SyDoM : un système de recherche d'information multilingue basé sur des connaissances

Bruno Pivano, Sylvie Calabretto, Catherine Roussey, Jean-Marie Pinon

SyDoM : un système de recherche d'information multilingue basé sur des connaissances

Bruno Pivano[1], Sylvie Calabretto[1], Catherine Roussey[2], Jean-Marie Pinon[1]

[1]LIRIS INSA of Lyon 7 avenue J. Capelle 69621 VILLEURBANNE Cedex
[2] LIRIS – Université Lyon 1, batiment Nautibus, 8 boulevard Niels Bohr,
43 Boulevard du 11 Novembre 69622 VILLEURBANNE CEDEX

Résumé : Dans cet article, nous présentons les réalisations que nous avons effectuées dans le cadre du développement d'un système de recherche d'information multilingue capable de retrouver un document répondant à une requête écrite dans une langue différente. Nous avons développé un prototype, intitulé SyDoM (Système Documentaire Multilingue), permettant de gérer des documents XML contenant des langues différentes. SyDoM est un système complet répondant aux différents besoins d'une bibliothèque numérique pour la gestion de son fond documentaire. Notre système de recherche d'information multilingue s'appuie sur une gestion des connaissances d'indexation. Nous affirmons qu'une bonne indexation ne peut se faire sans au préalable, et parallèlement, effectuer une gestion des connaissances du domaine. Ce système utilise un nouveau modèle de représentation des connaissances proche des graphes conceptuels de Sowa (SOWA, 1984) : les graphes sémantiques. Le modèle des graphes sémantiques distingue les connaissances du domaine, du vocabulaire utilisé pour présenter ces connaissances. L'ensemble de ces connaissances est stocké dans un thésaurus sémantique, qui allie à une ontologie, plusieurs terminologies. Ce thésaurus sémantique est utilisé dans tous les modules de SyDoM.

Mots-clés : Recherche d'information multilingue, Bibliothèque numérique, Document XML, langage de représentation des connaissances, graphe conceptuel, ontologie.
Keywords: Multilingual information Retrieval, Digital Library, XML Document, knowledge representation language, conceptual graph, ontology.

1 Introduction

L'émergence de l'Internet a profondément transformé les moyens de communication, notamment en facilitant les échanges de documents entre les pays. Dès lors, les collections de documents se sont enrichies par des documents écrits dans différentes langues. Les bibliothèques ont du s'adapter à cette révolution technique pour devenir des bibliothèques virtuelles ou numériques, capables de gérer des collections multilingues de documents.

Pour faciliter l'accès à ces documents et en particulier, pour considérer leur aspect multilingue, il est nécessaire d'améliorer la représentation du contenu des documents. En effet, les mots présents dans les documents lorsqu'ils sont considérés individuellement, ne sont pas toujours suffisants pour exprimer la signification du document. Il faut donc travailler sur des éléments plus porteurs de sens que les mots à savoir les concepts. Une autre forme d'amélioration des représentations des documents consiste à ajouter des relations liant les concepts présents dans le document. Les concepts et les relations permettent d'enrichir les connaissances sur le document et favorisent la création de requêtes plus pertinentes. Par conséquent, la recherche d'information doit s'accompagner d'une fonction de comparaison adaptée qui tienne compte des connaissances implicites et explicites introduites dans ces nouvelles représentations. Pour les documents et les requêtes, ces représentations seront qualifiées de descriptions sémantiques.

En nous appuyant sur les graphes conceptuels de Sowa (SOWA, 1984), nous avons défini un nouveau modèle de représentation des connaissances : le modèle des graphes sémantiques qui distingue les connaissances du domaine du vocabulaire utilisé pour présenter ces connaissances. Nous avons également

développé un prototype, intitulé SyDoM (Système Documentaire Multilingue), permettant de gérer des documents XML contenant des langues différentes. Dans un premier temps, nous présentons les composantes du modèle des graphes sémantiques. Puis la deuxième partie développe la méthode d'indexation préconisée dans SyDoM pour construire une représentation du contenu documentaire. La troisième partie porte sur les algorithmes de recherches implémentés dans SyDoM pour comparer l'index d'un document avec une requête. La partie 4 présente les interfaces de SyDoM et toutes ces fonctionnalités. La cinquième partie propose une liste de systèmes comparables à SyDoM

2 Un nouveau modèle de représentation de connaissances

Notre système de recherche d'information multilingue s'appuie sur une gestion des connaissances d'indexation. Nous affirmons qu'une bonne indexation ne peut se faire sans au préalable, et parallèlement, effectuer une gestion des connaissances du domaine. Nous présentons dans le chapitre 2.1, le thésaurus sémantique, défini à partir du modèle des graphes sémantiques brièvement décrit dans le chapitre 2.2.

2.1 Thesaurus sémantique

Un thésaurus sémantique est un nouveau genre d'ontologie que nous avons défini dans (ROUSSEY, 2001b). Le thésaurus sémantique allie à une modélisation du domaine plusieurs terminologies. Ainsi les termes sont dissociés des notions qu'ils dénotent, ce qui permet de clarifier les relations entre les termes et les notions et d'identifier les relations terminologiques des relations sémantiques. De notre point de vue, un thésaurus sémantique définit deux niveaux de connaissances :

1. **Le niveau conceptuel** modélise le domaine d'étude formé de types de concepts ou de relations. Dans notre cas, il s'agit d'une conceptualisation du domaine résultant d'un consensus entre les différents acteurs d'un domaine particulier. Cette conceptualisation est utilisée comme langage pivot dans SyDoM, elle est équivalente au **support** du modèle des graphes conceptuels de Sowa (SOWA, 1984).
2. **Le niveau terminologique** est composé de l'ensemble des termes, le terme étant défini comme la manifestation linguistique d'un concept repéré dans un texte.

