CITOM : approche de construction incrémentale d'une Topic Map multilingue

Nebrasse Ellouze[1], Nadira Lammari[2], Elisabeth Métais[2], Mohamed Ben Ahmed[1]

[1] Ecole Nationale des Sciences de l'Informatique, Laboratoire RIADI
Université de la Manouba, 1010 La Manouba, Tunisie
{nebrasse.ellouze, mohamed.benahmed}@riadi.rnu.tn
[2] Laboratoire Cedric, CNAM
292 rue Saint Martin, 75141 Paris cedex 3, France
{metais, lammari}@cnam.fr

Résumé. Cet article décrit CITOM, une approche pour la Construction Incrémentale d'une TOpic Map Multilingue. Cette dernière sert à organiser un contenu multilingue composé de documents textuels. Elle a pour avantage de faciliter la recherche d'information dans le contenu. CITOM conjugue l'utilisation de trois sources d'information : (a) un ensemble de documents disponibles dans différentes langues, (b) un thésaurus du domaine sur lequel portent le contenu à organiser ainsi que (c) l'ensemble de toutes les sources d'interrogations possibles telles que les questions relatives aux documents sources qu'un expert du domaine ou un utilisateur quelconque peut poser, les foires aux questions (FAQ), etc.

Mots clés: Topic Map (TM), construction incrémentale, enrichissement, documents multilingues, thésaurus, requêtes des utilisateurs

1 Introduction

Les Topic Maps [1] conçues à l'origine comme un équivalent électronique d'index traditionnels, répondent à l'heure actuelle à un besoin d'organiser autour d'une vision métier des contenus et connaissances de différentes sources et de différentes langues. Grâce au réseau de liens sémantiques entre les sujets qu'elles représentent, elles permettent une navigation facile et sélective améliorant ainsi la recherche de l'information dans les contenus.

Un contenu à organiser étant très souvent volumineux et sujet à enrichissement perpétuel, il est pratiquement impossible d'envisager une création et gestion d'une Topic Map, le décrivant, de façon manuelle. Plusieurs travaux de recherche se sont penchés sur cette problématique. Beaucoup de propositions ont concerné la construction de Topic Maps à partir de documents textuels [2]. Cependant, aucune d'elles ne permet de traiter un contenu multilingue. De plus, bien que les Topic Maps soient, par définition, orientées utilisation (recherche d'information), peu d'entre elles prennent en compte les requêtes des utilisateurs.

Dans cet article, nous proposons CITOM, une approche évolutive et incrémentale d'élaboration d'une Topic Map multilingue. La construction d'une telle Topic Map offrira à l'utilisateur la possibilité de s'enrichir de connaissances se trouvant dans des documents écrits dans une langue autre que la sienne.

Outre les bases de documents disponibles dans différentes langues, CITOM conjugue l'utilisation de deux autres sources d'information : un thésaurus du domaine sur lequel portent le contenu à organiser ainsi que l'ensemble de toutes les sources d'interrogations possibles telles que les questions relatives aux documents sources qu'un expert du domaine ou un utilisateur quelconque peut poser, les foires aux questions (FAQ), les traces des discussions téléphoniques et des consultations directes avec les travailleurs du domaine.

CITOM vise à produire une Topic Map globale permettant de structurer sémantiquement des concepts dans plusieurs langues tout en prenant en charge une des spécificités du multilinguisme qui est l'absence éventuelle de termes sémantiquement équivalents d'une langue à une autre ; ce qui est assez fréquent lorsque les contenus sont issus de différentes cultures.

Dans CITOM, il est proposé de munir les topics de méta propriétés qui sont initialisées à la création de la Topic Map. Ces méta propriétés nous renseignent sur l'importance des topics et sur l'usage qu'on en fait lors de l'exploitation de la Topic Map. Ils sont utilisés pour la gestion des évolutions de la Topic Map, plus précisément pour l'élagage de topics considérés non pertinents.

Le reste du papier est organisé comme suit. La section 2 présente les caractéristiques du modèle de Topic Map utilisé. La section 3 détaille les différentes étapes de notre approche. La section 4 est dédiée à l'élagage de la Topic Map générée. La section 5 illustre notre approche à travers un exemple d'application dans le domaine de la construction durable. Enfin, la section 6 conclut ce travail de recherche et présente nos perspectives.

2 Caractéristiques du modèle de Topic Map

Le modèle Topic Map a été formalisé en norme ISO 13250. Il dispose d'un langage de spécification XTM définie par le consortium « Topic Maps.Org ». Il a été aussi inclus dans l'ODM (*Ontology Meta-Model Definition*) par l'OMG dans l'objectif de fournir un modèle standard TM-UML favorisant l'applicabilité des concepts du MDA (*Model Driven Architecture*) à l'ingénierie des Topic Maps. La figure 1 montre un extrait du méta modèle des Topic Maps proposé dans l'ODM.

Une Topic Map est une structure abstraite permettant de représenter les connaissances qu'on peut avoir sur des ressources (documents, bases de données, vidéos, etc.). Elle est organisée autour de topics représentant des sujets que le créateur de la Topic Map souhaite décrire et pour lesquels des ressources sont disponibles pour fournir de la connaissance sur ces sujets. Comme le montre la figure 1, un topic peut avoir un nom de base (*base name*) et des variantes (*variant names*). Tout topic a une ou plusieurs occurrences qui regroupent toutes les informations permettant d'accéder aux différentes ressources concernées par ce topic. Il peut être lié à un ou plusieurs

topics via des liens sémantiques appelés « associations ». Le rôle joué par ce topic dans une association constitue une de ses caractéristiques.

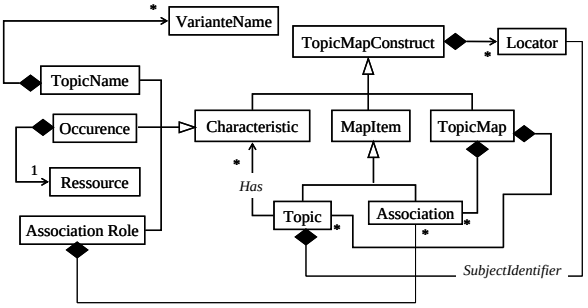

Fig. 1. Extrait du méta modèle des Topic Maps selon la spécification de l'OMG.

L'existence des liens sémantiques entre topics permet de naviguer dans la Topic Map et rend possible l'interconnexion des ressources via les topics qui les représentent. Tel qu'il est mentionné dans [3], il n'y a pas de limitation dans la définitions de ces liens. Ils sont spécifiés par le créateur de la Topic Map selon les besoins en information, les connaissances véhiculées par la Topic Map ainsi que l'application à laquelle elle est destinée.

