ANNOTATION SEMANTIQUE ET RECHERCHE DE DOCUMENTS FONDEES SUR LES GRAPHES CONCEPTUELS

Catherine Comparot,Ollivier Haemmerlé, **Nathalie Hernandez**
comparot@univ-tlse2.fr, haemmerl@univ-tlse2.fr, hernande@univ-tlse2.fr

IRIT, Université de Toulouse le Mirail, Département de Mathématiques-Informatique, 5 allées Antonio Machado, F-31058 Toulouse Cedex, France,

Mots clefs :
Recherche d'information, graphes conceptuels, activité de recherche, annotation sémantique, traitement de requêtes
Keywords:
Information retrieval, conceptual graph, semantic annotation, query processing.
Palabras clave :
Búsqueda de información, gráficos conceptuales, anotación semántica, tratamiento de preguntas

Résumé

Dans cet article, nous proposons un mécanisme d'annotation permettant d'interroger des documents à partir de la modélisation sémantique du contexte de l'interrogation. Ce contexte est modélisé par deux aspects dépendant des besoins de l'utilisateur : le sujet couvert par les document d'une part, les métadonnées associées aux documents d'autre part. Chacun d'eux est décrit par une ontologie. Les ontologies sont représentées à l'aide du modèle des graphes conceptuels.
Le mécanisme d'annotation sémantique construit automatiquement des graphes conceptuels en prenant en compte à la fois les caractéristiques du sujet et celles des métadonnées. L'interrogation s'appuie sur des patrons de tâches. Ces patrons représentent l'information (contenu ou métadonnée) pouvant intéresser les utilisateurs lors de leur activité de recherche. Ces derniers peuvent alors exprimer leur requête par le biais d'une interface graphique en remplissant le patron correspondant à leur besoin.

1 Introduction

De nombreux travaux en Recherche d'Informations (R.I.) cherchent à améliorer la méthode classique d'indexation de documents par mots-clés en lui substituant une indexation fondée sur les annotations sémantiques. Ces travaux visent à ajouter une couche d'annotation décrivant le sens des documents. Ces annotations sont fondées sur des ontologies [1] qui sont des descriptions formelles des concepts apparaissant dans les documents.

Le principe d'une annotation sémantique repose sur l'hypothèse selon laquelle le sens de l'information textuelle (et des mots qui composent un document) dépend des relations conceptuelles existant entre les objets du monde auxquels elle correspond plutôt que des relations linguistiques et contextuelles fournies par le document où elle se trouve [2]. L'objectif est de permettre le développement de moteurs de recherche sémantiques améliorant la précision et le rappel dans un processus de recherche d'informations.

La précision et le rappel peuvent également être améliorés en utilisant les métadonnées explicitement présentes dans les documents. Ces métadonnées sont fournies en utilisant un vocabulaire spécifique qui résulte souvent d'initiatives visant à faciliter l'intégration d'informations (e.g. Dublin Core pour l'archivage, LOM pour l'enseignement à distance). Dans ce cas, le sens des différents types de métadonnées [3,4] peut être fourni par une application s'appuyant sur une ontologie. Ainsi, si un utilisateur utilise dans une requête un concept pour qualifier le type d'une information recherchée, le moteur de recherche est capable de traiter toutes les métadonnées présentes dans les documents et décrites comme des spécialisations du concept.

Notre travail s'appuie sur un mécanisme d'annotation mettant en oeuvre plusieurs ontologies pour faciliter la réutilisation. Une telle approche est adoptée dans [5] pour améliorer la description et la recherche d'information de séquences de films médicaux. Les patrons d'indexation sémantique y sont développés en utilisant une ontologie qui formalise le contenu des documents (i.e. domaine médical), et une autre qui décrit les documents eux-mêmes (i.e. documents audio-visuels) en fonction des besoins de l'application (e.g. besoins d'archivage). Pour le développement de cours, [6] propose de faciliter la réutilisation en utilisant deux ontologies de façon à couvrir respectivement les aspects liés à la formalisation de la conception et du contenu des cours.

Dans notre approche, nous proposons de modéliser les documents par deux ontologies utilisées conjointement lors de l'annotation des documents. La première ontologie est l'*ontologie du thème* ; elle correspond au domaine traité par les documents. La seconde est l'*ontologie documentaire* ; elle correspond au type des documents eux-mêmes et permet de représenter les métadonnées. Les deux ontologies peuvent être réutilisées en fonction du type du corpus documentaire et du thème traité dans les documents. Nous proposons également de focaliser le mécanisme d'interrogation sur la tâche dans laquelle la recherche de documents s'effectue en utilisant des patrons de tâches définis sur la base de besoins en information spécifiques.

Une première étape de notre travail [7] présente l'utilisation de deux ontologies dans un contexte de recherche d'informations. Une des originalités de cette approche tient dans le choix automatique des concepts utilisés pour indexer un document. Ce choix est effectué suite à l'extraction des termes faisant référence aux concepts de l'ontologie, pondérés en fonction de leur représentativité dans le document. Une autre originalité est la façon d'accéder aux éléments d'information du corpus. Cet accès est réalisé en naviguant dans la collection via les deux ontologies.

Notre système de navigation est cependant limité :

- quand les ontologies sont constituées d'un très grand nombre de concepts et de relations, il est difficile pour l'utilisateur de choisir le point d'entrée pour naviguer ;
- la navigation ne peut pas satisfaire tous les besoins en information, notamment quand l'utilisateur est intéressé par plusieurs concepts qui ne sont pas directement liés à l'ontologie.

Nous présentons dans cet article la version préliminaire d'un moyen d'assister un utilisateur à formuler des requêtes.