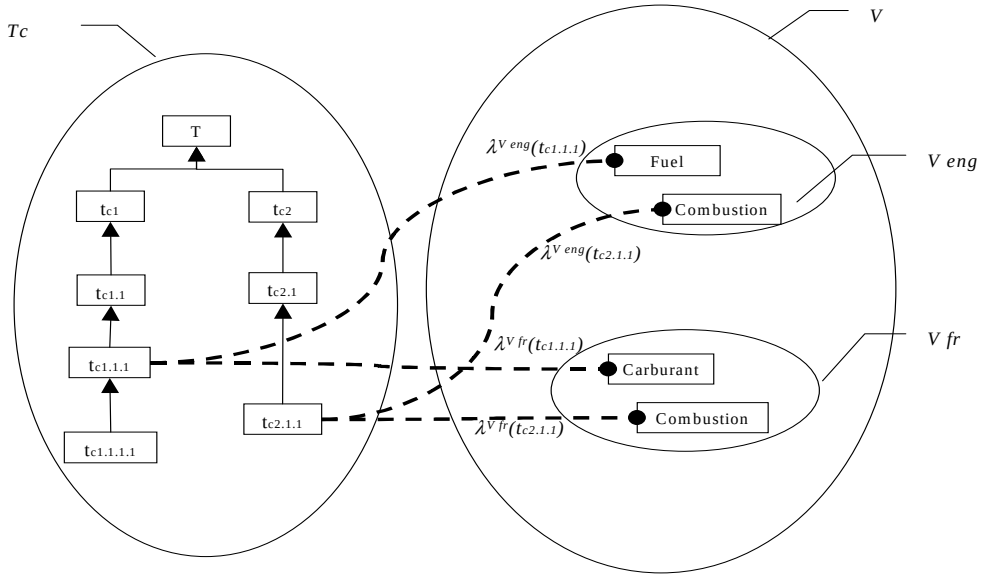


Fig. 1 – Un exemple de thésaurus sémantique.

Dans la figure 1, *V* se compose de deux vocabulaires, un vocabulaire anglais V_{eng} et un vocabulaire français V_{fr}. Pour chacun des vocabulaires V_i , il existe une fonction λ^{Vi} qui associe à chaque type au moins

un terme appartenant à V_i. Dans l'exemple de la figure 1, la fonction λ_C^{Veng} fait correspondre au type $t_{c1.1.1} \in T_C$, le terme anglais *"Fuel"* $\lambda_C^{Veng}(t_{c1.1.1})$= *"Fuel"* et la fonction λ_C^{Vfr} fait correspondre au même type $t_{c1.1.1}$, le terme français *"Carburant"* $\lambda_C^{Vfr}(t_{c1.1.1})$= *"Carburant"*.

2.2 Les graphes sémantiques

A partir de ce thésaurus sémantique définissant le vocabulaire et les connaissances du domaine, nous allons pouvoir définir les graphes sémantiques. Ces graphes peuvent être comparés aux graphes conceptuels de Sowa. La formalisation que nous proposons a pour but de mettre l'accent sur les couples de concepts reliés par une relation. Un exemple de couple de concepts est présenté dans la Figure 2 (dans le formalisme des graphes conceptuels). C'est pourquoi dans le modèle des Graphes Sémantiques est définie la notion d'arc : un **arc** se compose d'un couple de concepts étiqueté par un type de relations (Figure 3).

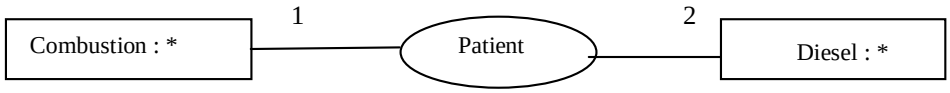


Fig. 2 – Un graphe conceptuel composé d'un couple de concepts relié par une relation

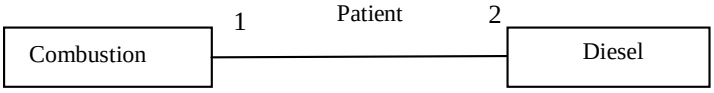


Fig. 3 – Un graphe sémantique composé d'un seul arc

Définition : Un *graphe sémantique Gs= (C, A, μ, labelc, ν, labelR),* défini sur un thésaurus sémantique *M*, est un multigraphe, non nécessairement connexe, où :

- C est l'ensemble des nœuds concepts[1] de *Gs.*
- $A \subset C \times C$ est l'ensemble des arcs de *Gs.* On notera un arc *a* comme équivalent au couple de concepts $(c, c') \in C \times C$: Par exemple, $a= (c, c') \in A$.
- $\mu : C \rightarrow T_C$ est une application qui à tout concept $c \in C$ associe une étiquette $\mu(c) \in T_C$, $\mu(c)$ est appelé le **type** de *c*.
- *labelC* est un ensemble d'applications $LabelC = \{labelC^{VL1}, \ldots, labelC^{VLj}, \ldots, labelC^{VLP}\}$ telle que $labelC^{VLj}: C \rightarrow V_{Lj}$ est une application qui à tout concept, $c \in C$, associe une étiquette correspondant à un terme de la langue *Lj*; $labelC^{VLj}(c) \in V_{Lj}$, $labelC^{VLj}(c)$ est appelé le **label** de *c* pour la langue *Lj*.
- $\nu : A \rightarrow T_R$ est une application qui à tout arc $a \in A$ associe une étiquette $\nu(a) \in T_R$, $\nu(a)$ est appelé le **type** de *a*.
- *labelR* est un ensemble d'applications $labelR = \{labelR^{VL1}, \ldots, labelR^{VLj}, \ldots, labelR^{VLP}\}$ telle que $labelR^{VLj}: A \rightarrow V_{Lj}$ est une application qui à tout arc $a \in A$ associe une étiquette correspondant à un terme de la langue *L*. $labelR(a)^{VLj} \in V_{Lj}$, $labelR^{VLj}(a)$ est appelé le **label** de *a* pour la langue *Lj*.

[1] Dans la suite de cet article, les termes "sommet concept", "nœud concept" et "concept" seront considérés comme équivalents.

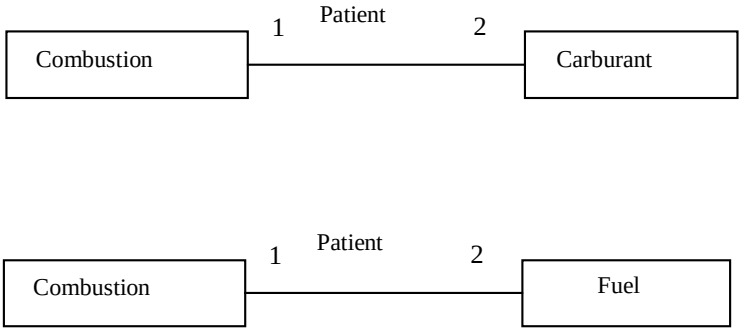


Fig. 4 – Des exemples de graphes sémantiques étiquetés par des labels.