L'utilisateur peut ainsi naviguer entre les topics en suivant des liens sémantiques. La Topic Maps créant des chemins sémiotiques fondés sur l'utilisation, l'utilisateur aura un service ajouté par rapport à une recherche non guidée dans un thésaurus ou par mots clés.

La classification de liens sémantiques entre concepts ou termes a fait l'objet de plusieurs résultats de recherche. A titre d'exemple, on peut citer la classification proposée par le ANSI (*American National Standards Institute*) dans le ANSI/NISO Z39.19-2005. Cette dernière considère trois catégories de liens : (1) les liens « d'équivalence » telles que la synonymie et la quasi-synonymie, (2) les liens hiérarchiques telles que la spécialisation et la généralisation et (3) les autres liens (nommés «associatifs ») tels que le lien « cause/effet » et le lien « action/cible ». Comme autre exemple de classification, on peut aussi citer l'ontologie proposée dans [4] pour la catégorisation des relations représentées par les verbes dans les phrases d'un texte.

Dans l'approche CITOM décrite dans la section qui suit , nous produisons, en plus des liens reliant un topic à ses ressources (les occurrences), deux autres catégories de liens : (a) les liens ontologiques et structurels et (b) les liens d'usage. Les liens ontologiques et structurels regroupent les liens de spécialisation, le lien de composition ainsi que les liens associatifs, tels que ceux définis dans le standard (ANSI/NISO Z39.19-2005), que nous pourrions identifier suite à l'analyse des documents à organiser. Pour cela, nous nous appuyons sur les outils linguistiques existants. Le lien d'usage est un hyper lien de type « répond à » (hyper lien questions/réponses) entre la question représentée comme un topic et les réponses associées, c'est-à-dire les topics référençant les documents qui permettent de répondre à la question. Nous prévoyons dans ce contexte de relier la question à chacun des mots clés la constituant via un hyper lien de type « est composé de ».

Le standard des Topic Maps dispose aussi du concept de scope (contexte) ou domaine de validité et du concept de facette. Le scope indique dans quel contexte tel topic aura tel nom, telles occurrences et tels rôles. La facette permet de compléter les informations à propos d'une occurrence en ajoutant des informations de type attributs-valeurs dans le composant occurrence qui référence le document concerné. Nous comptons exploiter ces deux concepts pour la prise en charge du multilinguisme dans l'élaboration d'une Topic Map. En effet, nous proposons de définir un scope pour chaque langue traitée dans la Topic Map avec la possibilité bien sûr d'attribuer à un topic une liste de noms dans différentes langues. Nous comptons aussi exploiter le concept de facette dans un objectif de filtrage des documents selon leurs langues. Pour ce faire, nous définissons un attribut « langue » dans le composant occurrence. La valeur prise par cet attribut dans une occurrence donnée correspondra à la langue du document référencé par cette occurrence.

Enfin, nous proposons dans le cadre de ce travail de recherche d'étendre le modèle TM-UML en rajoutant aux caractéristiques d'un topic des méta propriétés servant à mesurer la pertinence des topics dans le temps. Ces méta propriétés sont initialisées à la création de la Topic Map. Ils nous renseignent sur l'importance des topics et sur l'usage qu'on en fait lors de l'exploitation de la Topic Map. Ils sont utilisés pour la gestion des évolutions de la Topic Map, plus précisément pour l'élagage de topics considérés non pertinents.

3 Notre approche

Il existe dans la littérature plusieurs travaux sur la construction de Topic Maps [2]. Ils se distinguent principalement par les sources et la nature des techniques utilisées pour la production de Topic Maps. A titre d'exemple, dans les travaux de Reynolds et Librelotto [5][6], les auteurs proposent un processus automatisé exploitant en entrée des documents XML. Pepper [7][8] s'est intéressé à la traduction de méta-données RDF en Topic Map. Des techniques d'apprentissage et de traitement du langage naturel [9] [10] ont été appliquées pour l'élaboration de Topic Map à partir de documents textes. Notons aussi, que certains travaux comme ceux présentés dans [11][12][13][14] décrivent des processus de construction collaboratives de Topic Maps. Cependant, à notre connaissance, aucune des approches existantes, exceptée celle de Kasler [15] qui utilise à la fois des documents écrits en anglais et en hongrois, ne permet l'obtention d'une Topic Map à partir d'un contenu multilingue. Aussi aucune des approches existantes n'exploitent plusieurs sources d'informations pour la construction d'une Topic Map.

L'approche CITOM, que nous décrivons dans cet article, conjugue l'utilisation de trois sources d'information : un thésaurus du domaine sur lequel portent le contenu à organiser, l'ensemble de toutes les sources d'interrogations possibles telles que les questions relatives aux documents sources qu'un expert du domaine ou un utilisateur quelconque peut poser, les foires aux questions (FAQ), les traces des discussions téléphoniques et des consultations directes avec les travailleurs du domaine ainsi que les documents disponibles.

Intuitivement, il s'agit, dans CITOM, de construire de façon incrémentale une Topic Map TM_i correspondante à un ensemble de documents D= $\{d_1, d_2, \ldots, d_i\}$ en fusionnant la Topic Map TM_{i-1} correspondante à l'ensemble de document D-$\{d_i\}$ avec la Topic Map associé au document d_i. Chaque phase permettant de construire la Topic Map correspondante à un document d_i utilise comme source aussi bien le document lui-même mais aussi le thesaurus et, un ensemble de questions correspondantes à ce document et extraites de sources d'interrogations.

CITOM exploite des documents de différentes langues. Elle permet ainsi de produire une Topic Map où les concepts sont décrits dans plusieurs langues tout en prenant en charge une des spécificités du multilinguisme qui est l'absence éventuelle de termes sémantiquement équivalents d'une langue à une autre ; ce qui est assez fréquent lorsque les contenus sont issus de différentes cultures.

CITOM trouve aussi son utilité toutes les fois où le contenu à organiser est enrichi de un ou plusieurs documents, ou encore lorsque l'on souhaite introduire d'autres questions fréquemment posées. Elle contribue dans le processus d'évolution d'une Topic Map.

L'algorithme général de CITOM est le suivant :

***Entrée**: un ensemble de documents multilingues, un thésaurus et sources d'interrogations (questions des experts, requêtes des utilisateurs, FAQ, historique, discussions téléphoniques, consultations directes avec les travailleurs du domaine, ...)*

***Sortie**: une Topic Map globale*

Action 1. *Construire la racine de la Topic Map globale. La racine correspond au topic portant le nom du domaine sous les différentes langues.*

Action 2. *Traiter les sources d'interrogations et constituer, pour chaque document, un ensemble de questions potentielles.*

***Pour** chaque document i du contenu à organiser **faire** :*

Action 3. *Extraire les topics et les associations entre ces topics à partir du document i.*

Action 4. *Enrichir la Topic Map en ajoutant d'autres liens ontologiques et structurels à partir du thésaurus*

Action 5. *Enrichir la Topic Map à partir des questions potentielles correspondantes à ce document*

Action 6. *Validation de la Topic Map résultante par les experts*

Action 7. *Intégrer la Topic Map du document i dans la Topic Map globale.*

Fin pour

La validation de la Topic Map résultante de la phase 4 consiste à préciser la sémantique de certains liens ou encore à supprimer ou à ajouter des liens et/ou des topics. Elle nécessite la collaboration d'un ou plusieurs experts du domaine.

