Open semantic : Portail sémantique communautaire sur l'open data

Idée de projet
auteur : François Régnier
date : 25/06/13
nom : Open semantic

Résumé

Description courte : Portail sémantique communautaire sur l'open data.

1. Associer les catalogues dewey et les catalogues de données ouvertes
    1. On réutilise et on participe au travail de catalogage des sujets ou notions
    2. Utiliser les classements existants comme base d'apprentissage et de référentiel
2. Construire collectivement une interface :
    1. d'indexation dynamique, exploration des sources:
        1. ateliers de mécanismes d'indexation
        2. ateliers de présentation des résultats
        3. ateliers de connecteurs d'extractions des données depuis les sources (adaptation aux formats de stockage)
    2. catalogues sémantiques, associer un ou des sens à un contenu
    3. lexique des catégories de sens ou notions
3. Exploiter directement les données de data.gouv

http://fr.wikipedia.org/wiki/S%C3%A9mantique
http://fr.wikipedia.org/wiki/Classification_d%C3%A9cimale_de_Dewey

Sommaire

Table des matières

Projet de service

Description détaillée

- Portail sémantique open data (Open semantic)
  - Générateur de relations sémantiques
  - Usine de robot d'analyse sémantique
  - Index de sources de données
  - Index de recherche et de présentation sémantique
- Constats
  - Les annuaires et index de données sont nécessairement succincts en catégorisation et en explication
  - Leur contenu est très variable en forme et qualité
  - La recherche impose un rapatriement et une exploitation coûteuse en temps du fait des échecs répétés avant de trouver le bon document ou la bonne source.
  - L'archivage des versions des données est rare et ajoute une hétérogène de génération technologique, l'étude par comparaison et la recherche de tendances et de statistiques temporelles est extrêmement complexe.
  - La classification dewey utilisée pour le classement des ouvrages offre deux assurances :
    - elle permet de puiser dans un volume très important et en constant évolution de catalogage documentaire. Ce faisant elle définit des ensemble d'apprentissage sur chacun de ses codes et de la notion qu'il représente ;
    - elle constitue un référentiel commun normalisé et stable afin de rendre durable l'évolution et la transmission des travaux de catégorisation.
  - L'apprentissage et la transmission des catalogages est indépendante des systèmes de référentiel, ce qui perdure et est réellement utilisé est la relation entre une notion et un ensemble de documents traitant de cette notion. Le sens, la notion est un ensemble dynamique de relations liées en et à un contexte culturel. Ce graphe participe à la définition du contexte culturel autant qu'il est défini par lui.
- Besoins
  - Outil de présentation automatisé
    - Analyseur de contenu et de forme
    - Indexeur partagé
  - Outil d'extraction automatisé
    - Échantillonneur pour analyse de contenu
    - Interfaces / connecteur multi format
  - Outil de recherche et d'agrégation
    - Veille et acquisition des sources
    - Supervision de l'usage des données pour estimation de l'intérêt d'une source
    - Constitution de réseaux de diffusion de recherche par voisinage sémantique
  - Système d'archivage et de description des transformations
    - Outil de recherche et d'agrégation en profondeur sur les sources
    - Conservation d'une synthèse et d'un échantillon des versions sur une source
    - Outil de présentation et d'alerte sur la modification des données et des sources
  - Supervision communautaire des outils de présentation et d'extraction
    - Mise au point progressive des interfaces / connecteurs
    - idem des analyseurs
    - dynamisation des contributions par une visibilité sur leurs usages, résultats et retour
    - dynamisation des usages par une visibilité et un suivi des traitements et réponses aux

demandes et remontés

- Présentation polymorphe fusionnelle (convergence visuelle)
  - Présentation de données hétérogènes sous des formes synthétiques (paramétrage, contexte et alternatives en annexe)
  - Adaptation et présentation de données sous forme variées et contextualisées
  - Présentation multidimensionnelles dynamique
    - élastiques
    - projections sphériques
- Agrégations d'historiques d'information
  - recherche temporelle d'une information
  - accumulation d'archivage
- Services et dérivés métiers
  - Urbanisme
    - veille sur les demandes et publications de projets
    - agrégation des news, blogues et autres e émergences
    - superposition des géométries urbaines
      - voirie
      - équipement
      - cadastre
      - statistiques d'usage
        - transports
        - accidents
        - immatriculation
      - autorisations d'usage de la voirie
        - à des fins privées
        - à des fins communautaires
        - en revendication
    - synthèses et projections d'analyse statistique
      - temporelles
      - géographiques
  - Territoriale
  - Communauté d'usager / d'intérêt
  - Sociétales
    - données publiques de qualité de vie
      - revenus...
    - événement publiques ou publiés visibles sur le net
      - rss
      - robots d'agrégation
      - déclaration spontanées dans un système mixte
        - pose d'info sur la carte (appli web/mobile)
        - boîtes mail robots...