2.3 Opération sur les graphes

L'opération de pseudo-projection est l'opération de base du modèle qui nous permet de rechercher un morphisme entre graphes. Cette opération sera utilisée comme fonction de comparaison entre les documents et les requêtes représentés sous forme de graphe sémantique. En effet, il existe une pseudo-projection d'un graphe *H* dans un graphe *G* si l'information représentée par *H* est voisine d'une partie de l'information représentée par *G*. *H* est dit **comparable à** *G*. Dans l'exemple de la Figure 5, le graphe *H* décrit un besoin d'information portant sur "le développement du diesel" et le graphe *G* représente le contenu d'un document traitant "du développement d'un nouveau carburant pour les moteurs monocylindres". L'existence de la pseudo-projection de *H* dans *G* prouve que le document indexé par *G* répond à la requête représentée par *H*.

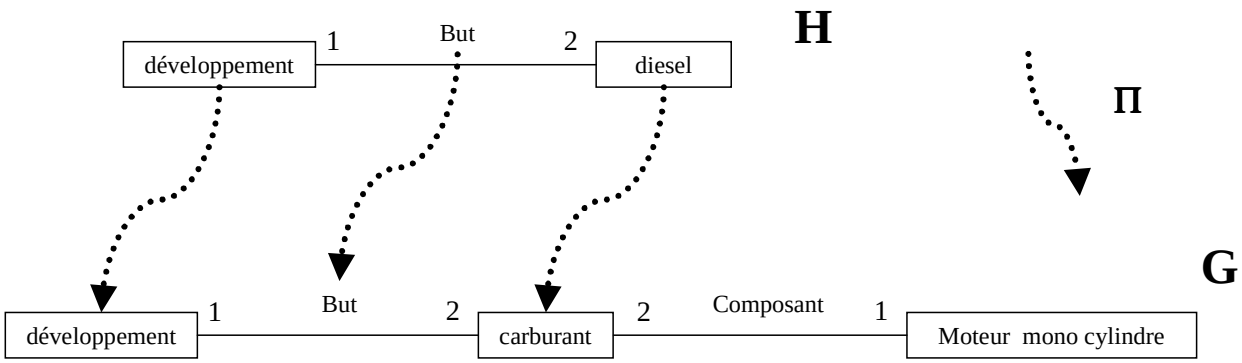


Fig. 5 – Exemple de pseudo-projection de *H* dans *G*.

La pseudo-projection est une opération moins contraignante que l'opération de projection définie par Sowa (SOWA, 1984). Par exemple dans la figure 5, il n'existe pas de projection de *H* dans *G* car *Carburant* n'est pas une spécialisation de *Diesel*. Plus formellement, nous obtenons la définition suivante pour l'opération de pseudo-projection :

Définition (pseudo-projection) : Une pseudo-projection d'un graphe $H = (C_H, A_H, \mu_H, labelC_H, \nu_H, labelR_H)$ dans un graphe $G = (C_G, A_G, \mu_G, labelC_G, \nu_G, labelR_G)$ est un couple de fonctions $\Pi = (f,g)$ où $f: A_H \rightarrow A_G$ établit la correspondance entre les arcs (relations) des deux graphes et $g: C_H \rightarrow C_G$ celle entre les sommets (concepts). Les propriétés de Π sont les suivantes:

- Π conserve les arcs : Pour tout arc $a = (c, c')$ de A_H, $f(a)=(g(c),g(c'))$ est un arc de A_G
- Π peut restreindre ou augmenter les étiquettes des arcs : Pour toute étiquette d'arc $\nu_H(a)$, $\nu_G(f(a)) \leq \nu_H(a)$ ou $\nu_H(a) \leq \nu_G(f(a))$.
- Π peut restreindre ou augmenter les étiquettes des sommets concepts : Pour tout sommet c de C_H , $\mu_G(g(c)) \leq \mu_H(c)$ ou $\mu_H(c) \leq \mu_G(g(c))$.

La pseudo-projection permet de comparer un graphe dont la totalité de son information est sémantiquement voisine d'une partie de l'information contenue dans un autre graphe. Il est important de définir un mécanisme permettant de comparer les graphes, partie par partie. Dans l'exemple de la Figure 5, nous avons montré qu'il existe une pseudo-projection de *H* dans *G*, mais inversement il n'existe pas de pseudo-projection de *G* dans *H* donc il ne sera pas possible de retrouver le document indexé par *H* avec une requête représentée par *G*. C'est pourquoi nous avons défini l'opérateur de pseudo-projection partielle dont la définition formelle est donnée par :

Définition (pseudo-projection partielle) : Il existe une pseudo-projection partielle de *H* dans *G* si, par définition, il existe *H'*, un graphe sémantique sous-graphe de *H* tel qu'une pseudo-projection de *H'* dans *G* pourra être trouvée.

Dans l'exemple de la Figure 5, il existe une pseudo-projection partielle du graphe *G* dans le graphe *H*.

Fonctions de similarités entre graphes

Les opérateurs de pseudo-projection (partielle) permettent de vérifier si deux graphes sont comparables. Le résultat de cette comparaison est booléen. Il existe une pseudo-projection ou il n'existe pas. Pour utiliser ces opérateurs comme fonction de comparaison d'un Système de Recherche d'Information, il est souhaitable que cette fonction retourne une valeur entre 0 et 1. Dans (ROUSSEY et al. , 2001c), nous avons défini des fonctions de similarités entre concepts et entre arcs pour aboutir à une fonction de similarité entre graphes. Ces fonctions de similarité sont basées sur les relations de spécialisation et de généralisation introduites dans le thésaurus sémantique.

3 La méthode d'indexation

Nous venons de définir un thésaurus sémantique contenant les connaissances utilisables comme descripteurs de documents, ces connaissances étant lisibles dans plusieurs langues. Maintenant nous allons indiquer comment nous utilisons ces descripteurs pour indexer les documents (ROUSSEY, 2001b). L'indexation tient compte des deux types de connaissances définies dans le thésaurus sémantique. Tout d'abord, nous annotons le document pour identifier les connaissances terminologiques. Puis, à partir de ces connaissances terminologiques trouvées dans le document, nous construisons un index contenant des connaissances du domaine.

