نتائج سابقة تحصلنا عليها في تجارب على وثائق تتحدث عن الحيوانات تتمثل في استغلال الهيكل لاستخراج المصطلحات والعلاقات الدلالية بينها. لقد تحصلنا عل نتائج مُرْضية باستعمال نفس الطريقة مع كتب الحديث. لقد قمنا بإدماج البرمجيات التي طورناها لبناء خريطة اجتماعية دلالية من كتب الحديث يمكن استغلالها في عملية البحث والإبحار في مختلف المواضيع. إن اختيارنا تم بناء على دراسة معمقة للخصائص المميزة لهذه الكتب حيث أنها وثائق ثرية من الناحيتين الاجتماعية والدلالية. غير أن هذا المشروع يحتاج لمزيد من الجهد يتمثل في تحسين التنسيق بين مختلف البرمجيات. على سبيل المثال ننوي الاستعانة بالتحليل اللغوي لإزالة حالات الالتباس في مرحلة استخراج أسماء الرواة. كما أننا ندرس المقاربات الممكنة من أجل عرض الخريطة وتفعيل وسائل البحث والإبحار فيها.

9. المصادر والمراجع

[الأنصاري و من معه، 2007]

Alansary S., Nagi M. & Adly N. Communicating in Arabic in Cyberspace. *Information and Communication Technology International Symposium (ICTIS07), Arabic Natural Language Processing Workshop*, Fez, Morocco, 3-5 April 2007.

[القباني و من معه، 2009]

Al-Qabbany A., Al-Salman A. & Almuhareb A. An Automatic Construction of Arabic Similarity Thesaurus. *The 3rd IEEE International Conference on Arabic Language Processing (CITALA2009)*, pp. 31-36, Rabat, Morocco, May 4-5, 2009.

[أوسنساك جيل و كوندامين، 2004]

Aussenac-Gilles N. & Condamines A. Documents électroniques et constitution de ressources terminologiques ou ontologiques. *Information-Interaction-Intelligence* 4(1): 75-94, 2004.

[برنرس-لي و من معه،2001]

Berners-lee T., Hendler J. & Lassila O. The semantic Web. *Scientific American*, Mai 2001, pp. 34-43.

[برونزل و سبيليوبولو، 2006]

Brunzel M. & Spiliopoulou M. 2006. Discovering Multi Terms and Co-hyponymy from XHTML Documents with XTREEM. *Proceedings of Workshop on Knowledge Discovery from XML Documents (KDXD 2006)*, pp. 22-32, Springer LNCS 3915, Singapur.

[بنتو و من معه، 2007]

Pinto D., Rosso P., Benajiba Y., Ahachad A. & Jiménez-salazar H.. Word Sense Induction in the Arabic Language: A Self-Term Expansion Based Approach. *The 7th Conf. on Language Engineering, The Egyptian Society Of Language Engineering, ESOLE-2007*, pp. 235-245, Cairo, Egypt, December 5-6, 2007.

[بولكنادل، 2006]

Boulaknadel S. 2006. Utilisation des syntagmes nominaux dans un système de recherche d'information en langue arabe. *Conférence Francophone en Recherche d'Information et Applications CORIA 2006*, pp. 341-346, Lyon, France, Mars 15-17, 2006.

[بونحاس و سليماني، 2009أ]

Bounhas I. & Slimani Y. A hybrid Approach for Arabic Multi-Word Term Extraction. *IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'09)*, pp. 429-436, Dalian, Chine, September 24-27, 2009.

[بونحاس و سليماني، 2009ب]

Bounhas I. & Slimani Y. A social approach for semi-structured document modeling and analysis. *International Conference on Knowledge Management and Information Sharing KMIS 09*, pp. 95-102, Madeira, Portugal, 6 - 8 October, 2009.

[بونحاس و سليماني، 2010]

Bounhas I. & Slimani Y. A hierarchical approach for semi-structured document indexing and terminology extraction. *International conference on information retrieval and knowledge management (CAMP'2010)*, (à paraître).

[حباش و من معه، 2009]

Habash N., Rambow O. & Roth R. MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. *The 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*, pp.102-109, Caire, Egypte, 22-23 Avril, 2009.