Compte tenu de l'espace imparti, nous nous limitons dans ce papier à la description, dans les paragraphes qui suivent, des phases 3, 4 et 5.

3.1 Extraction de topics et d'associations à partir d'un document

L'objectif de cette phase est d'extraire des topics et des associations entre ces topics à partir d'un document. Comme mentionné dans la section 3, les associations que nous souhaitons extraire correspondent aux liens ontologiques et structurels (lien « est-un », lien « partie-de ») et aux liens sémantiques associés au domaine de la

Topic Map. Pour cela, nous nous appuyons sur les techniques d'analyse linguistique de documents textuels pour l'extraction de concept et de relations entre concepts.

Dans la littérature, diverses méthodes ont été proposées pour résoudre ce problème. La majorité d'entre elles ont été développées dans le cadre de la création et de l'enrichissement d'ontologies à partir de données textuelles. Elles se distingue par le type de technique utilisée : technique statistique, technique syntaxique ou encore techniques de fouille de données. La plupart d'entre elles sont outillées.

Dans les méthodes basées sur les techniques statistiques, des mesures sont utilisées afin de sélectionner les concepts candidats. Parmi elles, on peut citer le nombre d'apparitions d'un terme au sein d'un corpus [16][17][18], l'information mutuelle, le tf-idf, le T-test ou encore les lois de distributions statistiques des termes [19][20][21]. Notons que toutes ces techniques ne permettent pas d'extraire des relations entre concepts.

Les méthodes basées sur l'analyse syntaxique, quant à elles, utilisent les fonctions grammaticales d'un mot ou d'un groupe de mots au sein d'une phrase. Certaines d'entre elles posent l'hypothèse que les dépendances grammaticales reflètent des dépendances sémantiques [22][23]. D'autres utilisent des patrons syntaxiques [24][25][26]. Ces méthodes ont l'avantage d'extraire aussi bien des concepts mais aussi les relations entre concepts. Toutefois, ces relations ne sont pas toujours étiquetées sémantiquement. Une intervention humaine est donc nécessaire pour nommer ces relations.

La dernière catégorie de méthodes exploite des techniques de fouille de données. A titre d'exemple, Han [27] et Neshatian [21] exploitent une ontologie et utilisent une technique de classification permettant de rapprocher des concepts candidats (contenus dans des documents) de concepts présents dans l'ontologie. Le principe est similaire dans les approches d'Agirre [16] et Parekh [18] qui regroupent, par une technique de clustering, des termes en fonction de leur nombre d'occurrences au sein du corpus. Chaque cluster représente alors la possibilité qu'une relation existe entre les concepts qu'il regroupe. D'autres méthodes proposent d'utiliser les corrélations fréquentes pouvant exister entre les termes d'un corpus. Ces approches consistent à extraire des règles d'association [28] entre des termes candidats [22][25][26]. A l'issu du processus, un ensemble de règles d'association est dérivé. Chacune des règles décrit l'existence d'une relation entre deux concepts. Un processus d'étiquetage manuel est par la suite exécuté afin de nommer les relations produites.

Il existe plusieurs outils d'extraction de concepts et/ou de relations entre concepts à partir des documents textuels. Parmi ces outils, on peut citer Nomino [29], Lexter [30], Fastr [31], Mantex [32], Likes [33], Acabit [34], Syntex [35], OntoGen [36] et Text2Onto [37]. A titre d'exemple, l'outil Syntex est un analyseur de texte. Il sert à identifier les dépendances syntaxiques entre concepts. Text2Onto est un outil conçu pour construire des ontologies à partir de textes de manière complètement automatique. Il est composé de modules qui extraient à partir des textes des concepts, des relations entre ces concepts (relation d'équivalence, hiérarchiques, etc.) et des instances de concepts. Il repose sur l'architecture GATE [38] pour prétraiter les textes. Les résultats sont dotés d'une mesure de confiance entre 0 et 1 obtenue à l'aide de différentes mesures combinables (TF.IDF, RTF, entropie).

Enfin, plusieurs architectures d'ingénierie du texte ont été développées pour les traitements linguistiques [39][40]. Nous citons par exemple, l'architecture GATE

(*General Architecture for Text Engineering*) qui vise généralement l'annotation linguistique et l'exploration de corpus pour l'extraction d'information.

Dans le cadre de notre travail, notre choix s'est porté sur la plate-forme GATE qui a l'avantage de proposer une solution générique pour le traitement linguistique des documents textuels à travers un ensemble de modules paramétrables. Ces modules peuvent être combinés, enrichis et adaptés selon nos besoins. GATE propose un module appelé *gazetter*, pour la reconnaissance d'entités nommées à partir de dictionnaires pré-établis. Ces dictionnaires peuvent être enrichis avec les termes du domaine étudié. De plus, GATE nous donne la possibilité d'intégrer une ressource externe représentée par le thésaurus du domaine pour construire la hiérarchie de topics et ajouter d'autres liens ontologiques dans la Topic Map.

3.2 Enrichissement de la Topic Map par d'autres liens ontologiques

L'objectif de cette étape est d'enrichir les topics issus des documents en leur ajoutant d'autres liens ontologiques et structurels. Pour cela, nous proposons d'utiliser les liens entre les termes qui existent dans le thésaurus. Rappelons qu'un thésaurus est fondé sur une structuration hiérarchisée d'un ou plusieurs domaines de la connaissance et dans lequel les notions sont représentées par des termes d'une ou plusieurs langues naturelles et les relations entre notions par des signes conventionnels (AFNOR 1987). Les normes (ISO 2788 et ANSI Z39) ont permis d'uniformiser leur contenu en termes de relations entre unités lexicales : équivalence, relations hiérarchiques et relations non taxonomiques (liens associatifs).

Plusieurs travaux de recherche décrivent des approches de construction d'ontologies à partir de thésaurus. Parmi elles, on peut citer celle d'Hernandez [41] qui propose de transformer un thésaurus en une ontologie de domaine. Les auteurs définissent une méthode qui permet d'extraire les éléments du schéma conceptuel de l'ontologie à partir d'un thésaurus et de documents textuels. Leur approche est fondée sur un ensemble de règles de transformation. Ces règles exploitent les liens «est plus spécifique que», «est plus générique que», « Utiliser plutôt » (UP) et « Utiliser pour désigner » (UPD) d'un thésaurus pour générer les concepts de l'ontologie, les labels associés à chacun de ces concepts et la hiérarchie de concepts.