3.1 Annotation

Le processus d'indexation comprend une étape d'identification des informations (l'étape d'annotation). Dans un premier temps, l'indexeur parcourt le document et sélectionne des mots ou expressions du document qu'il juge important. Ensuite, parmi ces mots ou expressions, il sélectionne ceux qui, de son point de vue, référencent une notion du domaine. Dans son contexte, l'occurrence d'un terme prend sens et ce sens sous un certain angle d'interprétation ou point de vue, rappelle une notion du domaine. Donc pour justifier cette interprétation, l'occurrence est associée à la notion du domaine par l'intermédiaire d'un nœud concept. De plus, une des fonctionnalités de SyDoM présentée dans le chapitre suivant utilise les annotations pour mettre à jour les connaissances terminologiques du thésaurus sémantique.

3.2 Indexation

Rappelons que l'index est une représentation précise et synthétique du contenu du document. Il remplace la liste des mots clés et le résumé, contenus dans la notice bibliographique du document. L'index représente, dans un langage de représentation des connaissances, les informations choisies, par les documentalistes, comme pertinentes. Ces informations devenues connaissances sont représentées par un graphe sémantique.

Le graphe sémantique représentant l'index du document est plus riche que le graphe sémantique résultant de la fusion des graphes d'annotation. Même si la construction de l'index se base sur les annotations, les annotations sont le résultat d'un premier niveau d'analyse du document, l'indexeur a

identifié les informations, alors que l'index est le résultat d'un second niveau d'analyse, l'indexeur a sélectionné les informations pertinentes. C'est à dire que l'indexeur déduit les concepts représentatifs du document, par sélection, spécialisation ou généralisation des concepts référencés dans le document par les annotations. Pour ne pas contraindre l'interprétation de l'indexeur et son analyse, nous lui laissons aussi l'opportunité de choisir comme représentatifs du contenu du document des concepts, non référencés par les annotations. Comparés aux annotations, les index explicitent les relations sémantiques qui existent entre les concepts.

4 Algorithmes de recherche

Avant de donner les grandes lignes de nos algorithmes de recherche (ROUSSEY et al. , 2001c), nous présentons la forme des requêtes utilisateur.

4.1 Formulation des requêtes

Une requête est définie elle aussi par un graphe sémantique. Par exemple, un utilisateur français cherche des informations concernant les lubrifiants de moteurs dans une collection de documents anglais. En naviguant dans les hiérarchies écrites en français de notions et de liens sémantiques, il compose un graphe sémantique requête comme suit :

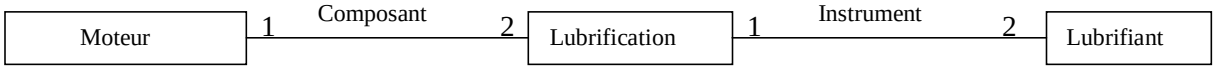


Fig. 6 – Un exemple de graphe requête en français

Le système transforme ce graphe, à l'aide du thésaurus sémantique, pour obtenir une requête en langage pivot.

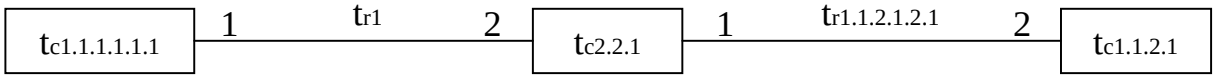


Fig. 7 – Un exemple de graphe requête en langage pivot

Ainsi le système pourra comparer le graphe sémantique représentant la requête avec les graphes sémantiques indexant des documents quelle que soit leur langue en utilisant les opérateurs de pseudo-projection et les fonctions de similarité.

4.2 Algorithmes

Nous utilisons une fonction de comparaison pour retrouver les documents correspondant à la requête. Cette fonction évalue les similarités entre le graphe requête et les graphes index. Tous les composants de ces graphes sont étiquetés par leur type. Donc les requêtes et les index sont représentés en langage pivot. La recherche d'une projection est un problème NP-complet. Il existe différentes solutions pour calculer l'existence d'un graphe de projection avec un temps de traitement raisonnable. La thèse de Guinaldo (GUINALDO, 1996) présente plus en détail ces travaux. En particulier, I. Ounis (OUNIS, 1995) propose d'utiliser des fichiers inverses et des tables d'accélération. Notre proposition s'inscrit dans la lignée des travaux de I. Ounis. Lors de l'indexation, nous mémorisons pour chaque graphe sémantique représentant un document, tous les sous-graphes requêtes possibles **comparables** à ce graphe index. Trouver les documents répondant à une requête revient à identifier les sous-graphes de la requête actuelle correspondant à un sous-graphe requête possible, stocké préalablement dans la base de données. Les graphes sémantiques étant

composés d'arcs et de sommets concepts, ils constituent par conséquent les deux types d'entités d'indexation. Nous avons donc considéré que le contenu d'un document serait représenté par deux index différents : une liste d'arcs et une liste de concepts. A partir de ces deux index, le graphe sémantique indexant le document, appelé aussi **graphe index**, peut être reconstruit car nous ne travaillons que sur la forme normale des graphes sémantiques où chaque étiquette de concept est unique. Par exemple, un document intitulé "*Evaluation des lubrifiants dans un moteur à réaction*" est représenté par le graphe sémantique de la figure 8.

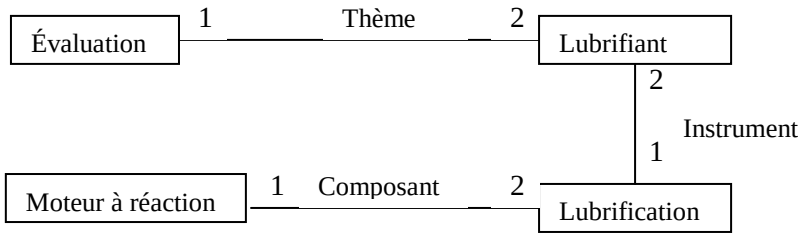


Fig. 8 – Le graphe sémantique représentant le document "*Evaluation des lubrifiants dans un moteur à réaction*"

Ce graphe étant sous forme normale, il se décompose de manière unique en trois arcs présentés dans la figure 9 et quatre nœuds concepts.