[دا كوستا بيريرا و باسي، 2007]

Da Costa Pereira C. & Pasi G. Fuzzy Indices of Document Reliability. *Applications of Fuzzy Sets Theory, Lecture Notes in Computer Science 4578/2007*, Springer Berlin / Heidelberg, 2007, pp. 110-117.

[دونينق، 1994]

Dunning, T. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1): 61-74, 1994.

[ذياب و من معه، 2004]
Diab M. T., Kadri H. & Jurafsky D. Automatic Tagging of Arabic Text: From raw text to Base Phrase Chunks. *The 5th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL04)*, pp.149-152, Boston, Massachusetts, Mai 2-7, 2004.

[رودريغيز و من معه، 2008]
Rodríguez H., Farwell D., Farreres J., Bertran M., Alkhalifa M., & Martí M. A. Arabic WordNet: Semi-automatic Extensions using Bayesian Inference. *Proceedings of the the 6th Conference on Language Resources and Evaluation LREC2008*, pp.1702-1706, Marrakech, Morocco, 28-30 May 2008.

[ريه، 2002]
Rieh, S. Y. Judgment of Information Quality and Cognitive Authority in the Web, *Journal of the American Society for Information Science and Technology*, 53(2):145-161, 2002.

[زاكلاد، 2007]
Zacklad M. Processus de documentarisation dans les Documents pour l'Action (DopA). *Babel - edit -, Le numérique: impact sur le cycle de vie du document*, ENSSIB, 2007.

[زيدي والعسكري ، 2005]
Zaidi S. & Laskri M. T. A cross-language information retrieval based on an Arabic ontology in the legal domain. *Proceedings of the International Conference on Signal-Image Technology and Internet-Based Systems (SITIS'05)*, pp. 86-91, Yaoundé, Cameroon, 2005.

[عطية و من معه ، 2008]
Attia M., Rashwan M., Ragheb A., Al-Badrashiny M., Al-Basoumy H. & Abdou S. A Compact Arabic Lexical Semantics Language Resource Based on the Theory of Semantic Fields. *Proceedings of the 6th international conference on Advances in Natural Language Processing*, pp.65 – 76, Gothenburg, Sweden, 2008.

[فالكات و من معه، 2004]
Falquet G., Jiang C. L. M.& Ziswiler J.C. Intégration d'ontologies pour l'accès à une bibliothèque d'hyperlivres virtuels. *14ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle (RFIA 2004)*, Toulouse, France.

[فلبوم و من معه ، 2006]
Fellbaum F., Alkhalifa M., Black W. J., Elkateb S., Pease A., Rodríguez H. & Vossen P. Building a WordNet for Arabic. *Proceedings of the the 5th Conference on Language Resources and Evaluation LREC2006*, May, 2006.

[فيقنيو، 2005]
Vignaux G. La recherche d'information: Panorama des questions et des recherches. *Research report*, CNRS-MSH, Paris Nord, 2005.

[فينغر، 1998]
Wenger, E. *Communities of Practice: Learning, Meaning and Identity*, Cambridge University Press, 1998.

[قروبر، 1993]
Gruber T. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5(2):199-220, 1993.

[كرمبس و رانواز، 2000]
Crampes M. & Ranwez S. Ontology-Supported and Ontology-Driven Conceptual Navigation on the World Wide Web. *HT'00, the 11th ACM Conference on Hypertext*, San Antonio, Texas, 2000.

[كسو و شان، 2006]
Xu Y & Chen Z. Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology*, 57(7):961 – 973, 2006.

[لاركي و من معه، 2002]
Larkey L. S., Ballesteros L. & Connell M., Improving Stemming for Arabic Information Retrieval: Light Stemming and Cooccurrence Analysis. *The 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002)*, pp. 275-282, Tampere, Finlande, Août 2002.

[مقبل و من معه، 2001]
Mokbel C., Hanna G., Charles S. & Mikko K. Arabic Documents Indexing and Classification Based on Latent Semantic Analsyis and Self-Organizing Map. *Proceedings of the IEEE workshop on Natural Langage Processing in Arabic*, Beirut, Lebanon, 2001.

