Techniques d’accès à l’information

Présentation

- Adrian.Chifu@irit.fr
- Thèse
  - Datamining et apprentissage l'adaptation des modèles de recherche
  - Allocation ministérielle de recherche + ATER UPS
  - Activité:
    - Laboratoire: IRIT
    - DCCE (Doctorant Contractuel Chargé d'Enseignement) + ATER
- Thématiques de recherche
  - Recherche d'information
  - Apprentissage
  - TAL (Traitement Automatique des Langues)
  - Bases de données

Organisation du cours

- Volume horaire
  - 16h CM/TD
  - 8h TP par groupe (discussion)
- Modalités de contrôle des connaissances
  - Examen/Evaluation de TP (30%)
  - Examen final (70%)

Introduction

- Objectifs du cours
  - Présenter les notions et concepts de TAI
    - Comment permettre à un individu d'accéder à l'information pertinente (répondant à ses besoins)?
  - Acquérir des bases pour accéder plus efficacement à l'information pertinente
    - Où et comment chercher?

Introduction (1/2)

- Thèmes abordés
  - Traitement documentaire
    - Quelques définitions
    - Représenter des documents (indexation)
    - Requêtes (besoin d'information)
    - Sélectionner les documents pertinents (recherche)
  - Outils d'accès à l'information
    - Moteur de recherche
    - Systèmes de recommandation
    - Outils de classification

Introduction (2/2)

- Thèmes abordés
  - Accéder à l'information
    - Méthodologie de recherche
    - Utilisation avancée des moteurs de recherche
  - Visibilité sur le Web
    - Bases du référencement
    - Exploiter les médias sociaux (plan professionnel et plan personnel)

Le Traitement Documentaire

Points abordés

- Quelques définitions
  - Notions: documents; collection; requête
  - L'accès à l'information
- Représenter les documents
  - Indexation manuelle
  - Indexation automatique
  - Stockage et représentation
- Sélectionner les documents pertinents

Documents, collections, requêtes

- Document (unité documentaire)
  - Unité minimale répondant aux besoins de l'usager
    - Article, page Web, cours, ligne, …
- Collection (fond documentaire)
  - Ensemble de documents
    - => Collection = { document }
- Requête
  - Demande, besoin d'information, mots clés
    - « prix Nokia Lumia », requêtes SQL, …
- ! Pas forcement du texte !

Types d'informations

- Distinguer le fond (contenu) et la forme
- Le contenu
  - Texte
  - Graphique
  - Image
  - Son
  - Vidéo
  - Référence, lien hypertexte
  - …

Types d'informations

- La forme
  - 3 types
    - structurée, non structurée, semi structurée
  - Structurée
    - Données factuelles sous forme tabulaire ou agrégative
    - Données élémentaires typées
    - Traite des questions précises
      - Langage strict
      - Peu lisible pour l'humain
      - Interprétable par un système informatique
    - Exemples
      - Bases de données relationnelles
      - Code source

Types d'informations

- C/C++:

```
int somme, a, b;
somme = a+b;
```

- Requête SQL:

```
CREATE TABLE facture (numero
Integer, date Date, montant
Double);
```

Types d'informations

- La forme
  - Non structurée
    - Autres informations
    - Structure
      - Non définie
      - Trop complexe pour être stockés dans une base de données
    - Exemple
      - Article de journal

Types d'informations

- La forme
  - Semi-structurée
    - Pas de schéma
    - Auto descriptives
      - Pas de séparation entre la donnée et son type
      - L'interprétation d'une donnée est faite indépendamment de toute autre information
    - Exemple
      - Document XML

Types d'informations

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<annuaire>

<personne class = "etudiant">
<nom>Pillou</nom>
<prenom>Jean-Francois</prenom> <telephone>555-
123456</telephone>
<email>jfpillou@toto.fr</email>
</personne>