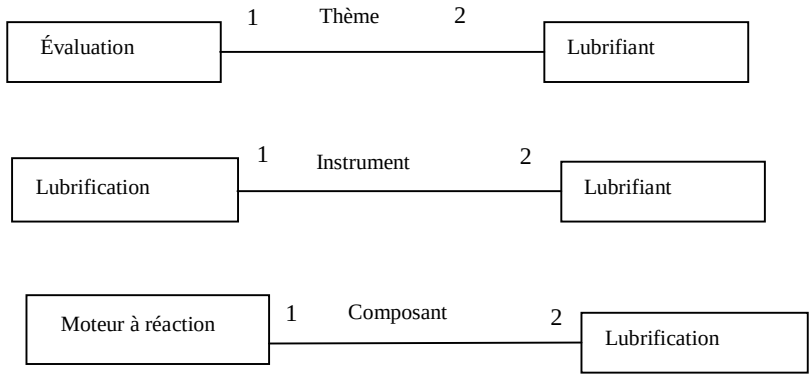


Fig. 9 – L'ensemble des arcs indexant le document "Evaluation des lubrifiants dans un moteur à réaction"

Nos algorithmes se basent sur l'association de fichiers inverses et de tables d'accélération (ROUSSEY et al. , 2001c). Les fichiers inverses associent, à chaque entité d'indexation (un arc ou un sommet concept), la liste des documents qu'elle indexe. Les tables d'accélération stockent, pour chaque entité d'indexation, la liste des entités qui leur sont comparables ainsi que le résultat de la fonction de similarité entre l'entité comparable et l'entité d'indexation. De cette manière, les tables d'accélération pré-calculent toutes les généralisations ou spécialisations possibles des entités d'indexation.

Lorsqu'un utilisateur formule une requête, celle-ci est décomposée de la même manière que pour les graphes index. On recherche ensuite les arcs du fichier inverse qui lui sont comparables, puis les sommets concepts. La fonction de comparaison entre graphes évalue ensuite une similarité entre les documents et la requête, ce qui permet de constituer la liste des documents jugés les plus pertinents. Cette liste de documents est ensuite affichée à l'utilisateur, chaque document étant pondéré par la similarité entre son graphe index et le graphe requête.

Nous présenterons plus en détail dans le chapitre suivant le prototype SyDoM (Système Documentaire Multilingue) qui a été développé pour concrétiser le modèle des graphes sémantiques.

5 Le système SYDOM

Nous avons développé un prototype intitulé SyDoM (Système Documentaire Multilingue) implémentant la méthode d'indexation et les algorithmes de recherche présentés plus haut. Dans le cadre de sa validation, nous avons collaboré avec la bibliothèque de l'INSA de Lyon (Doc'INSA). Doc'INSA nous a fourni un corpus de documents anglais : les pre-prints de la SAE (Society of Automotive Engineers) consacrés à la mécanique automobile. Ces documents ne sont disponibles actuellement que par une recherche sur les mots du titre, car les indexeurs ont des difficultés à trouver des descripteurs français correspondant aux termes anglais. Pour les aider dans cette tâche, nous avons construit un thésaurus sémantique pour le domaine de la mécanique comportant une trentaine de relations et une centaine de concepts tous associés à des termes français et anglais.

5.1 Principes généraux de SyDoM

L'interface de SyDoM a été développée en java à l'aide du JDK1.2.2. Pour le traitement des documents XML, nous avons utilisé l'ensemble des classes java XML4J développé par IBM. SyDoM est construit au-dessus du Système de Gestion de Base de Données relationnel ACCESS, mémorisant l'ensemble des données manipulées par SyDoM. Il se compose de différents modules, chacun de ces modules est dédié à une étape du processus d'indexation des documents XML (ROUSSEY et al. 2001a). SyDoM comprend (1) un module de gestion des thésaurus sémantiques. Ce module permet de construire un langage documentaire utilisé pour indexer et interroger les documents XML. Ce langage se compose d'une modélisation du domaine à laquelle sont associés plusieurs vocabulaires. (2) un module d'indexation manuelle de documents en XML. Ce module permet d'indexer les documents par des graphes sémantiques. (3) un module de recherche. Ce module permet de construire une requête sous forme de graphe sémantique et de récupérer la liste des documents répondant à cette requête.

5.2 Le module de gestion du thésaurus sémantique

Comme le thésaurus sémantique est l'élément central du système SyDoM, la création de ce thésaurus est une étape importante pour le processus d'indexation et de recherche. Ce module permet aux documentalistes, chargés de l'indexation d'un domaine, de représenter leur modélisation du domaine en créant une hiérarchie de types de concepts et de types de relations. La création des hiérarchies de types est accompagnée de la création des vocabulaires permettant de lier un ensemble de termes à un type.

5.3 Le module d'indexation

Le module d'indexation permet d'enrichir un document XML par de nouvelles balises sémantiques représentant des graphes sémantiques. Comme le montre la figure 10, ce module se compose, sur la partie droite, d'un navigateur permettant de parcourir le thésaurus sémantique, et sur la partie gauche d'une liste d'onglets. Ces onglets permettent de visualiser le document XML, ou un graphe sémantique indexant un document. La langue de présentation des informations peut à tout moment être changée, en sélectionnant une nouvelle langue à l'aide du bouton situé au-dessus de l'éditeur.