[نومان و رولكار، 2000]
Naumann F. & Rolker C. Assessment Methods for Information Quality Criteria. *International Conference on Information Quality (IQ)*, Cambridge, MA, 2000.

[هاجيك و من معه، 2005]
Hajic J., Otakar S., Buckwalter T. & Hubert J. Feature-Based Tagger of Approximations of Functional Arabic Morphology. *The Fourth Workshop on Treebanks and Linguistic Theories*, pp. 53-64 Universitat de Barcelona, Spain, December 2005.

[يوسفي و من معه، 2008]
Yousfi A., Aouragh H. & Allal J.. Modèle p-contexte de classe pour la génération automatique des phrases arabe. *International Conference on Web and Information Technologies (ICWIT '08)*, pp. 170-174, Sidi Bel Abbes, Algeria, 29 - 30 June 2008.

10. جدول الألفاظ

Web	الويب
Semantic Web	الويب الدلالي
Socio-semantic Web	الويب الاجتماعي الدلالي
Ontology	المكنز
Information reliability	اعتمادية المعلومة
Usage type	نوعية الاستخدام
Web search engine	محرك البحث
Explicit specification of a conceptualisation	توصيف علني لتصور
Knowledge representation	تمثيل معرفي
Document usage type	نوعية استخدام الوثيقة
Community of practice	جماعة ممارسة
Semi-structured documents	الوثائق شبه المنظمة
Hyper-document	وثيقة فائقة التداخل
Hyper-library	مكتبة فائقة التداخل
Level of granularity	مستوى الحبوبية
Terminological resources	الموارد المصطلحية
Semi-automatic	شبه آلية
Morphological analysis	التحليل الصرفي
Stemming	التشذيب

Light Stemming	التشذيب الخفيف
Hybrid method	الطرق الهجينة
Statistic method	طرق إحصائية
POS tagging	العنونة النحوية
Frequency	التردد
TFIDF : (Term Frequency, Inverse Document Frequency)	ت.ل.ع.ت.و(تردد اللفظة، عكس تردد الوثيقة)
Macro-transactions	معاملات كُلية
Micro-transactions	معاملات جزئية
Class	الفصيلة، الفئة
Expert interview	المقابلات مع الخبراء
Document logical structure	الهيكل المنطقي للوثيقة
Macro-logical structure	الهيكل المنطقي الكلي
Micro-logical structure	الهيكل المنطقي الجزئي
Context free grammars	قواعد اللغات الحرة
Qualitative indexing	الفهرسة النوعية
Le score LLR	القياس ل.ل.ر
Metamodel	نموذج عالي
Hypernym	المصطلح العام
Physical structure	الهيكل المادي
Prefixes	البوادئ
Suffixes	اللواحق
Semiotic production	إنتاج سيميائي
Cooccurrence	التساوق
Parameter	معلمة
Dynamic	تفاعلي

11. الخلاصة باللغة الإنجليزية

Shifting from the actual Web to the socio-semantic Web requires developing methodologies, approaches and tools for document representation and knowledge extraction. Considering the particularities of the Arabic language and research works in this field, introducing Arabic documents into the socio-semantic Web is still a difficult and challenging task. We present in this paper a meta-model for representing Arabic documents in a socio-semantic map. Besides, we propose an architecture allowing to analyse documents and organise their fragments according to this model. The proposed toolbox aim to study information reliability seen as a social task since it establishes confidence between information stakeholders. We extract documents' logical structure then we segment each document into coherent fragments. The goal is to present parts of documents in a way which helps users' practices. Thus, the logical structure have a great importance in the process of document analysis. Having as a goal to help build ontologies, we exploit the structure of documents for knowledge extraction. In this step, we consider that terms which appear in the main title of the document and in the titles of its parts reflect more the document's subject and sense. In addition, we exploit logical links between fragments to infer semantic relations

between terms. Merging all these elements in a unique architecture allowed us to build socio-semantic maps for the themes "marriage" and "drinks" from books of Arabic stories. The obtained results encourage us to continue working in this field by developing tools supporting search and navigation through the map.