Notre approche d'enrichissement de la Topic Map par d'autres liens ontologiques est proche de celle proposée dans [41] dans le sens où nous exploitons aussi les liens existant dans un thésaurus pour produire des liens ontologiques. En effet, les topics sont organisés hiérarchiquement à partir de la relation « est-un ». Ces liens hiérarchiques entre topics sont directement issus des liens «est plus spécifique que» (EPS) et «est plus générique que» (EPG) explicitement présents dans le thésaurus. Nous utilisons aussi les relations de type « Utiliser plutôt » (UP) et « Utiliser pour désigner » (UPD) pour ajouter un nouveau nom au topic ou encore regrouper des topics en un seul topic. Cependant la démarche que nous utilisons est algorithmique. Elle s'exécute en deux temps. Dans un premier temps nous exploitons les liens UP et UPD entre termes du thésaurus pour regrouper des topics. Dans un second temps, nous organisons les topics en hiérarchies.

Soit $SYN(Terme_i)$ l'ensemble constitué du $Terme_i$ et de tous les termes auquel est lié $Terme_i$ via le lien UP ou UPD dans le thésaurus. Notons par TP_i le terme préféré

dans SYN(Terme$_i$). L'algorithme permettant de regrouper des topics ou d'attribuer plusieurs noms à un topic est le suivant :

Pour tout Topic T_i faire
Si T_i est dans le thésaurus
Alors construire SYN(T_i), TP_i et FILS(TP_i)
T_i aura pour nom de base TP_i et pour autres noms les noms se trouvant dans SYN(T_i)
Fin Pour
Pour tout couple T1 et T2 de topic faire
Si SYN(T1) = SYN(T2)
Alors regrouper T1 et T2 en T3 tel que le nom de base de T3 soit TP_1, et les autres noms soient ceux faisant partie de SYN(T1). Ce regroupement implique bien sûr le regroupement de toutes les autres caractéristiques de T1 et T2 (Rôles et occurrences).
Fin Pour

Pour l'organisation des topics, nous utilisons deux techniques que nous avions proposées dans [42], en les adaptant à notre contexte pour la construction et la maintenance d'ontologie. La première technique, nommé «translation», permet de traduire une hiérarchie de concepts en contraintes entre concepts. La seconde technique, nommée «normalisation», permet, compte tenu d'un ensemble de contraintes entre concepts, de déduire une hiérarchie de concepts.

Dans le contexte particulier d'enrichissement de la Topic Map, nous considérons deux types de contraintes : (a) la contrainte d'exclusion de sémantique, notée $\not\leftrightarrow$, qui définie entre deux concepts T1 et T2, exprime le fait que T1 et T2 n'ont pas la même sémantique et la contrainte d'inclusion de sémantique, notée $\rightarrow$, qui définie d'un concept T1 vers un concept T2 (T1 $\rightarrow$ T2) exprime le fait que la sémantique de T1 contient celle de T2 (l'inverse n'étant pas vrai).

Pour extraire les contraintes entre termes représentés en topics dans notre Topic Map, nous appliquons la technique de translation sur le thésaurus. Nous nous basons pour cela sur les liens «est plus générique que» explicitement présents dans le thésaurus. Cette technique s'appuie sur les trois règles de translation suivantes :

- *R1* : si, dans le thésaurus, un concept T2 est plus générique qu'un concept T1, alors (T1 $\rightarrow$ T2).
- *R2 :* si, dans le thésaurus, il n'existe pas un concept T3 comme nœud de départ de deux chemins l'un allant vers T1 et l'autre allant vers T2 tel que ces chemins soient constitués uniquement de lien «est plus générique que » alors T1 $\not\leftrightarrow$ T2
- *R3* : si, dans le thésaurus, deux concepts T1 et T2 sont tous les deux reliés à un concept T3 par le lien «est plus générique que» alors T1, T2$\rightarrow$T3

Soit *FILS(T)* l'ensemble de termes du thésaurus récoltés lors du parcours de tous les chemins constitués uniquement de liens EPG et ayant pour noeud de départ T, l'algorithme permettant d'extraire les contraintes entre topics à partir du thésaurus se résume en ce qui suit :

Pour tout topic T1 faire
Pour tout T ∈ FILS(T1)
T→T1
Fin Pour
Fin Pour
Pour tout couple T1 et T2 dans la Topic Map faire
Si FILS(T1)≠FILS(T2) alors T1 ↮ T2
Sinon si FILS(T1)∩FILS(T2) = {T3} alors T1, T2→T3
Fin Pour

Une fois les contraintes déduites, nous appliquons sur notre Topic Map la technique de normalisation afin d'organiser sous forme d'hiérarchie l'ensemble des concepts de notre Topic Map. Intuitivement, il s'agit :

- dans un premier temps, de construire un graphe complet non orienté ayant pour nœuds l'ensemble des topics,
- dans un second temps, d'éliminer tout lien entre deux topics T1 et T2 si T1 ↮ T2,
- dans un troisième temps, de déduire toutes les cliques[1] possibles respectant l'ensemble des contraintes d'inclusion et d'organiser ces cliques en un graphe d'inclusion,
- et enfin d'inverser les liens d'inclusion pour obtenir les liens d'héritages entre topics et d'éliminer les redondances en supprimant tout concept d'une sous-classe se trouvant dans sa super-classe.

La figure 2 montre un exemple de déroulement de l'algorithme de normalisation:

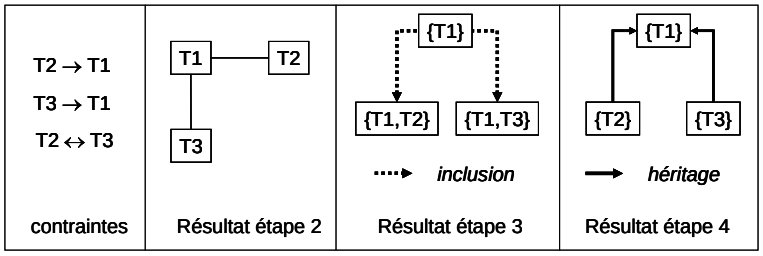

Fig. 2. Exemple de déroulement de l'algorithme de normalisation.

3.3 Enrichissement de la Topic Map à partir des requêtes

L'intérêt principal de la Topic Map globale est de guider l'utilisateur dans ses interrogations. Elle doit lui permettre de naviguer au sein de sa structure et d'accéder aux documents associés aux topics explorés.