Les annotations

Nous rappelons que la première étape de la méthode d'indexation consiste à annoter les documents. Le documentaliste enrichit le document en associant des parties du document, qu'il juge important, à des types de concepts appartenant au thésaurus sémantique. Pour construire une annotation, l'utilisateur sélectionne une partie de texte dans le document et un type de concept dans le thésaurus sémantique. Puis, il sélectionne dans le menu annotation la fonction *insérer une annotation*. Comme le montre la figure 10, une fenêtre apparaît pour présenter les informations relatives à une nouvelle annotation. Dans cet exemple, le documentaliste annote la partie de texte « engine » avec le concept ayant pour terme français « moteur ». Cette annotation est représentée dans le document XML par une balise TERM encadrant la partie de texte « engine ». Cette balise possède un attribut *semkey* indiquant le type de concept C1.1.1.1.1.1 associé à la partie de texte. Les annotations vont permettre de mettre à jour les connaissances terminologiques du

thésaurus sémantique. En effet, si la partie de texte sélectionnée dans l'annotation n'existe pas comme label du type de concept dans le thésaurus sémantique, un nouveau terme peut être ajouté dans la liste des labels du type de concept. Pour se faire, il suffit de sélectionner le bouton *dans la liste label* de la fenêtre *insérer une annotation*

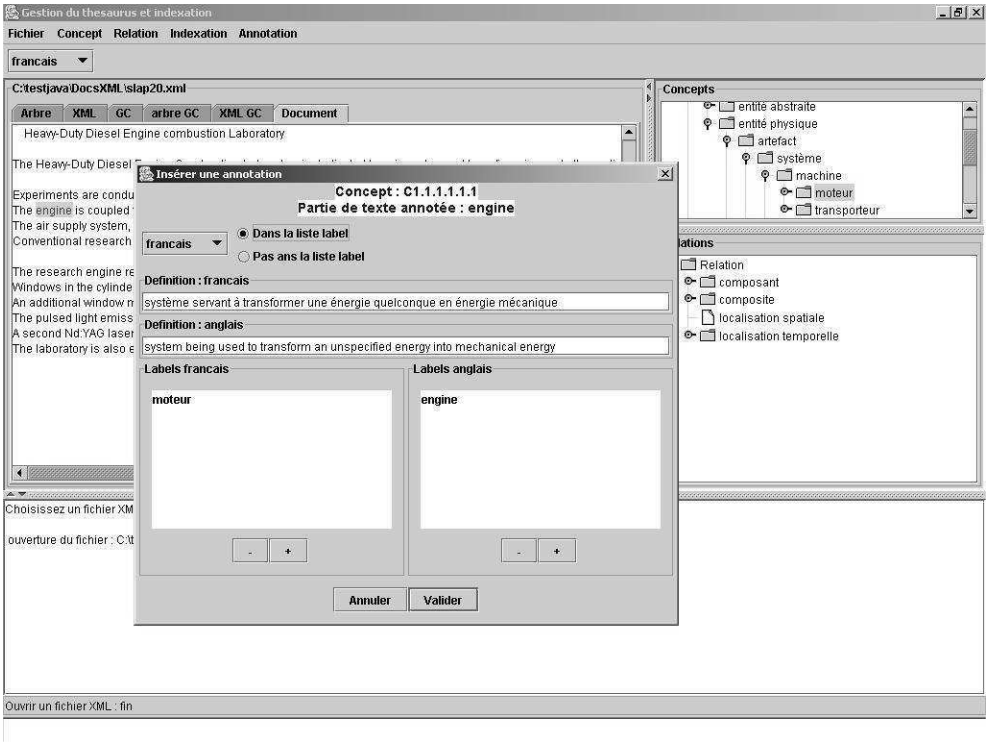


Fig. 10 – Le module d'annotation de SYDOM

Les index

Les parties de textes annotées sont mises en évidence dans le document. Ainsi ces annotations vont permettre au documentaliste de visualiser rapidement les termes importants du document et aider à la construction de l'index. Afin de construire un graphe index, le documentaliste parcourt le thésaurus sémantique dans la langue de son choix, et sélectionne des types de concepts ou les types de relations. Pour chaque type choisi, apparaît dans l'éditeur de graphe un nœud concept ou une relation, instance de ce type. A l'aide de cet éditeur, le documentaliste lie les nœuds concepts les uns aux autres, avec les relations. Une fois le graphe sémantique construit, le document XML est automatiquement enrichi en insérant une série de balises sémantiques au début du document (Pour plus de détails voir (ROUSSEY et al. 2001a)). Ces balises représentent le graphe sémantique dont chaque nœud concept et chaque relation est identifiée par son type. Ensuite, la base de données contenant les index est mise à jour grâce aux algorithmes présentés dans la quatrième partie.

5.4 Le module de recherche

Le module de recherche contient les mêmes composantes que le module d'indexation, c'est à dire un éditeur de graphes et un navigateur pour parcourir le thésaurus sémantique. L'éditeur de graphes permet à l'utilisateur de construire une requête dans la langue de son choix en sélectionnant des types, dans le thésaurus sémantique. Par exemple, la figure 11 présente un graphe requête correspondant au besoin d'information sur "la génération de bruit dans les moteurs diesels". L'utilisateur visualise les nœuds des graphes à l'aide des labels, tandis que le système ne travaille que sur les types des nœuds. Dans la première version du prototype SyDoM, la fonction de comparaison entre les graphes index et les graphes requêtes ne compare que les arcs des graphes. Autrement dit, les graphes sont décomposés uniquement en un ensemble d'arcs et seul l'algorithme de recherche sur les arcs a été implémenté. Le poids des documents résultat de la

figure 11 correspond uniquement à la similarité entre arcs du graphe requête et des graphes index. Dans la deuxième version de SyDoM est implémentée la totalité des algorithmes de recherche, c'est à dire que la fonction de comparaison entre graphes tient compte de la similarité entre les arcs et de la similarité entre les concepts. En résumé, ces fonctions de comparaison évaluent une similarité entre les documents et les requêtes, ce qui permet de constituer, pour chaque requête, la liste des documents jugés les plus pertinents. Cette liste de documents est ensuite affichée à l'utilisateur, chaque document étant pondéré par la similarité entre son graphe index et le graphe requête.

