

Quelle évaluation pour la Recherche d'Information Sémantique

Haïfa Zargayouna

LIPN, Université Paris 13
haifa.zargayouna@lipn.univ-paris13.fr
<http://www.lipn.univ-paris13.fr/~zargayouna/>

Atelier RISE@CORIA
15 mars 2011

Historique (le mien !)

- 2005 [Zargayouna, 2005] : évaluation RIS avec un petit corpus CACM
 - résultats : ...
- à partir de 2008 : évaluation de l'acquisition de ressources sémantiques (terminologies, ontologies)
 - Quel protocole? quelles métriques pour évaluer l'acquisition?
[Zargayouna & Nazarenko, 2010]

Questions d'évaluation

- Pourquoi évaluer ?
- Qu'est ce qu'on évalue ?
- Comment évaluer ?
- Qui évalue ?

1 Tour d'horizon

- Web Sémantique
- Recherche Information Sémantique

2 Méthodologie

- Types d'évaluation
- Amorce de benchmark : Cooking Contest Data
- Suite Benchmark

LUBM

- LUBM : A Benchmark for OWL Knowledge Base Systems [Guo et al., 2004]
 - Quantité de données : passage à l'échelle et performances.
 - Capacité de raisonnements nécessaires pour répondre aux exigences de sémantique des applications.
 - Prendre en compte des requêtes extensionnelles
 - Volume de données (A-Box) variable
 - Des ontologies de complexité et de taille modérée.
 - Des requêtes en SPARQL
 - Critères d'évaluation : Temps de chargement, Taille, Temps de réponse, "completeness" et "soundness" (rappel/précision)
- Voir aussi The Berlin SPARQL Benchmark [Bizer & Schultz, 2009]

Collections

- Les campagnes existantes :
 - ImageCLEFmed-2005 [Radhouani, 2008]
 - Les tracks de TREC : Chemical IR Track, Legal Track, Medical Track
- Les petites collections : ADI (science de l'information), TIME (tous sujets), CISI (library science), MED (médical), CACM (informatique) [Seydoux, 2006]

Chemical Track@TREC

- Deux tâches : **Technical Survey Task (TS)**, Prior art Search Task [Lupu et al., 2009]
- Collection de documents : 1 185 012 brevets + 59 000 articles scientifiques
- Requêtes : 18 requêtes proposées par des experts en brevets
- Jugement de pertinence : 300 documents par requête
 - En deux phases : 1) Deux étudiants évaluent la pertinence (pertinence graduée : -2 "unsure" à 2 "highly relevant"), 2) Présentation aux évaluateurs experts
 - L'analyse du désaccord entre étudiants (jusqu'à 90%!) a permis de déterminer les requêtes difficiles ou ambiguës.
- 8 participants avec des méthodes allant de RI classique à des méthodes spécifiques à la chimie (extraction d'entités nommées et synonymes..).
- Meilleur résultat : RI classique en faisant varier les schémas de pondération ...

Réutilisation de TREC pour la RIS

- Né du besoin d'une évaluation spécifique de la Recherche d'Information Sémantique [2009]
 - Collection de documents : TREC WT10G
 - Requêtes et jugements de pertinence
 - TREC 9 et TREC 2001 corpus de test (100 requêtes avec leurs jugements de pertinence)
 - 20 requêtes ont été sélectionné et adapté pour être utilisé par PowerAqua (le module de QR proposé)[2011]
 - Ontologies
 - 40 ontologies publiques qui couvrent un sous-ensemble des domaines TREC et des requêtes
 - 100 repositories (2GB de RDF et OWL) stockés et indexés avec un portail WS
 - Base de connaissances : Parmi les 40 ontologies sélectionnées, quelques unes enrichies semi-automatiquement à partir de Wikipedia
- Requêtes disponibles en ligne
- Disponibilité des jugements de pertinence? (à voir)

Types d'évaluation

Selon le but de l'évaluation

- Comparative : permet de comparer différents systèmes sur une tâche définie (ou application).
- Selon des caractéristiques intrinsèques : permet d'établir un certain nombre de critères pour un système donné.