L'objectif de cette étape est d'introduire dans la Topic Map des connaissances sur les questions les plus fréquemment posées sur les documents par le biais de liens d'usage. Cela suppose bien sûr que les questions les plus fréquemment posées ont été

[1] Une clique est un sous graphe complet

préalablement recensées et triés par document. Le recensement de ces questions potentielles est fait suite à l'analyse de toutes les sources d'interrogations disponibles (questions relatives aux documents sources qu'un expert du domaine peut poser, les foires aux questions, les journaux de reprise des serveurs de données, les traces des discussions téléphoniques et des consultations directes avec les travailleurs du domaine, etc.). Cette analyse s'effectue à la phase 2 de notre processus.

Toute question recensée est donc représentée en un topic que l'on relie via le lien « répond à » à ses réponses c'est-à-dire les topics référençant les documents qui permettent de répondre à cette question. Cette même question est aussi reliée à chacun des mots clés la constituant via un hyper lien de type « est composé de ». Le stockage des liens « est-composé-de » d'une question (i.e. phrase en langage naturel) vers les termes la composant permet d'une part une recherche par navigation et d'autre part une recherche automatique de "question proche" en reconstituant le vecteur de Salton [43] correspondant à la question, grâce à l'ensemble des termes composants.

L'insertion des réponses à une question suppose que le processus d'enrichissement est capable de sélectionner les mots clés d'une question, de retrouver les documents réponses associés à une question et enfin de rechercher dans ces documents les mots clés les représentant. Pour ce faire, nous utilisons les techniques linguistiques existantes dans la littérature [16][22][25].

L'insertion d'un topic obtenu par enrichissement peut se traduire soit par une action vide dans le cas où le topic existe tel quel dans la Topic Map, soit par le rajout d'un nom à un topic existant dans le cas où le topic existe sous un autre nom soit par l'insertion effective de ce topic dans le cas où il n'existe pas sous aucune des formes citées. Pour l'étude de ces possibilités, nous nous appuyons sur les techniques de fusions existantes [44][45][46].

4 Elagage évolutif de la Topic Map

Un problème important à traiter est celui de l'élagage de la Topic Map. Une Topic Map étant utilisée essentiellement pour l'organisation d'un contenu et pour la recherche d'information dans ce contenu, il est souhaitable de réviser sa structure de façon périodique afin qu'elle puisse répondre de façon efficace aux besoins de recherche d'information et qu'elle puisse évoluer en accord avec les évolutions effectuées sur le contenu.

Pour faciliter la gestion des évolutions d'une Topic Map, il nous a paru judicieux de collecter des connaissances qui nous renseignent sur l'importance des topics et sur l'usage qu'on en fait lors de l'exploitation de la Topic Map. Elles peuvent servir dans l'évaluation de la qualité d'une Topic Map. Nous les matérialisons, comme énoncé dans la section 2 de ce papier, par des méta propriétés.

Dans un premier temps, nous proposons comme méta propriété la note d'un topic qui nous renseigne sur son importance dans la Topic Map. Elle est initialisée dès la création de la Topic Map. Elle peut être (a) très bonne dans le cas où le topic concerné est obtenu à partir des trois sources (les documents, le thésaurus et les requêtes) utilisées pour élaborer la Topic Map globale, (b) moyenne dans le cas où le topic est issu de deux sources ou encore (c) moins bonne s'il a été extrait d'une seule source.

L'attribution d'une note à un topic pourrait aussi servir comme critère de choix de visualisation. En effet, un topic ayant une note très bonne pourrait être considéré comme topic principal et de part ce fait apparaître dans une visualisation par défaut de la Topic Map.

Nous projetons, dans nos prochains travaux de faire une étude permettant de réunir les critères de qualité d'une Topic Map afin de déterminer de façon plus ou moins exhaustive la liste de méta propriétés utiles à la gestion des évolutions d'une Topic Map.

5 Application au domaine de la construction durable

Nous expérimentons CITOM sur un corpus du domaine de la construction durable. Ce dernier est composé de 105 documents html écrits en français et en anglais. Nous utilisons à cet effet le thésaurus bilingue (français/anglais) CTCS [47] (*Canadian thesaurus of construction science and technology*) développé par l'université de Montréal. Ce thésaurus comporte actuellement 15331 termes répartis sur 10 niveaux de hiérarchies. Chacun de ces termes est décrit dans un fichier html. Les relations entre termes sont des hyperliens dans ce fichier html.

La figure 3 montre un exemple de Topic Map construit à partir des trois ressources (documents, thésaurus et requêtes) selon les phases 3, 4 et 5 de notre processus d'élaboration de Topic Map. Cet exemple permet de monter la partie de la Topic Map qui vient du text mining et qui ne se trouve pas déjà dans le thesaurus et vice versa.

Fig. 3. Exemple d'élaboration d'une Topic Map à partir d'un document du domaine de la construction durable.

Pour illustrer les phases 3 à 5 présentées dans ce papier, considérons un extrait d'un document D1 en français («http://www.climamaison.com/climatisation-chauffage-ventilation.php») et un autre extrait issu d'un document D2 en anglais (http://www.aceee.org/). Ces deux documents portent sur les solutions pour l'économie d'énergie. En utilisant l'outil GATE dans lequel nous avons intégré quelques fichiers du thésaurus CTCS, nous avons identifié à partir des deux extraits 47 topics et 16 associations entre ces topics. Cette première Topic Map contient des topics ayant un nom en français et un autre en anglais s'il existe un et vice versa. Pour cela, lors de la création de la Topic Map, nous avons défini deux contextes (scope) selon la langue, un contexte « français » et un autre « anglais » (voir figures 4 et 5). Nous avons utilisé l'outil TM4J [48] pour la visualisation de la Topic Map multilingue générée. Comme le montre la figure 6, les topics ayant un nom dans les deux langues traitées, au niveau de l'interface de présentation de la Topic Map, ils seront affichés avec les deux noms en français et en anglais. Nous avons, ensuite,

enrichi cette première Topic Map moyennant le thésaurus CTCS. De cet enrichissement, d'autres liens ontologiques ont été ajoutés. A titre d'exemple, nous pouvons citer la relation hiérarchique « est-un » entre « chauffage » et « chauffage écologique ».

Nous avons ensuite extrait la question (exprimée en français) « Quels sont les moyens de chauffage économique ? » à partir d'une liste de la FAQ du site «http://www.climamaison.fr/» et comme le montre la figure 7, nous l'avons intégré dans notre Topic Map. Suite aux deux phases d'enrichissement 4 et 5, si un topic ou un lien (issu du thesaurus ou des questions) existe déjà dans la Topic Map sous un nom français alors nous rajoutons à ce topic un nouveau nom en anglais (s'il existe un équivalent bien sur) et de même pour l'anglais. Dans le cas où le topic identifié n'existe pas alors il est inséré dans la Topic Map avec un nom dans chaque langue. Nous produisons ainsi une Topic Map où les concepts sont décrits dans plusieurs langues en prenant en compte une des spécificités du multilinguisme qui est l'absence éventuelle de termes sémantiquement équivalents d'une langue à une autre ; ce qui est assez fréquent lorsque les contenus sont issus de différentes cultures.