Principes et contexte

Éléments de réflexion, éclairage sur la sémantique dans notre contexte

Contexte :

Forme et contours sémantiques :

Trois niveaux de reconnaissance de forme sémantique[1]
(reconnaissance de forme : analyse automatique, nous parlons ici de méthode de classement par appartenance, proximité ou relation)

- Catégorisation sémantique :
  - Indication interne claire d'appartenance à une idée et une identité domaniale *(contenu)*
  - Indication par des données fortement structurées (attache à un domaine...) *(forme)*
  - Catalogage ou identification volontaire *(intervention)*
- Étiquetage sémantique :
  - Reconnaissance par rapprochement (relation lexicale forte aux thésaurus)
  - Identification catégorielle (catégorisation forte)
  - Probabilité réaliste (évaluation forte)
- Évaluation sémantique :
  - statistique lexicale correspondant à une indexation domaniale par thésaurus
  - contextualisation (flot, conversation, origine, documents liés, analyse de réputation)
  - analyse de forme et d'entités identifiables (personnes, citations, événements relatés, lieux)
  - analyse de structure et de nommage, catégorisation des noms et formats de champs, rapprochement à des conventions de nommages, des systèmes d'indexation

1 Sémantique ensembliste à des fins d'indexation globales d'une source, d'un document, d'un enregistrement. Nous abordons la sémantique pour répondre à un besoin de classement et de rapprochement des données. Extraire une notion de domaine et d'idées maîtresses du fond et de la forme d'un volume de données. Il ne s'agit pas de traduction précise et d'analyse des subtilités d'un discours, il s'agit de reconnaître des schémas, des notions permettant d'étiqueter le sens d'un élément de nos ensembles.

Éléments techniques

Clés

Évaluateurs

- de nommage
  - strict : nom/type
  - lâche : schéma de nommage
    - segments de mot clé
    - forme d'agrégation de mot (ex : séparateurs « _ » majuscule en début de mot..., mot spécifique à un domaine : rue, égal, =, année lumière, cm...)
  - groupé :
    - strict : liste de nom/type
    - strict et suffisant : représentation n/m d'un ensemble
    - lâche : association de champs lâches corrélés à n/m dans un ensemble
- de type
  - strict : exemple géométrie SIG
  - combinaison n/m dans un ensemble de forme
- de contenu
- mixtes

Apprentissage

- phases
  - initiale : import de bases de références bibliographiques et autres sources validées d'éléments clés de catégorisation et d'indexation
  - continue :
    - import régulier des catalogues officiels
    - apparition d'autorité domaniale et réutilisation de leurs références avec un poids fort
    - remonté d'erreur et modification des poids d'autorité et de validité
    - affinage des espaces mitoyens entre notions et domaines
    - élagage et constitution de zones intermédiaires entre espaces de notions
- usage
  - présentation des schémas d'évolution :
    - temporelles
    - événementiels
  - auto complétion et auto apprentissage
- évolution
  - supervisée
    - cycles :
      - libre
      - contrôle
      - rééquilibrage supervisé
      - archivage de test pour passage au mode autonome
  - autonome
    - cycles :
      - libre
      - contrôle
        - sur retours d'échec (notification claire d'erreur ou d'absence de réponse)

        - sur estimation de circuit d'échec (reprise de question avec retour en arrière sur affinage sans situation de réponse claire)
    - rééquilibrage aléatoire sur les éléments en échec

Cache

- schémas de relation sémantique
    - http://archives.limsi.fr/Individu/habert/Cours/PX/BHabertIntroductionSemantique0102/BHabertIntroductionSemantique0102.html
    - Ensembles
        - appartenance
        - exclusion
        - proximité
        - union

Architecture logicielle

Phases de traitement

Éléments de la maquette fonctionnelle

Modules

- « scrapping » : extraction des données
  - extracteur d'indexes et de sommaires[2] : recensement des sources et des catalogues de données, sous forme pseudo normalisée à des fins d'agrégation et de recherches croisées
  - extracteur de données brutes[3] : pour la livraison finale[4], pour le traitement statistique[5] et pour la catégorisation avancée[6]
    - formats de support spécifique ex : xls-multi-pages vers csv et json
    - formats de contenu spécifique ex : champs structurés identifiables (auteur, citation, époque, lieu, bilan comptable, planification...)
- statistique mots et expressions
- indexation et sommaires
  - Bases/tables de clé/valeur permettant d'utiliser au mieux les capacités des caches et servir des outils de présentation avec des données json. L'indexation réelle est repoussée vers le contexte d'interface pour agréger les personnalisations et les données hétérogènes proposées.
  - Bases/tables d'agrégation résultant de l'apprentissage contextuel redescendu depuis les interfaces utilisateurs contextualisées.
  - Bases/tables de proposition en émergences produites par les constitutions de thésaurus et les agrégateurs[7] sémantiques, contextuels ou structurels
- archivage et suivi de modification
- présentation

---

2 Urlib, BeautifulSoup...

3 Catdoc xls2csv: http://www.maketecheasier.com/convert-xls-file-to-csv-in-command-line/2012/02/03, pdfedit...

4 Livraison finale au client/utilisateur ayant effectué une recherche suivi d'une sélection parmi les réponses

5 Traitement statistique permettant de compléter la description du document par un comptage brute puis par une catégorisation et une analyse extraction de thésaurus constitué des mots et expressions spécifiques au sujet et au traitement particulier du sujet dans le document.

6 Catégorisation avancée au delà des champs structurés, citations et analogie de thésaurus de la description originelle du document. Cf chapitre catégorisation avancée et ateliers de catégorisation.

7 Agrégateurs de constitution de relation entre les données ou les sources de données. L'agrégation sémantique se ferait par une multitude de voies liées aux statistiques de mots et étiquettes, elle se ferait aussi par l'exploitation des relation de flux comme des conversations, des références bibliographiques et autres listes identifiables pour un contexte de sujet ou de notion. L'agrégation de contexte est une réduction de la précédente en la privant des étiquetages de catégories sémantiques, on s'arrête au fait comme l'appartenance à un même auteur, un indexe. L'agrégation structurel est encore plus précise, c'est l'utilisation d'un lien direct ou un fil de liens présentant dans les données.