PROJET
De la recherche d'information au Web Sémantique

1. Présentation

Le projet consiste à mettre en place un système de recherche d'éléments XML. Contrairement aux systèmes de recherche d'information traditionnels qui retournent pour une requête utilisateur une liste de documents, le système que vous allez implémenter doit retourner une liste de **paragraphes** XML pertinents pour la requête.

Nous vous recommandons de développer le cœur du système en Java et de stocker les index dans une base de données MySQL. Vous pouvez également rechercher des modules existants et les intégrer dans la mesure où vous respectez les étapes détaillées dans la section suivante (voir annexe 1 pour quelques pointeurs).

Le projet s'effectuera en binômes.

2. Etapes du projet

Le projet est décomposé en plusieurs étapes :
Partie 1 : implémentation du système de recherche d'information (SRI) classique
- sous étape 1-a : implémentation du module d'indexation
- sous étape 1-b : implémentation du module de recherche
- sous étape 1-c : évaluation de votre système

Partie 2 : utilisation d'une ontologie de domaine pour la reformulation de requêtes dans un SRI
Partie 3 : génération de triplets RDF à partir de documents XML
Partie 4 : interrogation de triplets RDF à partir de SPARQL

Les parties 2 à 4 seront détaillées ultérieurement.

Les sous-étapes de la partie 1 sont détaillées dans la suite du document. Elles s'étaleront sur les 2 premières périodes de l'année

3. Ressources à votre disposition

3. a Collection de documents

Il s'agit de documents en Français contenant des récits de balades et voyages (en France ou à l'étranger), balisés au format XML. Le balisage de ces documents est simple, et ils ne contiennent pas de contenu mixte[1].
La collection est composée de 103 documents et on trouvera un exemple de document en annexe 2.
La collection est téléchargeable : http://www.irit.fr/~Nathalie.Hernandez/M2ICE/Collection.zip

3.b Collection de requêtes

Un ensemble de requêtes est à votre disposition pour évaluer votre système: http://www.irit.fr/~Nathalie.Hernandez/M2ICE/Queries.zip.

Pour chacune d'entre-elles nous vous fournissons également les jugements de pertinence, ie les éléments XML de type paragraphe (balise <p>) de la collection qui ont été jugés pertinents par des utilisateurs pour chacune des requêtes. Les résultats sont mis dans un fichier appelé « qrel.XX » où XX correspond au numéro de requêtes. On trouvera un exemple de fichier qrel en annexe 3. Les fichiers qrel sont disponibles à http://www.irit.fr/~Nathalie.Hernandez/M2ICE/qrels.zip
Attention, ce n'est pas parce que les termes de la requête ne sont pas présents dans un élément que cet élément est non pertinent. Les jugements peuvent être subjectifs.

[1] On parle de contenu mixte lorsque des éléments sont imbriqués dans du texte.
Par exemple :<section> aaa bbb aaa bbb kkkk aa ccc </section>

4. Détail des étapes de la partie 1

Lors de l'étape 1-a, vous allez mettre en place les différents mécanismes relatifs à l'indexation. *Attention de bien réfléchir à la structure de vos index et à la façon dont vous allez lier l'information textuelle et l'information structurelle, ainsi qu'à l'algorithme de recherche que vous allez utiliser dans l'étape suivante.*

Il s'agit de :

- parcourir les documents XML avec un parseur SAX ou DOM – Voir annexe 1 pour plus d'informations.
- récupération des informations nécessaires aux index
  - éléments de structure
  - information textuelle
    - reconnaître les mots dans une séquence de lettres ou des symboles composant l'élément. On considère que les espaces et toutes les ponctuations constituent un séparateur de mots.
    - nettoyer les mots composant les passages à partir d'une stopliste dont on trouvera un exemple à l'adresse suivante : http://www.irit.fr/~Nathalie.Hernandez/M2ICE/stopliste.txt
    - lemmatiser les termes (via des troncatures puisque les documents sont en Français)
    - implémenter les tables dans la base de données

On trouvera le .jar permettant la liaison avec la base de données à l'adresse suivante : http://www.irit.fr/~Nathalie.Hernandez/M2ICE/mysql-connector-java-3.0.15-ga-bin.jar

=> Vous devrez rendre le MCD de votre index ainsi que le code permettant l'indexation le dernier vendredi de la P1.

