Réédition de documents numériques guidée par un modèle utilisateur

Fady Farah

Laboratoire du Génie de la Conception (LGeCo)
Institut National des Sciences Appliquées (INSA)
24 boulevard de la Victoire, 67084 Strasbourg Cedex
fady.farah@insa-strasbourg.fr

RÉSUMÉ. *Notre travail se situe dans un contexte où une requête documentaire dans une base de documents XML d'un domaine spécifique fournit une masse de documents inexploitable par un humain. Un post traitement que nous appelons réédition est alors indispensable : il consiste à utiliser des unités d'information qui sont les éléments XML provenant des documents résultats de la requête pour composer de nouveaux documents. Une balise XML n'ayant pas de signification intrinsèque mais une interprétation donnée par son auteur, nous associons à chaque élément XML des connaissances spécifiques pour la réédition. Un ensemble conséquent de primitives de réédition ainsi qu'un langage de combinaison de tâches ont été définis, permettant l'utilisation des unités et des connaissances supplémentaires pour composer des documents. Enfin, un modèle utilisateur permet, grâce à des règles liées au domaine, de choisir et paramétrer les combinaisons de tâches de réédition à utiliser au moment de présenter les résultats du système de recherche d'information.*

ABSTRACT. *In this paper, we consider the problem of a request to an information retrieval system that leads to a lot of XML documents. To allow the user to exploit the results, a post process is necessary: it consists in composing information units (XML elements) extracted from resulting documents in order to obtain new documents that fit the needs. As an XML tag does not have an intrinsic significance, we add a complementary knowledge useful for republication to each tag. We have also defined an exhaustive set of republication primitives coupled to a tasks combination language that enable to use the information units and the associated knowledge to generate documents. Finally, a user model allows the system to choose a task combination after the request through domain association rules.*

MOTS-CLÉS : *annotation sémantique, unité d'information, réédition, profil utilisateur, système de recherche d'informations*

KEYWORDS : *semantic annotation, information unit, republication, user profile, Information Retrieval System*

1. Introduction

La recherche d'information permet la sélection de documents répondant à une requête généralement sous forme de mots clés. Il en résulte le plus souvent une masse de documents difficilement exploitable. Nous proposons pour pallier ce problème une approche similaire à (Cruzel et al., 2000) mais plus générique. Après une requête documentaire dans un domaine spécifique qui a fourni beaucoup de documents, nous effectuons un post traitement pour présenter les résultats. Ce traitement que nous appelons **réédition** utilise des **unités d'information** extraites des documents sélectionnés pour composer de nouveaux documents exploitables.

Notre approche se situe généralement dans un cadre structuré et s'applique à une base de documents XML ayant une certaine homogénéité. Nous commençons l'article par une description de ces documents ainsi que des connaissances qui leur font défaut dans une optique de réédition. Afin d'étendre notre travail aux documents semi voire non structurés, nous proposons dans la deuxième section un outil d'annotation XML. Dans la troisième partie, nous abordons la réédition qui est une combinaison de primitives de manipulation d'éléments XML. Enfin, nous détaillons la modélisation utilisateur permettant le choix d'une combinaison de tâches de réédition pour un utilisateur donné.

2. Grammaire des documents et connaissances pour la réédition

Les documents du Web étant très hétérogènes au niveau des formats et des thèmes, nous considérons une base documentaire XML d'un domaine spécifique ayant une DTD unique. Le format XML permet de structurer l'information contenue dans le document. Les éléments XML identifient des unités d'information, ou **fragments de document**, qui seront manipulées lors de la réédition.

Le premier problème de la réédition est de choisir les fragments pertinents entrant dans la composition d'un nouveau document. Cela nécessite de disposer de méta connaissances sur ces unités. Les DTD ou XML Schema permettent une description des balises XML mais se limitent aux relations père/fils, aux cardinalités et au typage. Nous proposons dans la suite une des connaissances complémentaires aux documents XML et à leur DTD et nécessaires pour la réédition.