Cycle de vie

Itérations des trois phases [Euzenat et al., 2005] :

- Spécifications *Plan phase*
- Mise en place *Experiment phase*
- Amélioration et recalibrage *Improve phase*

Un premier travail

Objectifs [Despres & Zargayouna, 2009]

- Construire une base de tests
- Evaluation des résultats obtenus par le système relativement à la base de tests
- Recommandations
 - Impact des ressources sur les résultats
 - Suggestions d'évolution des ressources

Données

- Corpus de recettes (1 489 recettes)
 - Titre de la recette
 - Ingrédients
 - Préparation
- Ensemble de requêtes test (5)
 - Liste de termes
 - Indication sur les éléments à privilégier **Main Focus** : « *type of meal and type of cuisine* »
 - Ingrédient
 - Type de plat
 - Type de cuisine
 - Régime
 - Type de préparation
 - Combinaison des précédents éléments

Constitution du gold standard (qrels)

- Sélection guidée par les connaissances
 - Issues de nos propres expériences
 - Extraites des sites relatifs à la cuisine, régime alimentaire, diététique, etc.
 - Issues d'encyclopédies en ligne
- Identification d'éléments dans la requête (e.g. *Asian Soup with Leek*)
 - Élément Central (EC) : Asian Soup
 - Élément Périphérique (EP) : Leek
- Actions menant à la sélection
 - Spécialisation de l'un des éléments (EC et/ou EP)
 - Modification de l'un des EP
 - Substitution par un EP
 - Ajout d'un EP
 - Retrait d'un EP

Jugement de pertinence

Recherche porte sur « tarte aux myrtilles »

- Caractérisation de la requête
 - EC : tarte
 - EC : myrtille
- Comment juger de la pertinence
 - **tarte** aux fraises
 - confiture de **myrtilles**

Jugement de pertinence gradué

Échelle de jugement **graduée** de 0 à 4

- TP : recette totalement pertinente
- PP : recette partiellement pertinente

EC	EP	Pertinence
N	N	0
N	PP ou TP	0
PP	N	1
PP	PP ou TP	2
TP	N	2
TP	PP	3
TP	TP	4

Enrichissement du benchmark

- Ajout de deux requêtes :
 - Prepare a cake with plum
 - I would like fruit salad with **kiwano**

Recette (id : titre)	Pert.	éléments explicatifs
1322 : <i>Summer fruit bowl</i>	3	Salade de fruit avec melon (Watermelon et cantaloupe)
1467 : <i>Winter fruit bowl</i>	3	Salade de fruit avec melon (Honeydew ou cantaloupe)
388 : <i>Cranberry fruit salad</i>	2	Salade de fruit sans melon
544 : <i>Festive fruit salad</i>	2	Salade de fruit sans melon
585 : <i>Frozen fruit salad</i>	2	Salade de fruit sans melon
904 : <i>Molded waldorf salad</i>	2	Salade de fruit sans melon
576 : <i>Fresh watermelon salsa</i>	2	Melon (en fait pastèque) mais pas salade de fruit
1447 : <i>Watermelon Punch</i>	2	Melon (en fait pastèque) mais pas salade de fruit

Bilan

- Constitution d'un "petit benchmark"
 - ouvert
 - collaboratif
- Ensemble de requêtes en fonction des fonctionnalités à évaluer (à étoffer)
- Tracer le jugement de pertinence (!)

Suite

- réutiliser les données d'autres campagnes (TREC par exemple)
- besoin de multiplier les expériences
- isoler l'impact des modules
 - impact de la ressource
 - impact de l'annotation sémantique
 - impact de la prise en compte des relations

Biblio

- [Bizer & Schultz, 2009] C. Bizer, A. Schultz *The Berlin SPARQL Benchmark* In : International Journal on Semantic Web & Information Systems, Vol. 5, Issue 2, Pages 1-24, 2009.
- [Despres & Zargayouna, 2009] S. Despres, H. Zargayouna *Evaluation of knowledge based applications : benchmark and guidelines* In Second International Workshop on Knowledge Acquisition, Reuse and Evaluation (KARE 2009)
- [Euzenat et al., 2005] J. Euzenat, M. Ehrig, R. Garcia Castro *Towards a methodology for evaluating alignment and matching algorithms* OAEI 2005
- [Guo et al., 2004] Y. Guo, Z. Pan, and J. Hefflin. *An Evaluation of Knowledge Base Systems for Large OWL Datasets* In Proc. of the 3rd International Semantic Web Conference (ISWC2004), 2004.
- [Lupu et al., 2009] M. Lupu, F. Piroi, J. Tait, J. Huang, J. Zhu *Overview of the TREC 2009 Chemical IR Track* TREC 2009
- [Radhouani, 2008] S. Radhouani *Un modèle de Recherche d'Information orienté précision fondé sur les dimensions de domaine* Thèse, University of Geneva, Geneva, Switzerland, Joseph Fourier University, 2008.
- [Seydoux, 2006] F. Seydoux *Exploitation de connaissances sémantiques externes dans les représentations vectorielles en recherche documentaire* Thèse EPFL, 2006.
- [Zargayouna, 2005] H. Zargayouna *Indexation sémantique de documents XML* Thèse, Université Paris-Sud, 2005.
- [Zargayouna & Nazarenko, 2010] H. Zargayouna, A. Nazarenko *Evaluation of Textual Knowledge Acquisition Tools : a Challenging Task* In the seventh international conference on Language Resources and Evaluation (LREC), 2010.