Nous proposons de représenter la connaissance de notre système dans le formalisme des graphes conceptuels [8]. Nous pensons que ce modèle de représentation de connaissances est bien adapté à notre application pour plusieurs raisons : (i) les algorithmes de recherche ont été largement étudiés et sont bien adaptés à la recherche d'informations ; (ii) ce modèle graphique permet à des scientifiques non informaticiens de formuler des requêtes plus sophistiquées qu'une simple conjonction de mots-clés, sans la complexité des langages de requêtes de type SQL ou FLWR : la saisie d'une requête peut être réalisée intuitivement au moyen d'une interface graphique.
De précédents travaux ont étudié l'utilisation des graphes conceptuels pour la Recherche d'Informations [9,10,11]. Les documents et les requêtes sont représentés par des graphes conceptuels et la recherche de réponses est réalisée par l'opération de projection qui permet de tester la relation de spécialisation entre graphes. L'opération de projection classique est souvent étendue pour prendre en compte les spécificités de la Recherche d'Information en vue de renvoyer aussi bien les réponses exactes que les réponses pertinentes. Une des faiblesses de ces approches est que les graphes conceptuels sont construits manuellement. Le nombre de documents annotés est donc limité. Dans notre système, nous proposons de générer automatiquement les graphes conceptuels en utilisant l'ontologie de thème, l'ontologie documentaire et un processus reposant sur une analyse syntaxique des documents. Nous proposons ensuite un mécanisme d'interrogation dans lequel l'utilisateur est guidé par des patrons de tâche. Un patron de tâche caractérise le type d'informations intéressant les utilisateurs dans leur activité de recherche.

Notre travail s'inscrit dans le projet français WebContent [12]. Ce projet a pour objectif de créer une plateforme logicielle réunissant les outils nécessaires pour exploiter efficacement et étendre le devenir de l'Internet : le Web Sémantique. L'objectif principal est de produire une plate-forme flexible et générique pour la gestion de contenus et d'intégrer les technologies du Web Sémantique en vue de tester leur efficacité dans des applications réelles ayant de forts impacts économiques et sociaux. La premier groupe d'applications de la plate-forme traite de la veille économique dans l'aéronautique, l'intelligence stratégique, la prévention des risques microbiologique et chimique dans les aliments, l'observation des événements sismiques. L'entrée de notre système est un ensemble de dépêches fournies par des agences de presse. Ces dépêches sont constituées d'un ensemble de métadonnées et d'un contenu exprimé en texte libre. Notre objectif est d'annoter chacune de ces dépêches pour mettre en œuvre un processus de Recherche d'Informations sur ce corpus.
Cet article est organisé comme suit. Dans la section 2, nous présentons les ontologies de notre système et leur représentation dans le formalisme des graphes conceptuels. Dans la section 3, nous présentons le mécanisme d'annotation. La section 4 présente le processus d'interrogation de la collection.

2 Un modèle fondé sur des ontologies

2.1 Ontologie de thème et ontologie documentaire

Notre système est fondé sur deux ontologies, l'ontologie du thème et l'ontologie documentaire. L'ontologie de thème vise à organiser la connaissance du domaine d'application. Par exemple, dans notre étude, nous considérons une ontologie liée au domaine de l'aéronautique. Elle a été construite de façon semi-automatique [13] à partir d'un thésaurus du domaine et a été enrichie par un processus semi-automatique qui extrait les nouveaux concepts apparaissant dans une collection de dépêches de l'aéronautique.

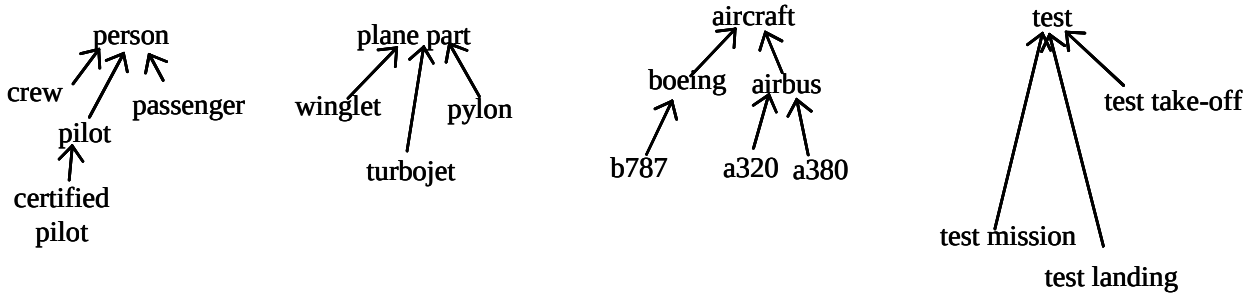

Fig. 1: Un extrait de l'ontologie de thème liée à l'aéronautique.

La figure Fig. 1 présente un exemple de l'ontologie de thème considérée. Cette ontologie structure la connaissance sur un avion : ses différentes parties (winglet, turbojet, pylon, …), les différents types de personne impliquées dans son fonctionnement (pilot, crew, passenger, …), les différents tests pouvant être effectués (test mission, test landing, test take off, …).

Parallèlement au domaine d'application représenté dans l'ontologie de thème, nous utilisons une ontologie documentaire. L'ontologie documentaire représente la connaissance associée aux métadonnées présentes dans les documents et leurs relations. Des efforts d'annotation réalisés au préalable sont ainsi pris en considération en exprimant la signification des métadonnées. Par exemple une ontologie documentaire relative à un corpus de publications scientifiques rend explicites les notions d'auteur, d'affiliation, de résumé, de référence qui sont habituellement étiquetées dans de tels documents. Dans notre étude, l'ontologie documentaire représente les métadonnées associées aux dépêches.

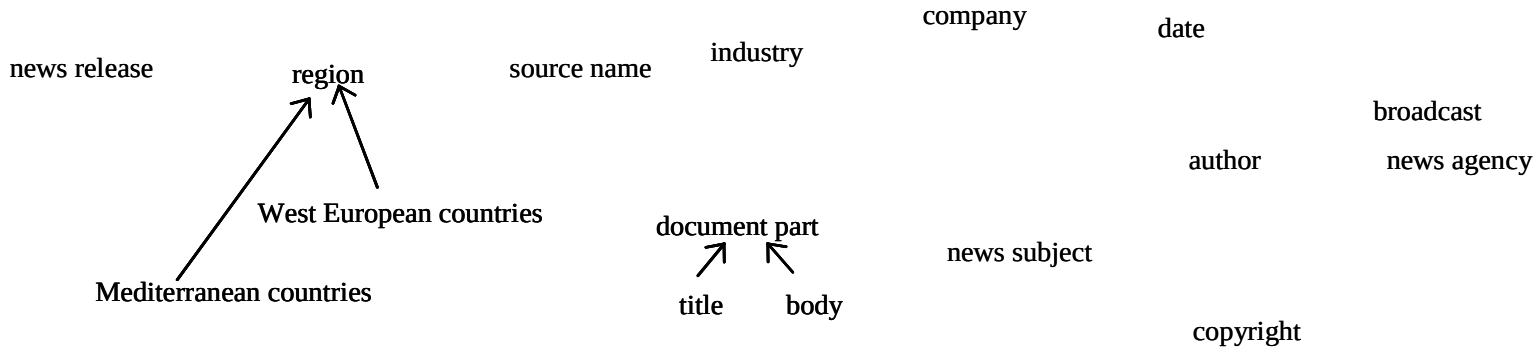

Fig. 2: Un extrait de l'ontologie documentaire liée aux dépêches.

