Université
franco-allemande
Deutsch-Französische
Hochschule

Semantic Snippets via Query-Biased Ranking of Linked Data Entities

Mazen ALSAREM

Résumé

Dans notre société fondée sur la connaissance, l'acquisition et le transfert de connaissances jouent un rôle principal. Les moteurs de recherche sur le Web sont en quelque sorte des outils d'acquisition et de transfert des connaissances du Web à l'utilisateur. La page de résultats d'un moteur de recherche (Search Engine Results Page - SERP) se compose principalement d'une liste de liens et de snippets (extraits à partir des résultats). Les snippets sont utilisés pour exprimer, aussi efficacement que possible, la façon dont une page Web peut être pertinente pour la requête.

Le Web sémantique ou "Web 3.0" est conçu pour transformer le Web de documents non structurés en un Web de données exploitable à la fois par les machines et les humains. Le Web de données obtenu et le Web de documents actuel coexistent et sont interconnectés via de multiples mécanismes, tels que les données structurées intégrées dan les pages Web, ou l'annotation automatique.

Dans cette thèse, nous introduisons un nouvel artefact interactif pour le SERP : le "Snippet Sémantique". Les snippets sémantiques s'appuient sur la coexistence des deux Webs pour faciliter le transfert des connaissances aux utilisateurs grâce à une contextualisation sémantique du besoin d'information de l'utilisateur. Ils font apparaître les relations entre le besoin d'information et les entités les plus pertinentes présentes dans la page Web.

La génération des snippets sémantiques repose principalement sur l'annotation automatique des entités de LOD dans les pages Web. Les entités annotées ont des niveaux d'importance, d'utilité et de pertinence différents. Les solutions de l'état de l'art pour l'annotation automatique des entités LOD dans les pages Web génèrent encore beau- coup de bruit sous la forme d'annotations erronées ou hors sujet. Par conséquent, nous proposons un algorithme biaisé-requête (LDRANK) pour l'ordonnancement de ces entités. LDRANK adopte une stratégie basée

sur la combinaison consensuelle linéaire de plusieurs sources de connaissances a priori (toute forme de connaissances contextuelles, comme les descriptions textuelles des nœuds du graphe) pour modifier un algorithme de type PageRank. Pour générer des snippets sémantiques, nous utilisons LDRANK pour trouver les entités les plus pertinentes dans la page Web. Ensuite, nous employons un algorithme d'apprentissage supervisé pour lier chaque entité sélectionnée à des extraits de la page Web qui mettent évidence la relation entre l'entité et le besoin d'information original.

Afin d'évaluer nos snippets sémantiques, nous les intégrons dans ENsEN (Enhanced Search Engine), un système logiciel qui améliore le SERP avec des snippets sémantiques. Enfin, nous utilisons le crowdsourcing pour évaluer l'utilité et l'efficacité de ENsEN.

Mots-clés: Semantic Snippets, Ordonnancement d'entités, Web de Données.