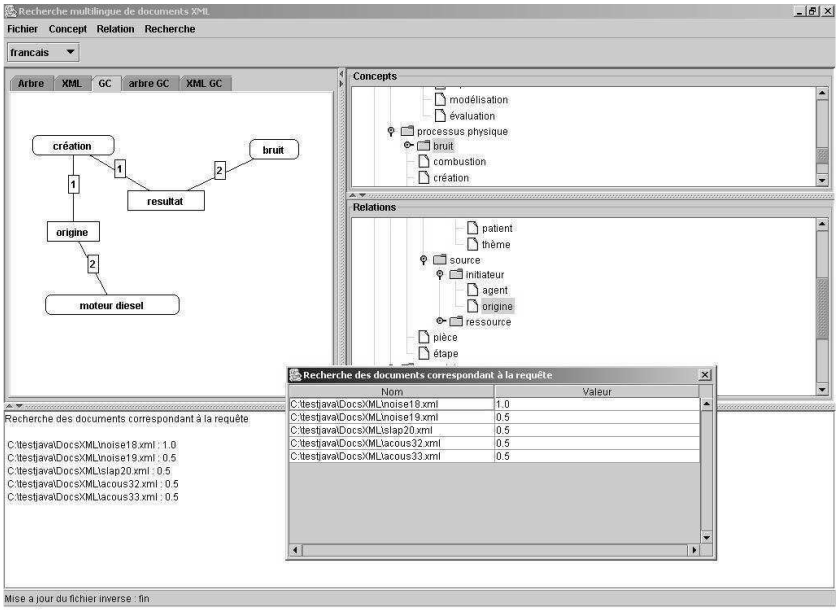


Fig. 11 – Un exemple de recherche

Visualisation des documents résultats

A partir de la liste des documents résultats, SyDom permet l'affichage d'un document sélectionné. En réalité, le document XML est mis en forme pour être présenté dans la langue de l'utilisateur. Les balises sémantiques de l'index sont remplacées par des termes choisis dans la langue de l'utilisateur, formant ainsi le résumé du document. Les annotations contenues dans le document pointent par l'intermédiaire de liens hypertextes sur les définitions des types de concepts dans la langue de l'utilisateur.

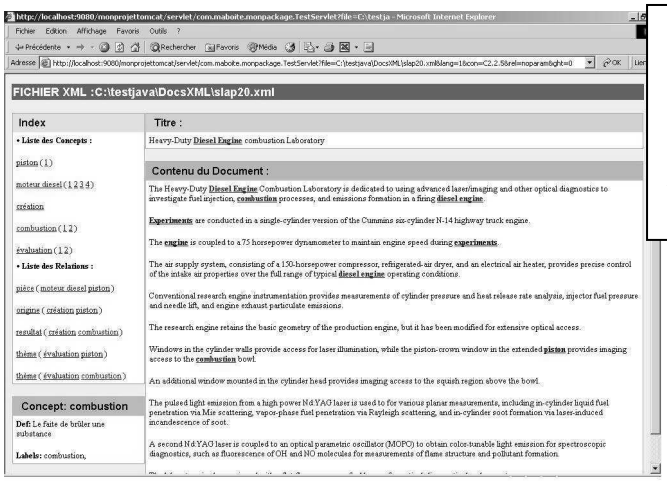
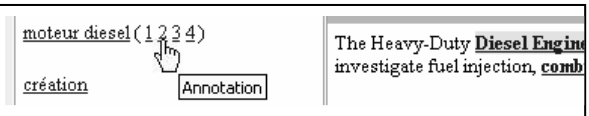
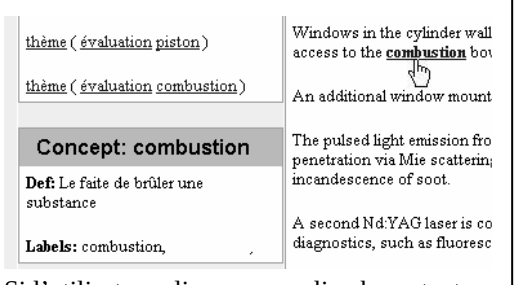


Fig. 12 – Un exemple de document résultat



Si l'utilisateur souhaite connaître l'endroit de la deuxième annotation de « moteur diesel » dans le texte, alors en cliquant sur « 2 » le système surligne l'occurrence dans le texte



Si l'utilisateur clique sur un lien hypertexte alors le système affiche la définition du concept ou de la relation sélectionné

6 Travaux apparentés

De nombreux travaux sont basés sur les Graphes Conceptuels : on peut citer les travaux de l'équipe de Michel Chein (CHEIN & MUGNIER, 1992) au LIRMM de Montpellier avec le système CoGiTant (GENEST, 2000), les travaux de l'IRIT de Toulouse (CHRISMENT et al., 1992), (PUGET, 1993), les travaux du CLIPS de Grenoble avec RELIEF (OUNIS, 1995) et ELEN (géniE logicieL et recherchE d 'informatioNs) (CHEVALLET, 1992), les travaux de l'INRIA Sophia-Antipolois avec les outils CGKAT et WebKB (extension de CGKAT au Web) (MARTIN & EKLUND, 2000). Ces systèmes diffèrent principalement par le niveau de prise en compte des graphes conceptuels : certains utilisent les graphes conceptuels uniquement au niveau de l'indexation, d'autres également au niveau de l'interrogation. Nous présentons trois outils qui nous semblent les plus proches de notre système : CoGitant, RELIEF et WebKB.

CoGiTant (GENEST, 2000) s'appuie sur une extension des graphes conceptuels permettant de réaliser un système de recherche d'information de type logique. Dans ses travaux, D. Genest s'est surtout intéressé à la réduction du silence produit lors de la phase de projection. En effet, une comparaison basée sur la projection est généralement binaire (la projection existe ou n'existe pas). De plus, le graphe conceptuel indexant un document doit être une spécialisation du graphe représentant la requête pour que le document soit jugé pertinent pour la requête. Malheureusement en recherche d'information, l'exactitude ou la précision n'est pas le critère unique d'un bon système de recherche d'information et s'oppose à un autre critère important le rappel. (GENEST, 2000) propose un mécanisme pour que des documents proches d'une requête (par exemple un document générique par rapport à la requête) soient aussi jugés pertinents. Pour améliorer les résultats de la recherche, il propose un ensemble de transformations sur les graphes conceptuels ainsi qu'un mécanisme d'ordonnancement de ces séquences de transformations. Le mécanisme de recherche détermine une séquence de transformation nécessaire à apporter aux graphes index pour qu'il existe une projection du graphe requête sur le graphe index. Les séquences de transformations étant ordonnées, ont obtient une liste des documents résultat ordonnée. D'une manière similaire, SyDoM propose une liste de documents résultats pondérés en fonction de la similarité entre le graphe index et le graphe requête.