La figure Fig. 2 met en évidence les différents concepts que nous considérons. Une dépêche est constituée de différentes parties : un titre et un corps. La source de la dépêche est caractérisée par son auteur, l'agence de presse qui la publiée, sa date de publication, son copyright. La publication prend en compte les différentes régions du monde, compagnies et domaines industriels potentiellement intéressés par la dépêche. L'objet de la dépêche fait référence à des sujets spécifiques qui ont été associés manuellement à la dépêche.
L'ontologie documentaire doit être réutilisable pour différents corpus partageant les mêmes métadonnées. Celle que nous proposons peut être réutilisée sur des corpus de dépêches traitant d'autres domaines que celui de l'aéronautique. L'ontologie de thème peut aussi être utilisée sur différents types de corpus traitant du même thème.
Dans cet article nous proposons de représenter les deux ontologies par une ontologie globale du système. Cette ontologie est exprimée dans le formalisme des graphes conceptuels. Nous utilisons la formalisation introduite dans [14], mais nous la modifions légèrement en donnant la possibilité d'ajouter un ensemble de synonymes à chaque type de concept ainsi qu'à chaque marqueur individuel, et en autorisant la représentation des valeurs numériques. La connaissance dans le modèle des GC est divisée en deux parties : la connaissance terminologique contenue dans le *support*, et la connaissance assertionnelle contenue dans l'ensemble des *graphes conceptuels*. Dans le modèle que nous proposons, les types de concepts liés au domaine d'application et les types de concepts liés aux métadonnées sont stockés dans le support. Les métadonnées associées à un type particulier de documents sont décrites au moyen d'un graphe conceptuel particulier, le *modèle de document*. Le modèle de document et les deux ontologies sont utilisés durant le processus d'annotation et au moment de l'interrogation, comme décrit dans les sections 3 et 4.

2.2 Le support

La base de connaissances exprimée dans le modèle des graphes conceptuels est construite sur un support qui contient la connaissance terminologique. Un support est un 5-uplet $S = (T_C, Syn_{TC}, T_R, M, Syn_M)$, T_C est l'ensemble partiellement ordonné de types de concepts, Syn_{TC} est l'ensemble des synonymes de types de concepts, T_R est l'ensemble partiellement ordonné de types de relation, M est l'ensemble de marqueurs individuels qui sont des instances de concepts, Syn_M est l'ensemble des synonymes de marqueurs individuels.
Dans notre application, les concepts de l'ontologie documentaire et ceux de l'ontologie de thème sont regroupés dans l'ensemble des types de concepts et sont partiellement ordonnés par la relation "sorte de". Deux types de concepts particuliers, *Document resource* et *Topic resource*, qui sont des sous-types immédiats du type de concept *Universal*, différencient les concepts relevant de l'ontologie documentaire - qui sont des spécialisations du type de concept *Document resource* - des concepts relevant de l'ontologie de thème - qui sont des spécialisations du type de concept *Topic resource*. La figure Fig. 3présente un extrait de notre ensemble de types de concepts.
Pour permettre l'annotation automatique de nos documents, nous mémorisons, pour chaque type de concept *t* spécialisation de *Topic resource*, un ensemble de synonymes noté $Syn_{TC}(t)$. Nous avons par exemple Syn_{TC}((A380) = {"Airbus A380" , "A380 aircraft" , "Airbus superjumbo" }.
Les types de relations associés à T_R représentent la nature des liens entre les concepts dans les graphes conceptuels. *for*, *loc*, *char*, *reg_dest*, *comp_dest*, *ind_dest*, *in*, *agt*, *origin*, …} sont des exemples de types de relations que nous utilisons dans notre application.
L'ensemble des marqueurs individuels *M* contient les instances des types de concepts. Il est partiellement défini *a priori* et est complété progressivement pendant l'annotation des documents, en extrayant par exemple les noms des auteurs d'une dépêche, considérés comme des instances du type de concept *Author*. Nous mémorisons pour chaque marqueur individuel *m* connu *a priori* un ensemble de synonymes noté $Syn_M(m)$. Cet ensemble est utilisé pour l'annotation sémantique durant l'analyse syntaxique des documents. Nous avons par exemple Syn_M (AFP) = {"Agence France Presse"}.