De plus, lors des différentes phases de création de la Topic Map, chaque topic est relié aux ressources (URL) concernées par ce topic. Selon la langue, nous attribuons chaque ressource au thème correspondant (français ou anglais).

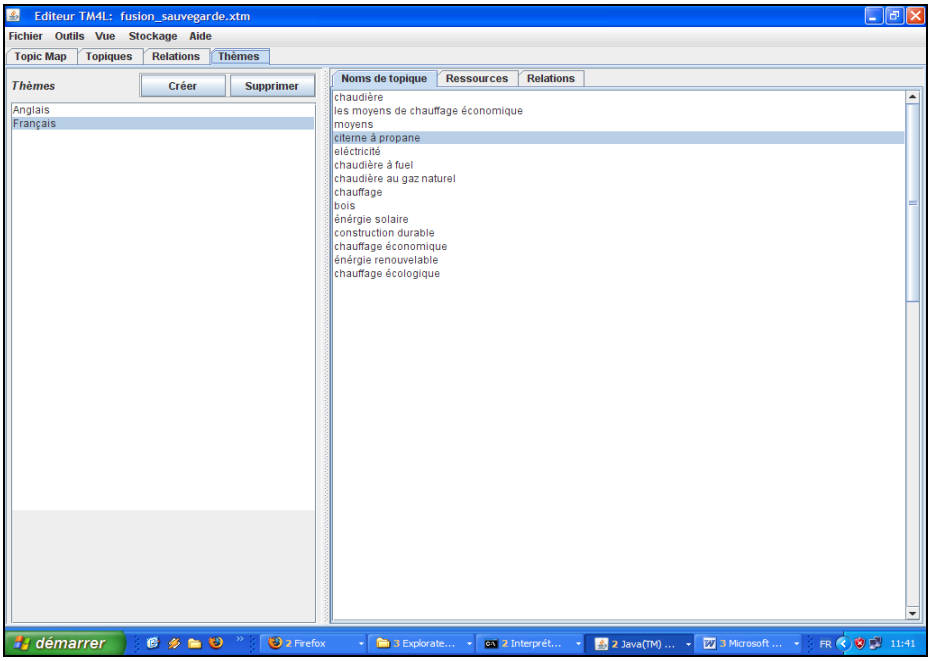

Fig 4. Etape de construction de la Topic Map multilingue : Liste des noms de topics dans le contexte (thème) français.

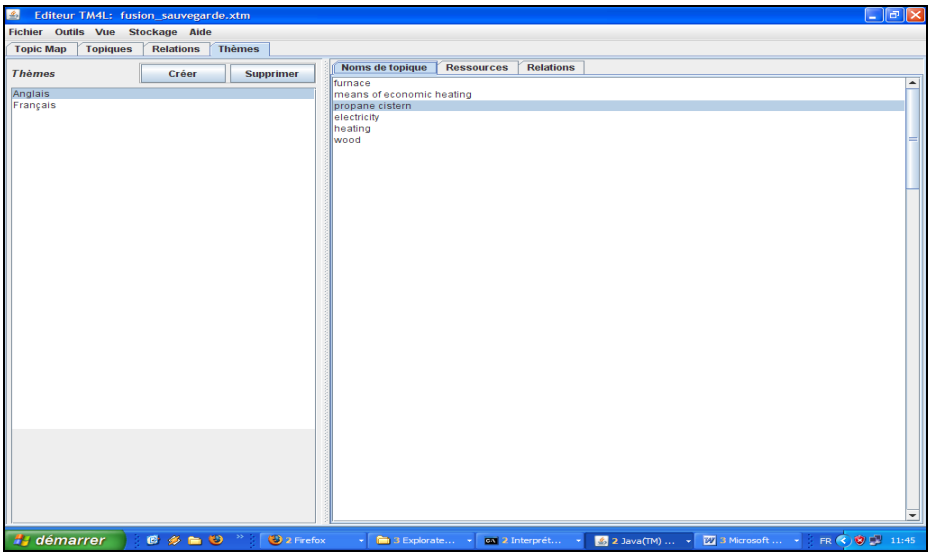

Fig 5. Etape de construction de la Topic Map multilingue : Liste des noms de topics dans le contexte (thème) Anglais.

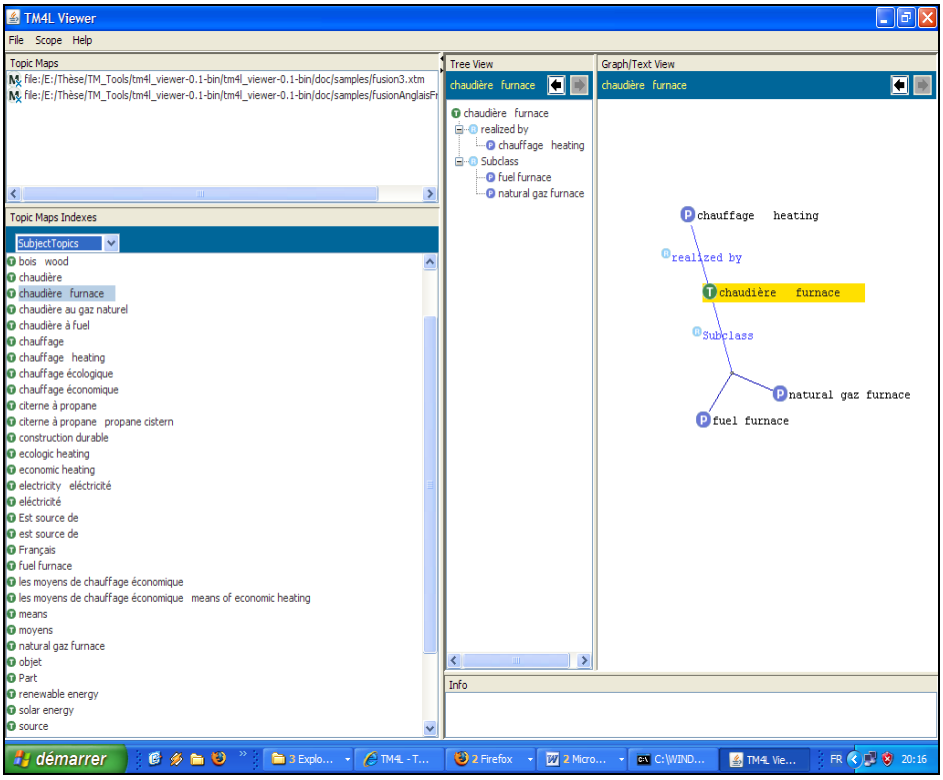

Fig. 6. La Topic Map multilingue après application des phases 3 à 5.