La première constatation issue de la littérature (Pédauque, 2003) est que tout document possède une structure liée à ses objectifs. Cette structure a trois dimensions : la **dimension logique** possède un aspect explicite (parties, sections…) et un aspect implicite (sens du contenu) et organise la lecture d'un document ; la **dimension physique** concerne la mise en page et la typographie et ne doit pas apparaître dans le document XML ; la **dimension temporelle** (révision du document…) décrit l'évolution du document.

L'unité d'information possède un rôle (titre, conclusion…) qui détermine sa position dans la structure (arbre XML). La réédition ayant pour finalité la production

de document(s) répondant à des objectifs, et la structure étant liée aux objectifs du document, nous introduisons la connaissance **rôle du fragment de document**.

La cohabitation et la position relative des fragments de documents font apparaître une notion supplémentaire : les unités peuvent posséder des relations. Nous proposons donc d'adjoindre aux connaissances liées aux ressources XML les **relations sémantiques entre fragments**. L'idée d'inclure ces relations est inspirée de la RST[1] qui utilise des relations entre unités textuelles pour générer de nouveaux textes. Nous définissons cinq types de relations : (1) **généralise**/**spécialise** ; (2) **est composé**/**compose** ; (3) **est contexte de**/**a pour contexte** unissant deux balises dont l'une décrit le contexte (date, …) de l'autre ; (4) **référence**/**est référencé par** associant deux balises dont le contenu de l'une fait référence au contenu de l'autre ; (5) **met à jour**/**est mis à jour** décrit une relation temporelle entre deux contenus (ou unités appartenant souvent à des documents différents) dont l'un est modifié par l'autre.

Les relations (1), (2) et (3) sont liées au sens des balises et sont connues par l'expert qui conçoit la DTD du domaine. L'expert fournit également un document au format XML spécifique qui contient ces relations ainsi que les rôles et s'associe à la DTD pour constituer une **DTD enrichie**. Les relations (4) et (5), quant à elles, dépendent du contenu des balises, donc des instances de la DTD. Elles apparaîtront seulement après lecture des documents et seront annotées directement dans les documents.

Exemple 1 : extrait d'une DTD des JO de la Commission européenne

```
<!ELEMENT Document  (Titre, Préambule, Dispositif, Annexe*)> …
<!ELEMENT Dispositif (Article | Point | Partie | Chapitre | Section)+>…
```

Exemple 2 : extrait de l'extension de la DTD ci-dessus (rôles et relations)

```
<Balise nom="Dispositif">
  <Rôle>dispositif de loi</Rôle>
  <Relation type="est composé" balise-cible="Article" /> …</Balise>…
```

3. Préparation des documents : annotation XML et extraction d'information

Le but est, après conception de la DTD enrichie, de préparer les documents à la réédition dès leur entrée dans la base (et avant toute requête au SRI) c'est à dire:

- convertir un document au format XML valide (selon la DTD) : ce point concerne la structuration logique par l'annotation XML. On se place alors dans un cadre semi voire non structuré. De nombreux travaux s'intéressent à l'annotation XML et proposent des approches originales (Chidlovskii et al., 2005) trop complexes pour nos besoins. Pour un domaine technique où le langage montre des régularités et peu d'ambiguïtés, nous avons réalisé un outil basé sur les expressions

1. Rhetorical Structure Theory : théorie de la structure rhétorique

régulières. Cet outil permet de définir une expression représentant l'entité à baliser avec certaines contraintes sur son contexte, ou de définir une zone à baliser en donnant une expression identifiant le début et une autre identifiant la fin de l'intervalle. L'expert s'appuie sur la grammaire suivante pour définir les règles d'annotation des documents et obtenir des documents complétés :