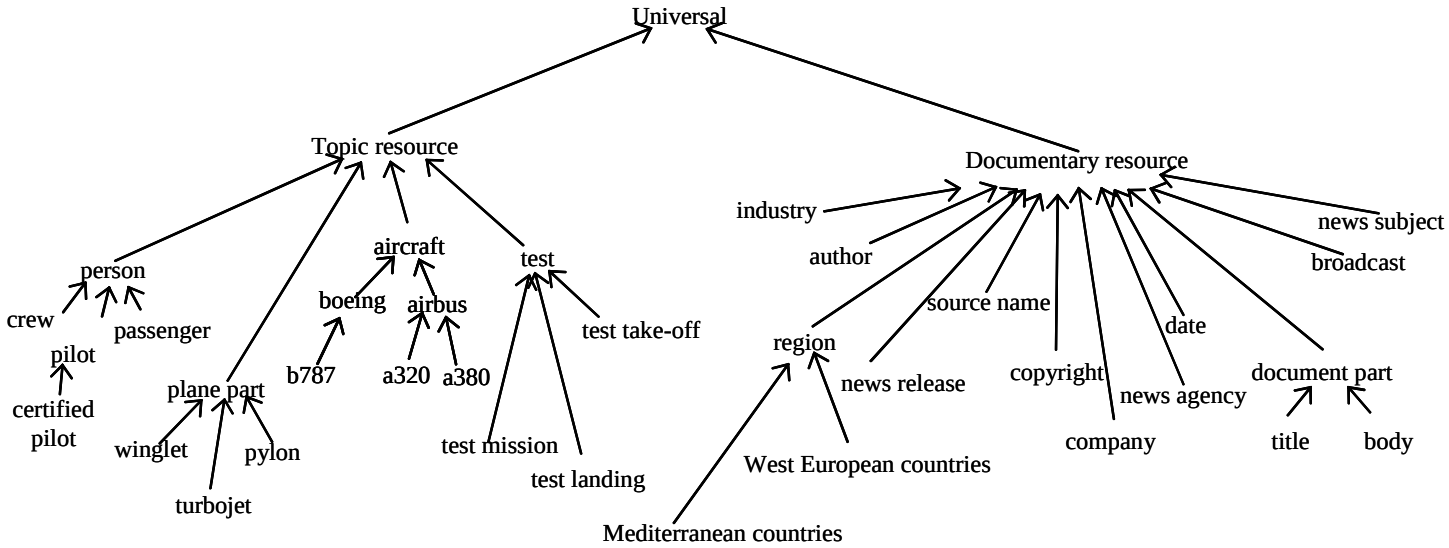

Fig. 3: Un extrait de l'ensemble des types de concepts de notre application.

2.3 Le modèle documentaire

Dans notre modèle, le *modèle documentaire* organise les métadonnées contenues dans le corpus.
Pour notre cas d'étude, nous avons choisi de représenter le modèle documentaire à l'aide des types de concepts suivants, spécialisations du type de concepts *Document resource* : *News release*, *Source*, *Broadcast*, *Title*, *Body*, *News subject*. Le graphe conceptuel M_{DOC} qui caractérise les relations sémantiques entre les concepts - et qui correspond donc à notre modèle documentaire - est représenté en Fig. 4.
Un ensemble de graphes conceptuels M^*_{DOC} est associé au modèle documentaire. Ces graphes conceptuels couvrent le graphe M_{DOC}. Ils représentent les granules d'informations élémentaires pouvant être retournés lors de l'étape d'annotation sémantique des documents. Dans notre exemple, l'ensemble M^*_{DOC} est le suivant :

[News release:*]-(origin)-[Source:*]
[Broadcast:*]-(obj)-[News release:*]
[News release:*]-(char)-[News subject:*]
[News release:*]-(part)-[Body:*]
[News release:*]-(part)-[Title:*]
[Source:*]-(char)-[Date:*]
[Source:*]-(char)-[Source name:*]

[Source:*]-(char)-[Author:*]
[Source:*]-(char)-[Copyright:*]
[Broadcast:*]-(comp_dest)-[Company:*]
[Broadcast:*]-(ind_dest)-[Industry:*]
[Broadcast:*]-(reg_dest)-[Region:*]

3 Annotation sémantique des documents

Contrairement à l'indexation par mots-clés [15], l'annotation sémantique se fonde sur l'hypothèse selon laquelle le sens d'une information textuelle (et des mots qui composent un document) dépend des relations conceptuelles entre les objets du monde auxquels elle fait référence plutôt que des relations linguistiques et contextuelles fournies par le document où elle se trouve [2].

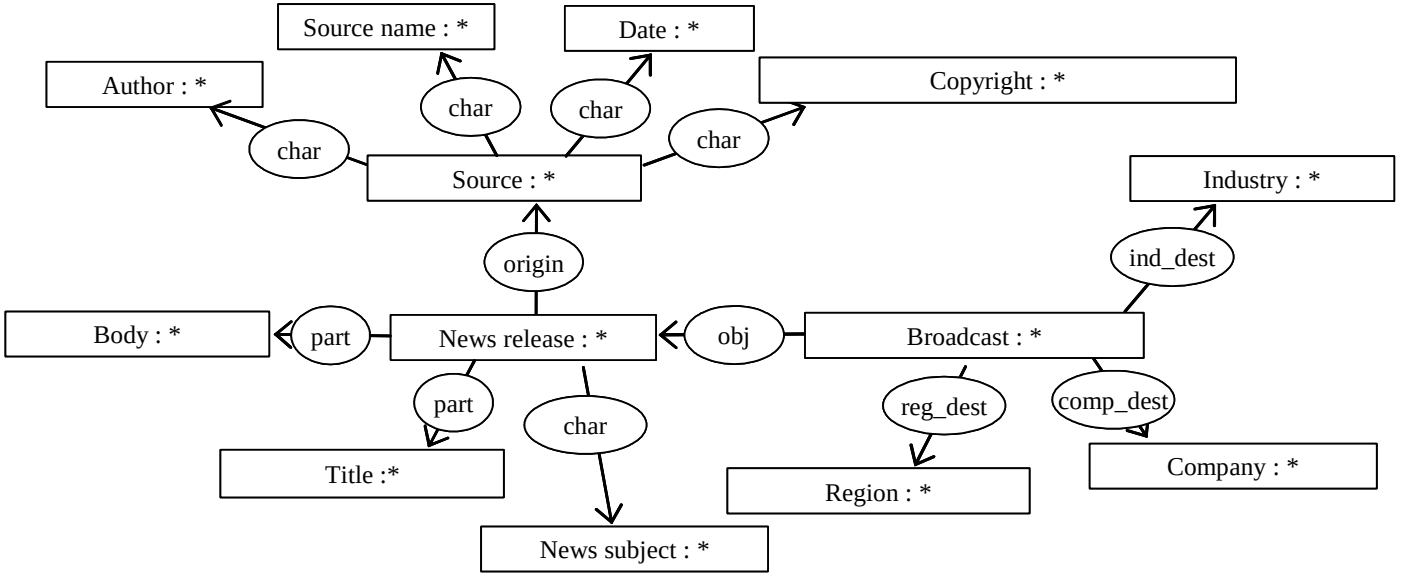

Fig. 4: Un exemple de modèle documentaire.

Dans notre système, chaque document est annoté par un graphe conceptuel qui décrit son contenu et ses métadonnées. Ce graphe conceptuel est construit en agrégeant des graphes conceptuels élémentaires appelés *motifs* qui sont retournés par les outils d'analyse de textes que nous utilisons. Ces motifs sont liés à la fois aux métadonnées et au domaine.
Nous présentons notre tâche d'annotation sur une dépêche réelle de l'AFP, partiellement représentée sur la figure Fig. 5.