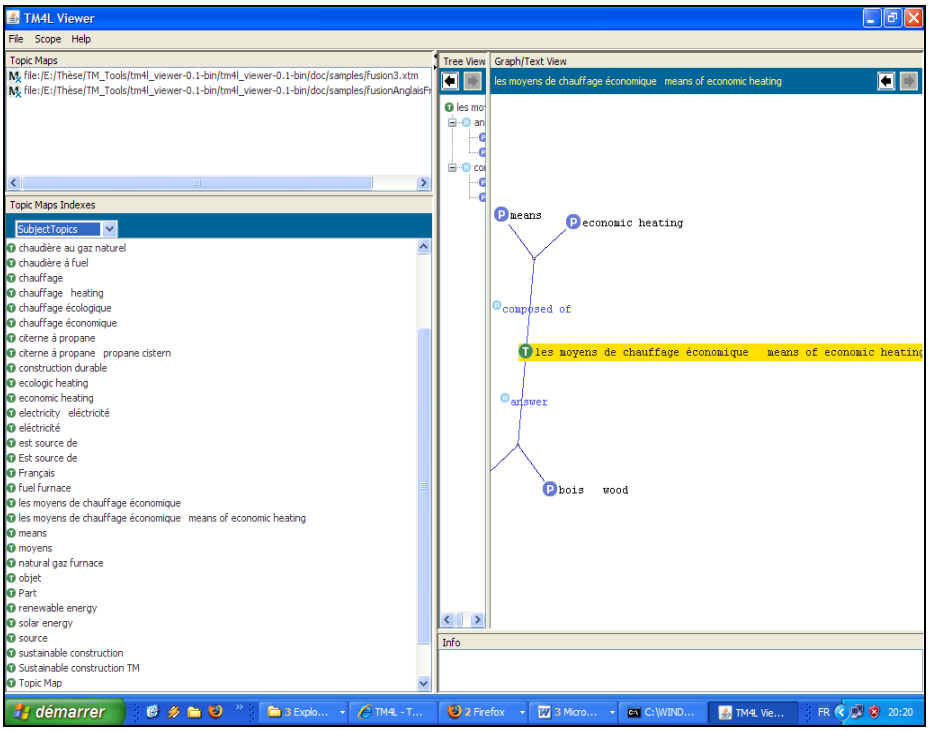

Fig. 7. La Topic Map multilingue après application des phases 3 à 5 : Intégration des topics issus des requêtes.

6 Conclusion et perspectives

Dans cet article, nous avons présenté CITOM, une approche incrémentale d'élaboration d'une Topic Map multilingue. L'approche a été expérimentée sur un corpus de document concernant la construction durable.

Notre approche a pour première originalité de prendre en compte l'usage de la Topic Map à travers la mise en œuvre de liens d'usage entre les questions potentielles extraites des sources d'interrogations disponibles et les réponses associées. Toute question potentielle (i.e. phrase en langage naturel) représentée sous forme de topic est aussi reliée à chacun des mots clés la constituant via un hyper lien de type « est composé de ». Le stockage des liens « est-composé-de » d'une question vers les termes la composant permet d'une part une recherche par navigation et d'autre part une recherche automatique de "question proche".

CITOM a aussi l'avantage de prendre en compte le multilinguisme des ressources qu'elle organise. Ainsi, un utilisateur pourra, lors de sa navigation, avoir accès à des documents qui ne sont pas dans sa langue d'origine. Le grand intérêt de cette approche par rapport à de simples traductions de réponses est de proposer à

l'utilisateur des documents correspondant à des concepts n'existant pas forcement dans sa langue ou dans sa culture. Ceci constitue à notre avis un enrichissement culturel.

De plus, CITOM permet l'instanciation de méta propriétés que nous avons intégré au modèle de Topic Map. Ces méta propriétés sont utiles pour la gestion des évolutions d'une Topic Map.

Enfin, de part son processus incrémentale, CITOM est réutilisable à chaque enrichissement du contenu qu'elle organise.

Nos futurs travaux porteront, à court terme, principalement sur l'étape de validation. A cet effet, nous comptons proposer une approche faisant collaborer plusieurs experts. Nous souhaitons également enrichir le processus d'élagage de la Topic Map, en intégrant, plus de critères (méta propriétés) permettant de juger de la pertinence d'un topic. Nous étudierons aussi la possibilité de leur généralisation aux autres concepts tels que le concept d'association.

Bibliographie

1. ISO/IEC :13250. Topic Maps: Information technology-document description and markup languages (2000): http ://www.y12.doe.gov/sgml/sc34/document/0129.pdf.
2. Ellouze N., Métais E., Ben Ahmed M., State of the Art on Topic Maps Building Approaches, In R.-D. Kutsche and N. Milanovic (Eds.): MBSDI 2008, Model Based Software and Integration Systems, CCIS 8, pp. 102–112, © Springer-Verlag Berlin Heidelberg, (2008)
3. Pepper S., Article for the Encyclopedia of Library and Information Sciences, http://www.ontopedia.net/pepper/papers/ELIS-TopicMaps.pdf, (2008).
4. Veda Storey C., Sandeep P., Understanding Relationships: Classifying Verb Phrase Semantics, Conceptual Modeling – ER 2004, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Volume 3288/2004, ISSN 0302-9743 (Print) 1611-3349 (Online), pp. 336-347, (2004).
5. Reynolds J., Kimber W.E. Topic Map Authoring With Reusable Ontologies and Automated Knowledge Mining. XML 2002 Proceedings by deepX (2002).
6. Librelotto G.R. Ramalho J. C., Henriques P. R.,. TM-Builder: An Ontology Builder based on XML Topic Maps. Clei electronic journal, vol. 7, no. 2, paper 4, (2004).
7. Pepper S. Topic Map Erotica RDF and Topic Maps "in flagrante" (2002): http://www.ontopia.net/topicmaps/materials/MapMaker_files/frame.htm
8. Pepper S. Methods for the Automatic Construction of Topic Maps (2002): http://www.ontopia.net/topicmaps/materials/autogen-pres.pdf
9. LeGrand B., Soto M. Topic Maps et navigation intelligente sur le Web Sémantique, AS CNRS Web Sémantique, CNRS Ivry-sur-Seine - October, (2002).
10. Folch H., Habert H. Articulating conceptual spaces using the Topic Map standard. Proceedings XML'2002, Baltimore, december (2002), 8-13.
11. Ahmed, K., "TMShare – Topic Map Fragment Exchange in a Peer-To-Peer Application", (2003):http://www. idealliance.org/papers/dx_xmle03/papers/02-03-03/02-03-03.pdf, 2003.
12. Lavik, S., Nordeng, T. W., Meloy, J. R. BrainBank Learning - building personal topic maps as a strategy for learning, In XML (2004) Washington.
13. Zaher L'H., Cahier J-P., Zacklad M. The Agoræ / Hypertopic approach. International Workshop IKHS - Indexing and Knowledge in Human Sciences, SdC (2006), Nantes.
14. Dicheva, D. & Dichev, C.. TM4L: Creating and Browsing Educational Topic Maps, British Journal of Educational Technology - BJET, 37(3), 391-404, (2006).