COMBINAISON := REGEXP CONTRAINTE* | ([|]) REGEXP REGEXP ([|])
CONTRAINTE := (et | et pas) REGEXP CONTEXTE
CONTEXTE := dans INTERVALLE de UNITE
INTERVALLE := [n m] ; n, m ∈ Z et n<=m
UNITE := mot | phrase | ligne | paragraphe | texte | REGEXP
REGEXP := une expression régulière entre parenthèses

- identifier des relations induites par le contenu des balises (relation (4) et (5)) : ces relations, à l'inverse des autres, ne peuvent être renseignées lors de la définition de la DTD enrichie. Ce point fait appel à l'extraction d'information pour obtenir les relations induites par le contenu textuel des balises. L'identification de ces relations est manuelle pour l'instant. Nous étendrons l'outil ci-dessus pour l'automatiser.

*Exemple 3 : relation de **modification** induite par le contenu pour des textes des JO*

Extrait de la Décision 2000/766/CE (texte de loi)

<Article><Intitule>Article 2</Intitule>
<Paragraphe> 1. L'interdiction visée dans l'article 1 ne s'applique pas à l'utilisation:
<Tiret>- de gélatine de non ruminants pour l'enrobage des additifs,</Tiret>
<Tiret>- de lait et de produits laitiers dans l'alimentation des animaux.</Tiret>
</Paragraphe></Article>

Extrait de la Décision 2002/248/CE (amendement modifiant le texte ci-dessus)

<Article><Intitule>Article premier</Intitule>
L'article 2 de la décision 2000/766/CE est modifié comme suit:
<Point>1) Le paragraphe 1 est modifié comme suit :
<Point>a) le dernier tiret est remplacé par le texte suivant:
<Tiret>- de lait, produits laitiers et oeufs</Tiret></Point> …<Article>

Le travail consiste à identifier le type de modification (en rouge) et ses objets (modifiant en bleu et modifié en vert). La cible modifiée est identifiée dans un premier temps par des informations permettant de construire une adresse (décision 2000/766/CE, article 2, paragraphe 1, dernier tiret). Cette adresse est résolue lors de la réédition et la modification (remplacement ici) est effectuée.

4. Formulation des demandes en réédition : combinaison de tâches de réédition

Nous disposons de documents XML valides pour la DTD du domaine et des connaissances supplémentaires présentées précédemment. Le travail consiste à utiliser à présent toutes ces informations pour la réédition. Nous nous basons sur un découpage en **tâches de base** ou **primitives de réédition** que nous pouvons

combiner grâce à un langage pour obtenir des **tâches complexes** ou **combinaisons de tâches de réédition**. Ce découpage est inspiré des sciences cognitives et propose trois groupes de tâches :

- **Sélection :** sélection des fragments documentaires selon divers critères (rôles, relations et description de leurs cibles, attributs – valeurs, texte contenu).

Exemple 4 : Dans les textes de l'exemple 3, choisir les fragments ayant une relation de modification (remplace, est remplacé, est supprimé ...) entre eux

- **Reformulation de l'information ou composition :** combinaison d'unités extraites (groupage, tri), création ou insertion de nouvelles unités correspondant à des rôles (table matières, titres), suppression d'unités. Une définition complète de la composition est donnée dans la thèse de Sylvie Ranwez (Ranwez, 2000).

Exemple 5 : Appliquer les modifications aux fragments cibles obtenus dans l'exemple 4 (remplacements, insertions et suppression de fragments).

- **Formatage** : mise en forme (pagination, style) et sauvegarde (format)

Exemple 6 : Sauvegarder les documents produits dans l'exemple 5 en Html

Un ensemble assez conséquent de tâches de base a été défini pour chaque groupe de tâches présenté ci-dessus. Le langage de combinaison défini permet le séquençage des tâches ainsi que les boucles et les conditionnelles. La section suivante présente la sélection d'une tâche complexe de réédition pour une requête effectuée par un utilisateur donné.