Airbus A380 takes off for round-the-world test flights

PublishDate : November 13, 2006
SourceName : ***Agence France Presse***

TOULOUSE, France, Nov 13, 2006 (AFP) -

The Airbus A380 will take off from France on Monday for a round-the-world test mission, in the final hurdle before the superjumbo becomes the largest passenger plane in service. (…) On-board engineers and certified test pilots will put the plane through its paces under simulated commercial conditions, including test landings at key airports, refueling practices and maintenance work. The 150 hours of flying, which are expected to be the last major tests before approval from regulators next month, come at a difficult time for Airbus amid a hailstorm of bad publicity for its star project. Airbus has been forced to push back its timetable for deliveries of the A380 three times because of problems encountered when wiring the cabins, with delays now estimated at about two years. (…) Its competitor Boeing, however, has gone from strength-to-strength on the back of buoyant demand for its 787 Dreamliner jet. The A380 will leave Airbus headquarters near Toulouse on Monday, heading for Singapore then the South Korean capital Seoul on Wednesday. A second test flight will take it to Hong Kong on Saturday, then Narita in Japan on November 19, while a third test flight is to encompass airports in China, namely Guangzhou on November 22, then Beijing and Shanghai on November 23.

ar/cos/gk

Regions:
Mediterranean Countries/Regions
Western European Countries/Regions

Companies:
Airbus S.A.S.

Industries:
Air Transport
Civil Aircraft

AFPR000020061113e2bd000dx

Fig. 5: Un exemple de dépêche utilisée comme entrée de notre système d'annotation.

3.1 Annotation selon le modèle documentaire

Les motifs utilisés pour annoter selon le modèle documentaire sont des spécialisations des graphes de l'ensemble M^*_{DOC} présenté dans la section 2.3. Des techniques d'extraction de connaissances sont utilisées pour extraire ces motifs à partir des métadonnées explicitement présentes dans les documents. Une correspondance manuelle est établie entre les balises et les types de concepts correspondant aux métadonnées définies dans la DTD des documents. Pour chaque chaîne de caractères apparaissant à l'intérieur d'une balise donnée, une instance du concept correspondant est créée.
Le résultat de cette analyse est un ensemble de motifs instanciés par les valeurs extraites des documents. Dans notre exemple, le processus retourne l'ensemble de motifs suivant :

[News release:AFPR00...]-(origin)-[Source:src_news#00]

[Broadcast]-(obj)-[News release:AFPR00...]
[News release:AFPR00...]-(part)-[Body:body_news#00]
[News release:AFPR00...]-(part)-[Title:Airbus A380 takes off for round-the-world test flights]
[Source:src_news#00]-(char)-[Date:13/11/06]
[Source:src_news#00]-(char)-[SourceName:AFP]
[Source:src_news#00]-(char)-[Copyright:(c) Agence France Presse 2006]
[Broadcast:bcst_news#00]-(comp_dest)-[Company:Airbus S.A.S]
[Broadcast:bcst_news#00]-(ind_dest)-[Industry:Air Transport]
[Broadcast:bcst_news#00]-(ind_dest)-[Industry:Civil Aircraft]
[Broadcast:bcst_news#00]-(reg_dest)-[Mediterranean countries]
[Broadcast:bcst_news#00]-(reg_dest)-[West European countries]

3.2 Annotation selon le modèle de domaine

Pour chaque partie du document (dans notre cas le titre et le corps), notre outil d'analyse du domaine retourne des motifs de la forme :

[Document part : *]-(subj)-[Topic resource]-(weight)-[Numerical_Value:x] où

[Topic resource] est un type de concept sous-type de *Topic resource*, et

x est un poids associé par notre outil au type de concept correspondant dans le document.

Les motifs sont extraits par un mécanisme d'indexation fondé sur les types de concepts et les instances du domaine. L'indexation sémantique se décompose en deux étapes : l'identification des types de concepts (ou instances) dans les documents et la pondération de ces concepts (ou instances) afin de déterminer leur représentativité pour les documents.

Nous utilisons l'analyseur syntaxique Syntex [16], qui extrait l'ensemble des syntagmes (mots ou groupes de mots) de chaque document. Le regroupement de mots se fonde sur une analyse grammaticale. Les syntagmes sont ensuite confrontés à l'ontologie en identifiant le type de concept - ou l'instance - auquel ils correspondent. Si un syntagme fait référence à plusieurs types de concepts (le syntagme dépend de plusieurs synsets de *Syn*), un mécanisme de désambiguïsation est utilisé. Il s'appuie sur la distance sémantique calculée entre les types de concepts (ou instances) candidats et les types de concepts et instances déjà identifiés dans ce contexte pour le syntagme. Le candidat le plus proche sémantiquement des autres objets du contexte est choisi (voir [7] pour plus de détails sur cette technique).

Quand cette étape est terminée, les types de concepts ou instances référencés dans l'ensemble des documents sont identifiés. Ils sont ensuite pondérés en fonction de leur représentativité pour chaque document. Cette représentativité est fondée sur une pondération statistique et sémantique. La représentativité statistique s'inspire de la mesure tf.idf utilisée en recherche d'informations pour établir le pouvoir discriminant d'un terme. Appliquée aux types de concepts ou instances, cette mesure permet de sélectionner l'élément qui apparaît le plus souvent dans un document mais avec une faible fréquence dans le reste du corpus. La représentativité sémantique prend en compte le lien dans l'ontologie entre l'élément considéré et les autres éléments identifiés dans le document. Ce principe se fonde sur l'idée selon laquelle plus un élément est sémantiquement proche des autres éléments référencés dans le document, plus il est représentatif de l'ensemble des sujets du document. Pour combiner les mesures statistique et sémantique de la représentativité, chacune d'elle est normalisée par rapport au corpus global. Pour chaque document, l'ensemble des éléments identifiés (types de concepts et instances) et leurs représentativités respectives sont fournis par notre outil. La formule utilisée dans le processus d'indexation et l'évaluation de ce processus sont présentées dans [7].