Lors de l'étape 1-b, vous allez devoir choisir un modèle de pondération des paragraphes. Pour simplifier le travail demandé, utilisez les facteurs *tf* , *idf* et/ou *ief*, et éventuellement l'information portée par les autres éléments des documents. Vous devez réaliser le processus de recherche en vous fondant sur les résultats d'indexation que vous avez produits précédemment. Le module doit prendre en entrée une requête composée de mots clés et produire une liste ordonnée d'éléments XML de type <P> comme réponse.

L'étape 1-c consiste à évaluer les performances de votre système en termes de rappel et de précision. L'évaluation sera effectuée sur toutes les requêtes fournies. Vous devrez pour cela programmer un 'évaluateur' qui prendra en entrée un fichier de type *qrel* et un fichier de résultats. L'évaluateur proposera en sortie les **précisions à 5, 10 et 25 éléments pour chaque requête**, ainsi que les **précisions moyenne à 5, 10 et 25 éléments pour toutes les requêtes**.

=> Vous devrez rendre votre moteur de recherche ainsi qu'une analyse critique des résultats obtenus sur les qrel le dernier vendredi de la P2.

Annexe 1 : Documentation

Documentation Java 1.5:
http://java.sun.com/j2se/1.5.0/docs/api/

Documentation parseurs XML (SAX, DOM, …) :
Pour débuter :
SAX : http://java.developpez.com/faq/xml/?page=sax
http://smeric.developpez.com/java/cours/xml/sax/
DOM : http://java.developpez.com/faq/xml/?page=dom
XPath et Java: http://java.developpez.com/faq/xml/?page=xpath

Et le site du W3C pour en savoir plus….

Documentation JDBC :
http://www.dil.univ-mrs.fr/~massat/ens/java/jdbc.html
http://jguillard.developpez.com/JDBC/

Annexe 2 : Exemple de document XML de la collection de test

```
<?xml version="1.0"?>
<!DOCTYPE BALADE SYSTEM "balades.DTD">
<BALADE>
   <PRESENTATION>
      <TITRE>Ascension du Mont-Blanc</TITRE>
      <DESCRIPTION>
         <P>Récit de mon ascension du Mont Blanc par le refuge du Goûter (voie de St Gervais).
C'est la voie la plus parcourue du Mont Blanc, elle est principalement utilisée pendant la
période estivale d'ouverture du refuge du Goûter. La première ascension du Mont Blanc remonte
au 8 août 1786 par Jacques Balmat et le docteur Michel Paccard. Cet exploit, pour l'époque, a
marqué les débuts de l'alpinisme. Aujourd'hui, tout montagnard espère un jour conquérir le
plus haut sommet de la chaîne des Alpes.</P>
      </DESCRIPTION>
   </PRESENTATION>
   <RECIT>
      <SEC>
         <SOUS-TITRE>Le sommet :</SOUS-TITRE>
         <P>
            <LISTE>
               <ITEM>Nom : Mont-Blanc</ITEM>
               <ITEM>Altitude : 4808 mètres</ITEM>
               <ITEM>Cartographie : IGN 3531 ET et IGN 3630 OT</ITEM>
            </LISTE>
         </P>
      </SEC>
      <SEC>
         <SOUS-TITRE>L'itinéraire choisi :</SOUS-TITRE>
         <P>
            <LISTE>
               <ITEM>Nom : Arête de bosses, dite voie normale</ITEM>
               <ITEM>Type d'escalade : rocher, neige</ITEM>
               <ITEM>Difficulté : PD+</ITEM>
               <ITEM>Orientation : Ouest</ITEM>
               <ITEM>Altitude de départ : 2372 mètres</ITEM>
               <ITEM>Dénivelé positif : 2436 mètres</ITEM>
            </LISTE>
         </P>
      </SEC>
      <SEC>
         <SOUS-TITRE>Le parcours en 2 jours :</SOUS-TITRE>
         <P>1er jour : Les Houches (1008 m) - Bellevue (1801 m) - Nid d'Aigle (2372 m) - Tête
Rousse (3167 m) - Refuge du Goûter (3817 m)</P>
         <P>Depuis Les Houches, prendre le télécabine jusqu'à Bellevue pour poursuivre avec le
tramway du Mont-Blanc jusqu'au Nid d'Aigle. Puis suivre le sentier pour regagner le refuge de
la Tête Rousse. De là, remonter les pentes de neige pour gagner la rive droite du "Grand
Couloir". Traverser ce dernier (chutes de pierre fréquentes) et regagner les rochers rive
gauche. Remonter alors cet éperon par une succession de ressauts et d'éboulis moins raides
pour gagner le refuge du Goûter).</P>
         <P>2eme jour : Refuge du Goûter (3817 m) - Aiguille du Goûter (3863 m) - Dôme du
Goûter (4304 m) - Refuge du Vallot (4362 m) - Arrête des Bosses (4550 m) - Mont-Blanc (4808
m)</P>
         <P>Depuis le refuge du Goûter, gravir la calotte glacière haute d'une vingtaine de
mètres surplombant ce dernier pour rejoindre l'arête. Suivre cette arête pour rejoindre les
pentes ouest du dôme du Goûter. Les gravir et couper dans les pentes sommitales du dôme,
versant Grand Mulet, pour passer sous le sommet. Descendre environ une cinquantaine de mètre
de dénivellation et traverser le col du Goûter. Un certain nombre de piquets en bois
permettent de se repérer en cas de brouillard. Remonter sur 100 m les pentes ouest et
rejoindre le refuge Vallot. Gravir ensuite les ressauts successifs de l'arête des bosses
proprement dite. La première bosse est assez raide (35°) et parfois crevassée. La troisième
bosse présente une arête pouvant être effilée. Tracer alors sur sont flanc gauche. Gravir
ensuite deux murs et terminer l'ascension par une longue arête effilée menant au sommet.
Retour aux Houches le même itinéraire.</P>
      </SEC>
      …
   </RECIT>
</BALADE>
```