RELIEF (OUNIS, 1995) est un système de recherche d'image basé sur le modèle des graphes conceptuels. Ce système utilise l'opérateur de projection pour comparer un document à une requête. Un aspect intéressant de ces travaux est de proposer un algorithme rapide et efficace de recherche de projection entre graphes. Actuellement, l'inconvénient majeur de l'utilisation de l'opérateur de projection comme fonction de comparaison entre document et requête réside dans le fait que les algorithmes de projection sont coûteux en temps de traitement. C'est pourquoi, le mécanisme de recherche proposé par Ounis effectue une série de pré-traitements au moment de l'indexation. Un fichier inverse et une série de tables d'accélération sont générés, à l'indexation, pour enregistrer toutes les spécialisations de sous-graphes requêtes possibles. Au moment de la recherche, il suffit d'identifier les sous graphes composant la requête pour retrouver tous les documents spécialisant la requête. Pour nos algorithmes de recherche, nous avons adapté la technique des fichiers inverses et des tables d'accélération de RELIEF pour l'utiliser avec notre opérateur de pseudo-projection. De plus nos algorithmes calculent une similarité entre graphes ainsi nous pouvons ordonnés nos documents résultats.

L'ensemble des outils de WebKB permet à ses utilisateurs de stocker, d'organiser et de retrouver des connaissances modélisées à l'aide des graphes conceptuels (MARTIN & EKLUND, 2000). Ces outils ne sont pas dédiés à la recherche d'information car leur but n'est pas de retrouver des documents mais de retrouver, à partir d'une requête, des connaissances répondant à cette requête stockées dans une base de connaissances. Plus précisément, WebKB propose une série d'outils pour l'acquisition de connaissances dont le but est de créer une base de connaissances illustrée par une documentation (un ensemble de textes au format HTML) qui ont permis de valider les connaissances de la base.

7 Conclusion et Perspectives

Nous avons défini un nouveau modèle de représentation de connaissances, le modèle des graphes sémantiques, pour représenter la sémantique des contenus documentaires et effectuer de la recherche de documents multilingues. Nous avons également proposé un nouvel opérateur de projection qui permet de comparer un graphe dont une partie seulement de l'information est voisine de l'information contenue dans un autre graphe. Nos travaux sur les graphes sémantiques ont abouti au développement du système SyDoM (Système Documentaire Multilingue). Ce système a été développé dans le cadre d'une collaboration avec la Bibliothèque de l'INSA (Doc'INSA). Nous l'avons évalué sur un corpus de documents anglais fourni par Doc'INSA : les pre-prints de la SAE (Society of Automotive Engineers) consacrés à la mécanique automobile. Nous explorons actuellement le passage à l'échelle de notre méthode d'indexation et de recherche, d'une part en éprouvant nos algorithmes sur des corpus volumineux, d'autre part, en essayant d'automatiser l'indexation. Une solution serait de réutiliser les connaissances des indexations antérieures. Par exemple, les connaissances terminologiques, stockées dans le thésaurus sémantique, permettraient de proposer un ensemble de concepts, à partir d'un terme et de son contexte d'utilisation. De la même manière, la base des index peut servir à proposer différentes relations déjà utilisées pour lier deux concepts donnés. Ainsi l'indexeur serait aidé dans sa tâche, et de plus la cohérence des indexations serait mieux maintenue.

Références

CHEIN, M. & MUGNIER, M-L. (1992). Conceptual Graphs: Fundamental Notions. In *Revue d'Intelligence Artificielle*, Vol. 6-4. p 365-406.

CHEVALLET, J-P. (1992). *Un modèle logique de recherche d'informations appliqué au formalisme des graphes conceptuels - Le prototype ELEN et son expérimentation sur un corpus de composants logiciels.* Thèse de doctorat en informatique : Université Joseph Fourier, Grenoble, France. 202 pages.

CHRISMENT, C., PUGET D., SOULE-DUPUY C. (1992). *Information Retrieval System Design.* Journées Information Technology and Applications. Actes édités Hemisphere Publishing Company (Whashington, USA), HIAST, Damas (Syrie).

GENEST D. (2000). *Extension du modèle des graphes conceptuels pour la recherche d'information*. Thèse de Doctorat en Informatique : Université de Montpellier II. 159 pages.

GUINALDO, O. (1996). Conceptual graphs isomorphism: Algorithm and use. In *Proceedings of the 4th International Conference on Conceptuals Structures (ICCS'96), Sydney*. p. 160-174. (Lecture Notes in Artificial Intelligence, Vol. 1115)

MARTIN, P., EKLUND P. (2000). Knowledge Indexation and Retrieval and the Word Wide Web. [On-line]. In *IEEE Intelligent Systems, special issue "Knowledge Management and Knowledge Distribution over the Internet"*. Disponible sur internet : <URL : http://meganesia.int.gu.edu.au/~phmartin/WebKB/doc/ papers/>

OUNIS I. (1995). *Une Dénotation pour les Graphes Conceptuels : comparaison avec les Logiques Terminologiques en Recherche d'Informations.* In Actes du XIIIe Congrès Inforsid, Grenoble. p. 147-164.

PUGET D. (1993). *Aspects sémantiques dans les Systèmes de recherche d'informations.* Thèse de doctorat en Informatique de l'Université Paul Sabatier – Toulouse 3, n°1658.

ROUSSEY C., CALABRETTO S., PINON J-M. (2001a). SyDoM: A Multilingual Information Retrieval System for Digital Libraries . In *Proceedings of the 5th International ICCC/IFIP Conference on Electronic Publishing (ELPUB'2001);Canterbury*, UK. p 150-164.

ROUSSEY C. (2001b). *Une Méthode d'Indexation Sémantique adaptée aux Corpus Multilingues*. Thèse de Doctorat en Informatique : INSA de LYON. 197 p.

ROUSSEY C., CALABRETTO S., PINON J-M. (2001c*). A new model of Conceptual Graph Adapted for Multilingual Information Retrieval Purposes . Proc.* Of the 12th International Conference on Database and Expert Systems Applications DEXA'2001,Technical University of Munich (Germany), Lecture Notes in Computer Science LNCS N°2113, pp. 92-101

SOWA J. (1984). *Conceptual Structures: information processing in mind and machine.* The System Programming Series. Addison Wesley publishing Company.