Sur notre exemple, cette étape fournit les motifs suivants :

[Body:body_news#00]-(subj)-[a380]-(weight)-[Numerical_Value:0.8]
[Body:body_news#00]-(subj)-[Test mission]-(weight)-[Numerical_Value:0.6]
[Body:body_news#00]-(subj)-[Test landing]-(weight)-[Numerical_Value:0.4]
[Title:Airbus A380 takes off for round-the-world test flights]-(subj)-[Test flight]-(weight)-[Numerical_Value:0.6]
[Title:Airbus A380 takes off for round-the-world test flights]-(subj)-[a380]-(weight)-[Numerical_Value:0.6]

3.3 Construction du graphe conceptuel d'annotation

Un graphe conceptuel non nécessairement connexe est donc construit par une somme disjointe[1] puis une normalisation[2]. Le graphe conceptuel *F* présenté en figure Fig. 6 est le résultat du processus d'annotation sémantique appliqué à l'exemple présenté dans cette section.

4 Mécanisme d'interrogation

Dans notre système, nous offrons la possibilité aux utilisateurs d'exprimer leurs requêtes en remplissant des patrons de tâches. Pour chaque tâche de recherche identifiée sur un corpus, des patrons de tâches définissent l'information principale dont un utilisateur peut avoir besoin en fonction de caractéristiques liées au domaine couvert et de caractéristiques liées aux documents. Par exemple, pour notre cas d'étude, deux tâches ont été identifiées : l'analyse de la couverture médiatique de sujets particuliers et l'analyse de dépêches traitant de certains sujets.

La figure Fig. 7 présente un patron de tâche liée à la couverture médiatique de sujets particuliers. Dans ce patron nous faisons intervenir une dépêche, les parties de document traitant des sujets identifiés et les caractéristiques de diffusion de la dépêche en fonction de différentes régions, compagnies et industries potentiellement intéressées par la dépêche.

Lors de l'expression de leur requête, les utilisateurs choisissent le patron qui correspond à leur tâche. Ils peuvent ensuite spécialiser les sommets concepts : (i) en remplaçant un concept générique par un concept individuel (en remplaçant par exemple le sommet [Industry : *] par [Industry : Air Transport]), et/ou (ii) en spécialisant un type de concepts (en remplaçant par exemple [Region : *] par [Mediterranean countries : *]). Ils peuvent aussi supprimer certaines parties du graphe qui ne les intéressent pas. Finalement, ils peuvent ajouter autant de types de concepts du domaine qu'ils le souhaitent pour réaliser une recherche multi-critères sur ces sujets.

La figure Fig. 8 fournit un exemple de requête R_1 qui concerne la tâche de couverture médiatique présentée dans ce papier. L'utilisateur cherche une dépêche dont le corps traite d'A380 et de mission de tests, et qui est destinée aux pays d'Europe de l'Ouest ou du pourtour méditerranéen et aux entreprises de l'aéronautique.

Notre processeur de requêtes fonctionne de la façon suivante. L'opération de projection cherche des spécialisations du graphe de requête dans la base de données. Dans notre exemple le graphe de requête R_1 peut être projeté sur le graphe *F*. Les utilisateurs peuvent ensuite accéder aux documents annotés par *F* à l'aide de sa référence. Ils obtiennent par ailleurs des informations complémentaires (titre de la dépêche, compagnies potentiellement intéressées par la dépêche…). Finalement le système extrait le poids associé aux types de concepts ou instances du domaine (ici, 0.8 pour ‘A380’ et 0.6 pour ‘test mission’). Il

[1] La somme disjointe consiste en une juxapposition de deux graphes conceptuels résultant en un graphe conceptuel non connexe.
[2] La normalisation consiste à fusionner les sommets concepts qui correspondent au même marqueur individuel

retourne une mesure d'adéquation du document avec la requête pour les sujets recherchés. Le système calcule le poids moyen pour retourner un poids global pour le document (ici, 0.7).