Annexe 3 : Exemple de fichier QRel

```
d001.xml        /BALADE[1]/PRESENTATION[1]/DESCRIPTION[1]/P[1]        0
d001.xml        /BALADE[1]/PRESENTATION[1]/DESCRIPTION[1]/P[2]        0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[1]                       0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[2]                       1
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[3]                       1
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[4]                       0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[5]                       0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[6]                       0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[7]                       0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[8]                       1
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[9]                       0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[10]                      0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[11]                      0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[12]                      0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[13]                      0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[14]                      0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[15]                      1
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[16]                      0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[17]                      1
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[18]                      0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[19]                      0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[20]                      0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[21]                      0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[22]                      0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[23]                      0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[24]                      0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[25]                      0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[26]                      0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[27]                      0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[28]                      0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[29]                      0
d001.xml        /BALADE[1]/RECIT[1]/SEC[1]/P[30]                      0
d002.xml        /BALADE[1]/PRESENTATION[1]/DESCRIPTION[1]/P[1]        0
d002.xml        /BALADE[1]/RECIT[1]/P[1]                              0
d002.xml        /BALADE[1]/RECIT[1]/P[2]                              0
d002.xml        /BALADE[1]/RECIT[1]/P[3]                              0
d002.xml        /BALADE[1]/RECIT[1]/P[4]                              0
d002.xml        /BALADE[1]/RECIT[1]/P[5]                              0
d002.xml        /BALADE[1]/RECIT[1]/P[6]                              0
d002.xml        /BALADE[1]/RECIT[1]/P[7]                              0
d002.xml        /BALADE[1]/RECIT[1]/P[8]                              0
d002.xml        /BALADE[1]/RECIT[1]/P[9]                              0
d002.xml        /BALADE[1]/RECIT[1]/P[10]                             0
d002.xml        /BALADE[1]/RECIT[1]/P[11]                             0
d002.xml        /BALADE[1]/RECIT[1]/P[12]                             0
d002.xml        /BALADE[1]/RECIT[1]/P[13]                             0
d002.xml        /BALADE[1]/RECIT[1]/P[14]                             0
d002.xml        /BALADE[1]/RECIT[1]/P[15]                             0
d002.xml        /BALADE[1]/RECIT[1]/P[16]                             0
```

La troisième colonne indique si l'élément est pertinent (1) ou non (0).