5. Modèle utilisateur et sélection d'une tâche de réédition complexe prédéfinie

La modélisation de l'utilisateur a pour but de permettre au système de choisir, après une requête thématique ayant fourni une base de documents, une tâche complexe de réédition prédéfinie correspondant à l'utilisateur. Une solution est de définir une matrice d'association pour lier un but utilisateur et son niveau d'expertise dans le domaine à un patron prédéfini (template) de document résultat (Cruzel, 1999). Nous généralisons cette idée en considérant que selon le domaine, des règles établiront un lien entre un modèle utilisateur et des tâches complexes de réédition paramétrées. Les règles doivent donc être conçues pour chaque domaine de manière à utiliser un modèle utilisateur pour choisir les combinaisons de tâches de réédition et leurs paramètres.

Le modèle retenu qui n'a pas encore fait l'objet d'expérimentations contient six éléments à valeurs discrètes : **métier de l'usager** qui détermine les types de documents à produire (donc les 'bonnes' combinaisons de tâches) ; **niveau de compétence de l'usager dans le domaine** qui influe sur le paramétrage des tâches ; **niveau de privilège** qui autorise ou non la définition de tâches complexes ; **expérience des SRI** qui informe sur la précision des requêtes ; centres **d'intérêts** qui indiquent une hiérarchie de thèmes intéressant l'utilisateur ; **préférences** qui prennent en compte le choix du format de sortie, les habitudes de mise en page et de

typographie. Le choix de ces attributs a été effectué dans l'idée de conserver un système adaptable à tout domaine, ce qui distingue nos travaux de ceux de Cruzel.

6. Résultats et perspectives

Cet article illustre notre approche pour résoudre le problème de la présentation de l'information obtenue à la suite d'une requête à un SRI. Un prototype a été réalisé dans notre laboratoire et des expérimentations ont été menées sur des documents provenant des journaux officiels de la Commission européenne. Parmi ces expérimentations, l'application à un texte règlementaire - obtenu lors d'une requête - de ses révisions (textes d'amendements) pour fournir un document à jour à une date donnée (exemple 7) a démontré l'intérêt de nos travaux. Cependant, le modèle utilisateur et les règles d'association de ce dernier avec un type de réédition n'ont pas encore été mis en œuvre. Ils feront l'objet des prochaines expérimentations et permettront de rendre le système plus intelligent. Une difficulté rencontrée réside cependant dans l'évaluation de nos résultats. Ces résultats étant des documents, comment vérifier s'ils répondent bien à un besoin utilisateur aussi bien pour la présentation, que pour la structure et le contenu du document ?

Exemple 7 : extrait du résultat de l'application de l'amendement au texte de loi

Article 2
1. L'interdiction visée à l'article 1 ne s'applique pas à l'utilisation:
- de gélatine de non ruminants pour l'enrobage des additifs,
- ~~de lait et de produits laitiers dans l'alimentation des animaux d'élevage~~. (a)
- de lait, produits laitiers et oeufs (b)

(b) remplace (a)

7. Bibliographie

Chidlovskii B., Fuselier J. « A Probabilistic Learning Method for XML Annotation of Documents » *IJCAI, 19th International Joint Conference on Artificial Intelligence,* Edinburgh, Scotland, August 2005. Ref. : 2005/005

Lainé-Cruzel S., « PROFILDOC : Filtrer une information exploitable » *BBF Paris*, T. 44 n°5, 1999

Lainé-Cruzel S., Guinet E., « Fragmentation et enrichissement de textes scientifiques sous forme électronique », *Document Numérique*, Editions Hermès, Vol. 4, n°1-2, 2000, pp.59-84

Pédauque R. T., « Document, signe et médium, les re-formulations du numérique », 2003, STIC-CNRS

Ranwez S., « Composition Automatique de Documents Hypermédia Adaptatifs à partir d'Ontologies et de requêtes intentionnelles de l'Utilisateur », Thèse de doctorat, université de Montpellier II, décembre 2000