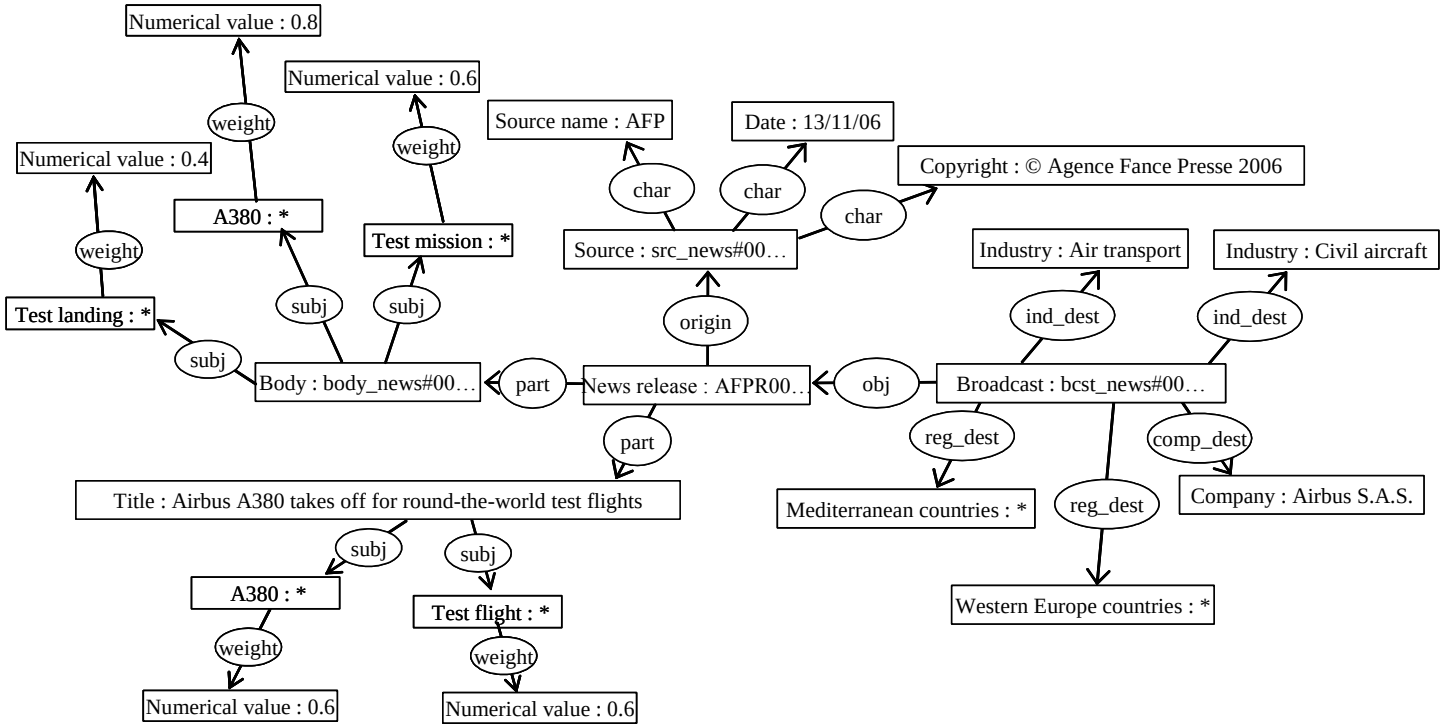

Fig. 6: Le graphe conceptuel F qui annote notre exemple de dépêche

Afin de compléter les documents qui répondent exactement à la demande, notre processeur de requêtes retourne les documents potentiellement pertinents. Par exemple, le graphe de requête R_2 présenté sur la figure Fig. 9 ne peut être projeté sur F.
Les réponses pertinentes sont obtenues comme suit. Le graphe requête est décomposé en sous-graphes partiels qui le couvrent. Ces sous-graphes partiels, appelés *sous-graphes requêtes*, sont décomposés selon l'ensemble M^*_{DOC} de graphes conceptuels qui couvrent le modèle de document M_{DOC} (chaque sous-graphe requête est ainsi une spécialisation d'un des graphes de l'ensemble M^*_{DOC} associé au modèle de document).

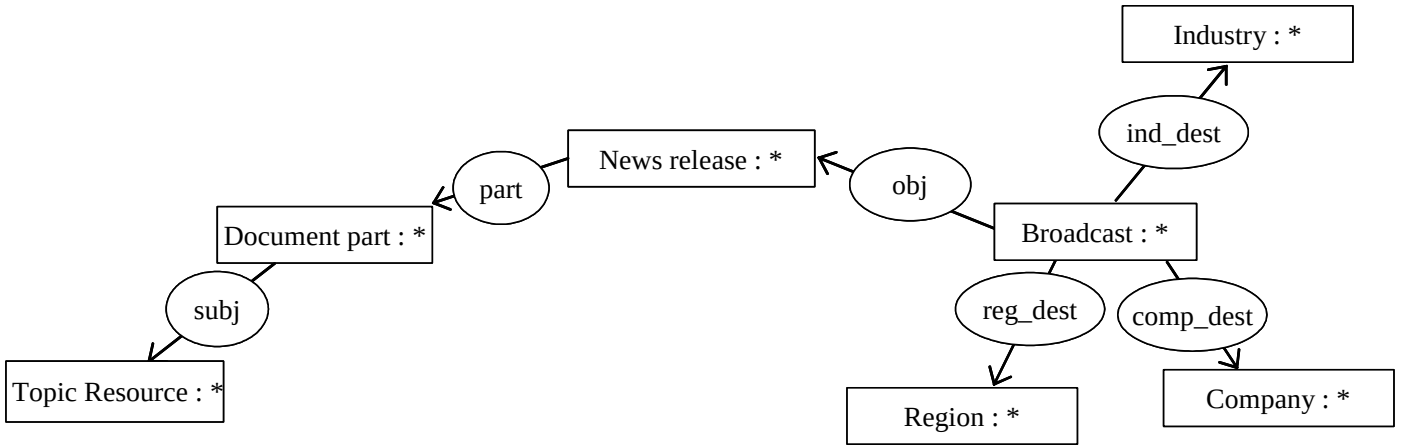

Fig. 7: Un graphe de patron.

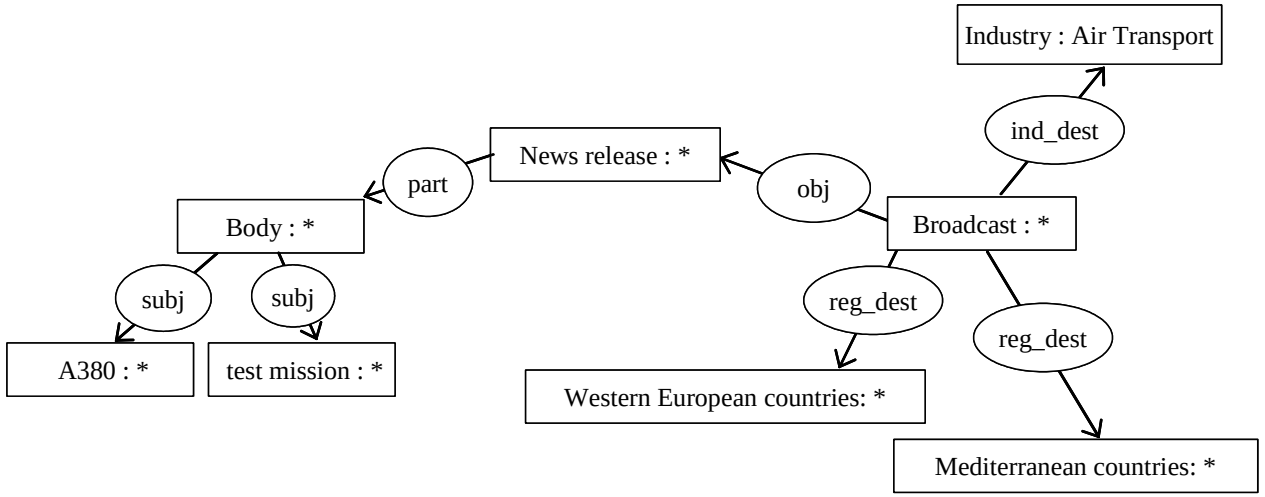

Fig. 8: Le graphe de requête R_1.

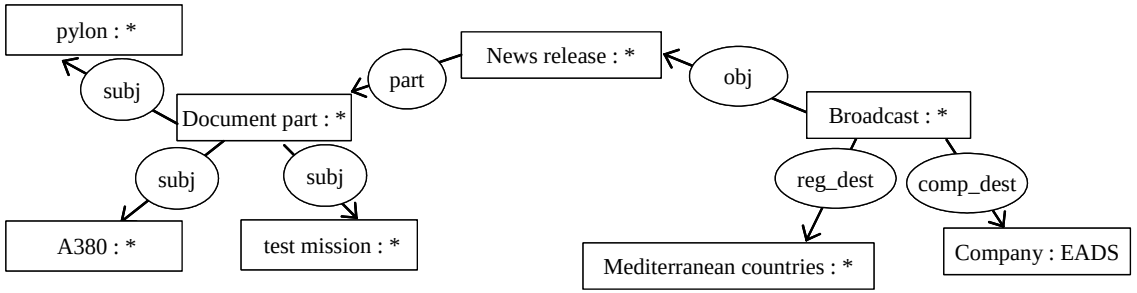

Fig. 9: Le graphe de requête R_2.