15. Kasler L., Venczel Z., Varga L.Z. Framework for Semi Automatically Generating Topic Maps. TIR-06. Proceedings of the 3rd international workshop on text-based information retrieval. Riva del Grada, 24-30, (2006).
16. Agirre E., Ansa O., Hovy E., Martinez D., Enriching very large ontologies using the WWW, In ECAI 2000 workshop on Ontology Learning, Berlin, Germany, (2000).
17. Faatz A., Steinmetz R., Ontology enrichment with texts from the WWW, In the Semantic Web Mining Conference WS'02, (2002).
18. Parekh V., Gwo J-P.,Finin T., Mining Domain Specific Texts and Glossaries to Evaluate and Enrich Domain Ontologies , In International Conference of Information and Knowledge Engineering, (2004).
19. Velardi P., Missikof M., Fabriani P., Using text processing techniques to automatically enrich a domain ontology , In Proceedings of ACM- FOIS, (2001).
20. Xu F.,Kurz D., Piskorski J., Schmeier, S., A domain adaptive approach to automatic acquisition of domain relevant terms and their relations with bootstrapping , In the 3rd international conference on language resources and evaluation, (2002).
21. Neshatian K., Hejazi, M. R., Text categorization and classification in terms of multi-attribute concepts for enriching existing ontologys , In 2ndWorkshop on Information Technology and its Disciplines, p. 43–48, (2004).
22. Bendaoud R., Rouane Hacene M., Toussaint Y., Delecroix B., Napoli A., Construction d'une ontologie à partir d'un corpus de textes avec l'ACF , IC (2007).
23. Roux C., Proux D., Rechermann F., Julliard L., An ontology enrichment method for a pragmatic information extraction system gathering data on genetic interactions , In Proceedings of the ECAI2000 Workshop on Ontology Learning, OL (2000).
24. Hearst M.A., Automatic acquisition of hyponyms from large text corpora, Rapport technique S2K-92-09, 1992.
25. Maedche A., Staab S., Mining ontologies from text, Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management. volume 1937, Springer-Verlag, (2000).
26. Stumme G., Hotho A., Berendt B., Semantic web mining: State of the art and future directions. Web Semantics: Science, Services and Agents on the World Wide Web, 4(2):124–143, June (2006).
27. Han E-H., Karypis G., Centroid based document classification: Analysis and experimental results , In The 4th European Conference of Principles of Data Mining and Knowledge Discovery, pages 424–431, (2000).
28. Agrawal R., Srikant R., Mining generalized association rules, Future Generation Computer Systems, 13(2–3): 161–180, (1997).
29. Dumas L., Plante A., Plante P., ALN: Analyseur Linguistique de ALN, vers.1.0 . ATO, UQAM, (1997).
30. Bourigault D., LEXTER, a Natural Language Processing tool for terminology extraction, Proceedings of the 7th EURALEX International Congress, Goteborg,(1996).
31. Jacquemin C., Bourigault D., Term Extraction and Automatic Indexing , in Mitkov R. (ed), The Oxford Handbook of Computational Linguistics, Oxford University Press, pp. 599-615, (2003).
32. Frath P., Oueslati R., Rousselot F., Identification de relations sémantiques par repérage et analyse de cooccurrences de signes linguistiques, in Ingénierie des connaissances. Évolutions récentes et nouveaux défis. Eds. Jean Charlet, Manuel Zacklad, Gilles Kassel, Didier Bourigault, Eyrolles, Paris, pp 291-304, (2000).
33. Rousselot F., Frath P., Oueslati R., Extracting concepts and relations from Corpora. In Proceedings of the Workshop on Corpus-oriented Semantic Analysis, European Conference on Artificial Intelligence, ECAI 96, Budapest, (1996).
34. Daille B., Identification des adjectifs relationnels en corpus, in Actes de la Conférence de Traitement Automatique du Langage Naturel (TALN'99), Cargèse, (1999).

35. Bourigault D., Fabre C., Frérot C., Jacques M.-P., Ozdowska S., Syntex, analyseur syntaxique de corpus , in Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles, Dourdan, France, (2005).
36. Fortuna B., Grobelnik M., Mladenic D., Semi-automatic data driven ontology construction system . In Proceedings of the 9th International multiconference Information Society IS-2006, Ljubljana, Slovenia, (2006).
37. Cimiano P., Volker J., Text2onto - a framework for ontology learning and data-driven change discovery . In A. MONTOYO, R. MUNOZ & E. METAIS, Eds., Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB), volume 3513 of Lecture Notes in Computer Science, p. 227–238, Alicante, Spain, Springer, (2005).
38. The GATE platform: http://gate.ac.uk/
39. Ferruci D., Lally A., UIMA: an architecture approach to unstructured information processing in a corporate research environment. Natural Language Engineering, 10(3-4), p. 327–348, (2004).
40. Muller H.-M., Kenny E. E., Sternberg P. W., Textpresso: an ontology based information retrieval and extraction system for biological literature , PLoS Biology, 2(11), 1984–1998, (2004).
41. Hernandez N., Mothe J., D'un thesaurus vers une ontologie de domaine pour l'exploration d'un corpus , In Actes de la conférence Veille Stratégique Scientifique & Technologique VSST (2006).
42. Lammari N., Métais, E., Building and Maintaining Ontologies: a Set of Algorithms , Data and Knowledge Engineering, 48(2): 155-176, (2004).
43. Salton, G., Buckley, C. Term-weighing approaches in automatic text retrieval. In Information Processing & Management, 24(5): 513-523, (1988).
44. Calvanese D., Giacomo G. D., Lenzerini M., A framework for ontology integration, In Proc. of the First Semantic Web Working Symposium, (2001).
45. Noy N. F., Musen M. A., Prompt: Algorithm and tool for automated ontology merging and alignment, in Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, (AAAI Press / The MIT Press, (2000)
46. Buneman P., Davidson S. B., Kosky A., Theoretical aspects of schema merging, in EDBT '92: Proceedings of the 3rd International Conference on Extending Database Technology, (Springer-Verlag, London, UK, (1992).
47. Canadian Thesaurus of Construction Science and Technology: http://irc.nrc-cnrc.gc.ca/thesaurus/.
48. TM4J website: http://tm4j.org/.