<personne>
...
</personne>
</annuaire>
```

Types d'informations

- Les métadonnées
  - Informations complémentaires sur le document
    - Contexte
  - Exemples
    - Date de rédaction, de publication
    - Auteur(s)
    - Support de publication
    - …

Accès à l'information pertinente

- Permettre l'accès à l'information
  - Fournir des réponses pertinentes, même si elles sont approximatives voire incomplètes, à des requêtes (besoins d'information) exprimées en langage naturel

Représenter les documents

- Comment représenter les documents?

La CSG pour financer la dépendance ?

lefigaro.fr Mis à jour le 19/09/2012 à 08:36 | publié le 19/09/2012 à 08:01
Réactions (2)

J'aime 0 Tweeter 10 +1 Share Recommander

Le budget 2013 ne comportera «aucune disposition» visant spécifiquement les retraités, a promis Jean-Marc Ayrault ce matin sur RTL. Le premier ministre répondait à une question sur la suppression de l'abattement de 10% sur les revenus pour «frais professionnels», dont bénéficient aussi les retraités, et sur l 'alignement du taux de CSG payé par les retraités sur celui des actifs.

En revanche, Jean-Marc Ayrault a laissé entendre que ces mesures, **préconisées par la Cour des comptes,** pourraient être prises dans le cadre d'une réforme de la dépendance. «Le jour où nous ferons cette réforme - et nous le ferons - il faudra voir comment on la finance», a déclaré le chef du gouvernement, en appelant à la «solidarité intergénérationnelle».

Le processus d'indexation

- Construire une représentation numérique des documents
  - Extraire des caractéristiques sur le contenu sémantique
  - Phase primordiale pour la qualité de l'accès à l'information
  - Analyser le contenu textuel
    - Sélection de termes significatifs/représentatifs du contenu (contenu sémantique)
    - => Langage d'indexation

Le processus d'indexation

- Langage d'indexation
  - Libre
    - Construit au moment de l'indexation
    - Termes directement extraits des textes
    - Peut inclure toute la variété du langage naturel
  - Contrôlé
    - Construit généralement avant la phase d'indexation
    - Termes prédéfinis et limités
  - Mixte

Indexation manuelle

- Réalisée par des documentalistes
  - Experts
  - Caractériser au mieux les idées contenues dans les documents
  - Effort cognitif et intellectuel important
- Indexation
  - Performante
  - Mais subjective
    - Dépend des connaissances des documentalistes
  - Très couteuse

Indexation automatique

- Indexation
  - Applicable à de grandes collections
  - Rapide
  - Limite les représentations des documents aux « entrées » (termes) utiles

Indexation automatique

- Vue globale de l'indexation automatique
  - 3 étapes
    - Extraction des mots simples
    - Normalisation des mots extraits
      - Troncature
      - Lemmatisation
      - Radicalisation
    - Pondération des mots normalisés

Indexation automatique

- Segmenter le texte
  - Extraire les mots
    - Mots simples
    - Groupe de mots ou de symboles (idéogrammes)
  - Eliminer les signes de ponctuation

    Le budget 2013 ne comportera « aucune disposition » visant spécifiquement les retraités a promis Jean-Marc Ayrault
    - => Texte = { mot }

Le	budget	2013	ne	comportera	aucune	disposition	visant	spécifiquement

les	retraités	a	promis	Jean-Marc	Ayrault

Indexation automatique

- Eliminer les mots-vides
  - Anti dictionnaire
    - Mots n'ayant qu'un rôle syntaxique (pas de sens)
    - Mots ordinaires ou athématique
    - Exemples
      - « de », « à » en français
      - En anglais, liste de plus de 275 mots

budget	2013	comportera	aucune	disposition	visant	spécifiquement

retraités	promis	Jean-Marc	Ayrault

Indexation automatique

- Normaliser les mots extraits
  - Désaccentuer

Le	budget	2013	ne	comportera	aucune	disposition	visant	specifiquement

les	retraites	a	promis	Jean-Marc	Ayrault

  - Minusculiser

le	budget	2013	ne	comportera	aucune	disposition	visant	spécifiquement

les	retraités	a	promis	jean-marc	ayrault

Indexation automatique

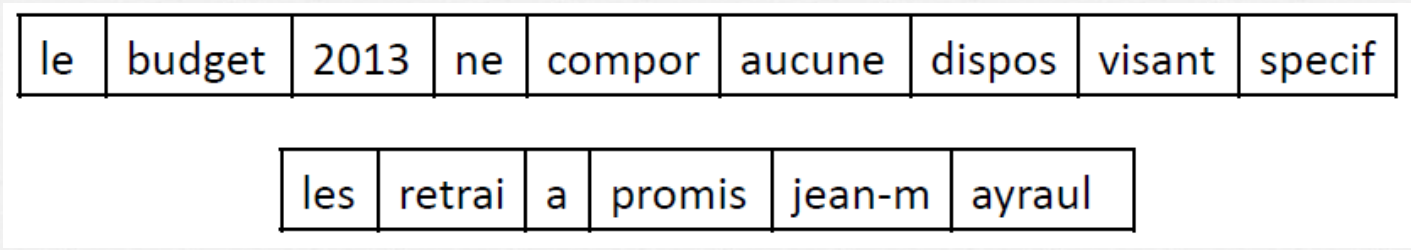
- Lemmatisation
  - Verbe: ce verbe à l'infinitif
  - Autre mot: le mot au masculin singulier
    - Petit, petite, petits, petites => petit
    - Avoir, ai, a, avons, eu, eussions => avoir

budget	2013	comporter	disposition	viser	specifique

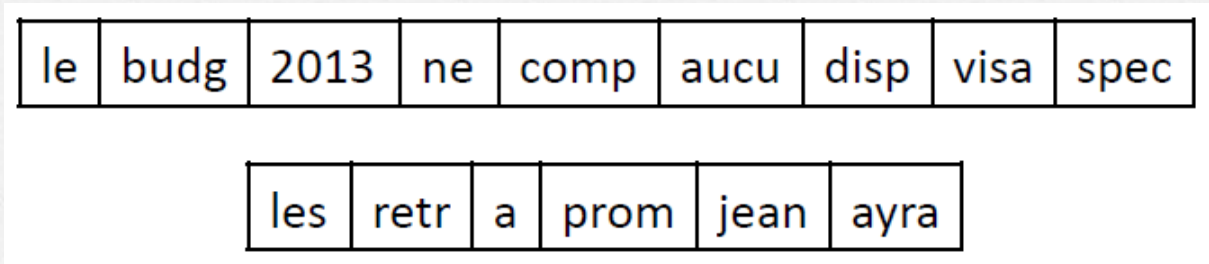
retraite	avoir	promettre	jean-marc	ayrault

Indexation automatique

- Troncature
  - Couper le mot au-delà d'un certain nombre de caractères
    - Troncature à 6



    - Troncature à 4



Indexation automatique

- Radicalisation
  - Forme morphologique d'un mot

Préfixe	Radical	Désinence

  - Radicalisation

Le	budget	2013	ne	comport	aucun	disposit	vis	spécifiqu

les	retraité	a	promi	Jean-Marc	Ayrault

Indexation automatique

- Conséquences liées à la normalisation