Par exemple, le graphe R_2 est décomposé comme suit :

r1 [Broadcast:*]-(obj)-[News release:*]
r2 [News release:*]-(part)-[Body:*]
r3 [Broadcast:*]-(comp_dest)-[Company:EADS]
r4 [Broadcast:*]-(reg_dest)-[Mediterranean countries]
r5 [Body:*]-(subj)-[A380:*]
r6 [Body:*]-(subj)-[Test mission:*]
r7 [Body:*]-(subj)-[Pylon:*]

5 Conclusion

Nous avons présenté dans cet article un mécanisme d'annotation automatique de documents. Les annotations sont construites à partir de deux ontologies : l'ontologie du domaine qui représente la connaissance dans le contenu des documents et l'ontologie documentaire qui structure les métadonnées associées à un certain type de documents. Nous avons choisi le formalisme des graphes conceptuels pour exprimer les ontologies. Ce formalisme est utilisé aussi bien pour annoter les documents que pour exprimer des requêtes sur le corpus. A nos yeux, ce travail présente plusieurs originalités :

- le processus d'annotation combine deux ontologies qui peuvent être réutilisées séparément dans d'autres contextes ;
- comparé aux autres travaux, nous proposons une construction automatique des graphes conceptuels utilisés pour annoter les documents. Notre objectif bien entendu n'est pas de construire des graphes conceptuels qui représentent complètement le contenu des documents, mais qui permettent une indexation satisfaisante du document ;
- nous proposons des patrons de tâche qui permettent aux utilisateurs d'exprimer facilement leurs requêtes en fonction de tâches qui leur sont propres.

Notre approche est actuellement expérimentée dans le cadre du projet WebContent. Nous avons annoté environ 150 dépêches fournies par la société EADS[3], partenaire du projet. Les dépêches traitent de divers événements qui se sont produits ces deux dernières années dans le domaine de l'aéronautique. Nous nous

[3] Nous remercions ESIS, EADS Shared Information System qui nous ont fourni le corpus sur lequel nous travaillons.

sommes focalisés sur deux tâches pouvant être réalisées sur ce corpus : l'analyse de la couverture médiatique d'un événement particulier et l'analyse de dépêches traitant de certains sujets. Cet article présente une première étape de notre travail. Nous travaillons à l'heure actuelle à caractériser les différents patrons de tâches en vue de faciliter leur conception et leur réutilisabilité. Nous souhaitons également enrichir les types de graphes conceptuels que nous construisons en analysant plus précisément le contenu textuel des documents. Enfin, ce travail sera intégré à la plateforme WebContent et évalué sur quatre domaines d'application : la veille économique dans l'aéronautique, l'intelligence stratégique, la prévention des risques microbiologique et chimique dans les aliments, l'observation des événements sismiques.

6 Bibliographie

[1] R.V. GUHA AND R. MCCOOL AND E. MILLER, *Semantic search*, Proceedings of the 12th International World Wide Web Conference, 2003, p 700-709

[2] H.M. HAAV AND T.L. LUBI, *A Survey of Concept-based Information Retrieval Tools on the* Web, Proceedings of the 5th East-European Conference ADBIS, 2001, p 29-41

[3] M. FISHER AND A. SHETH, *Semantic Enterprise Content Management*, Practical Handbook of Internet Computing, Chapman & Hall CRC Press, 2004

[4] M. DOERR, J. HUNTER AND C. LAGOZE, *Towards a Core Ontology for Information Integration*, Journal of Digital Information, Volume 4, Number 1, 2003.

[5] A. ISAAC AND R. TRONCY, *Using Several Ontologies for Describing Audiovisual Documents : A Case Study in the Medical Domain*, Workshop on Multimedia and the Semantic Web, Second European Semantic Web Conference (ESWC 2005), Springer-Verlag, p 95-104

[6] C. KNIGHT, D. GASEVIC AND G. RICHARDS, *Ontologies to integrate learning design and learning content*, Workshop Journal of Interactive Media in Education (Advances in Learning Design. Special Issue), 2005.

[7] N. HERNANDEZ, J. MOTHE, C. CHRISMENT AND D. EGRET, *Modeling context through domain ontologies*, Journal of Information Retrieval, Special Topic Issue on Contextual Information Retrieval, Volume 10, Number 2 / avril 2007, p. 143-172

[8] J.F.. SOWA, *Conceptual structures - Information processing in Mind and Machine*, Addison-Welsey, 1984.

[9] O. CORBY, R. DIENG-KUNTZ AND C. FARON-ZUCKER, *Querying the Semantic Web with Corese Search Engine*, Proceedings of ECAI, 2004, p.705-709.

[10] D. GENEST AND M. CHEIN, *A content-search information retrieval process based on conceptual graphs*, Knowl. Inf. Syst., Volume 8, Number 3, 2005, p.292-309.

[11] J.P. CHEVALLET, *E.L.E.N. : Un Système d'interrogation d'une base de logiciel*, in Congrès INFORSID 1991, p.1-20.

[12] WEBCONTENT, *The WebContent project*, 2000.

[13] N. HERNANDEZ AND J. MOTHE, *TtoO: Mining a thesaurus and texts to build and update a domain ontology*, Data Mining with Ontologies: Implementations, Findings, and Frameworks, Idea Group Inc., 2007.

[14] M. CHEIN AND M.L. MUGNIER, *Conceptual graphs, fundamental notions*, Data Revue d'Intelligence Artificielle, Volume 6, Numéro 4, 1992, p. 365-406

[15] C.J. VAN RIJSBERGEN, *A non-classical logic for information retrieval*, The computer journal, Volume 29, Numéro 6, 1992, p. 481-485

[16] D. BOURIGAULT AND C FABRE, *Approche linguistique pour l'analyse syntaxique de corpus*, Cahiers de Grammaire, 25, Université Toulouse le Mirail, 2000, p.131-151