La représentation du contenu et l'extraction des connaissances à partir de textes : l'apprentissage automatique de langue, les systèmes de gestion de graphes, l'apprentissage automatique pour l'analyse des graphes et les approches Web sémantique

PANEL DOING - ROCED

Modératrices :
Catherine Roussey, INRAE
Genoveva Vargas-Solar, CNRS
Mirian Halfeld Ferrari, Université d'Orléans

Participant.e.s :
Donatello Conte (*Polytech Tours, LIFAT*)
Nathalie Hernandez (*Université de Toulouse, IRIT*)
Agata Savary (*Université Paris-Saclay, LISN*)
Nicolas Travers (*ELSIV, Centre de Recherche Da Vinci*)

Contexte

L'analyse de texte consiste à traiter des textes afin d'en extraire des faits lisibles par une machine. L'objectif de l'analyse de texte est de créer des données structurées à partir de contenus textuels libres. On peut considérer que le processus consiste à découper des tas de documents non structurés et hétérogènes en éléments de données faciles à gérer et à interpréter. Du point de vue informatique, ce processus demande l'utilisation de techniques allant du traitement semi-automatique de langue pour décoder l'ambiguïté du langage humain, la représentation de connaissances, la détection des modèles et des tendances représentant des connaissances véhiculées dans les textes (interrogation au sens large).

Le contenu des textes définit à la fois un maillage syntaxique et un maillage de concepts qui peuvent être structurés sous forme de graphes. L'interrogation de ces contenus peut être envisagée à plusieurs niveaux selon le degré d'abstraction représentée par ce graphe, allant d'un simple guidage sur le texte brut, à la construction d'une base de données (graphe) respectant un certain nombre de contraintes structurelles.

Selon les caractéristiques des graphes utilisés, il est possible d'interroger le contenu des textes de différentes manières : la recherche d'information nous offre des méthodes par mots-clés ; l'interrogation des graphes via de langages de requêtes permet de trouver des patrons structurels et de calculer des agrégations ; les méthodes d'apprentissage automatique et fouille de données nous font découvrir des motifs ; l'intelligence artificielle rend possible la découverte des nouveaux liens sémantiques entre les concepts ... L'analyse de textes et l'extraction de connaissances sont donc abordées de différents points de vue peut être complémentaires (?) selon les objectifs d'exploitation à travers des applications.

Dans ce panel nous souhaitons discuter sur :

(i) La façon dont le contenu des textes est extrait et représenté par des modèles de données différentes comme les matrices avec des fréquences de termes, des ontologies, par la conception de bases de données à graphs.

(ii) Les implications du choix de modélisation sur les possibilités d'interrogation, de mise à jour et découverte de connaissances.

La discussion s'organise autour d'une question *leitmotiv* :

Quelle est la place de la modélisation/interrogation du contenu des textes dans le traitement automatique de langue, le Web sémantique, la conception des BD pour les SGBD à Graphes, et l'apprentissage automatique/fouille de données ?

Déroulement du panel

Après une première table ronde où les panélistes partageront leurs points de vue préparés à partir d'un ensemble de questions provocatrices, nous ouvrirons une discussion pour faire le suivi de certaines questions sur l**a représentation du contenu et l'extraction des connaissances à partir de textes**. Enfin, nous ouvrirons le débat aux questions du public. Pour clore le panel, nous vous demanderons de conclure sur les perspectives de recherche.

--

Questions provocatrices – Vous pouvez choisir de développer sur une ou plusieurs questions.

1. Quelle est la place de la modélisation/interrogation du contenu des textes dans le Web sémantique, la conception des BD Graphes et l'apprentissage automatique ou la fouille de données ? Quels verrous et défis ? Applications phares ? Expériences qui ont été représentatives ou qui ont posé des défis particuliers ?

2. Comment traiter les informations manquantes, c-à-d, *des annotations incomplètes ou de mauvaise qualité ou contradictoires* sur les graphes de propriétés et/ou ontologies représentant le contenu des textes ? Quels impacts dans l'interrogation et la maintenance ? Quels verrous y a-t-il à relever ?

3. Maintenance et mise à jour des représentations du contenu des textes sous forme de graphes de propriétés ou des ontologies :
    - Comment faire évoluer les annotations en fonction de l'évolutions des référentiels, des ontologies et des corpus ?
    - Comment valider les nouvelles connaissances déduites, découvertes ou insérées explicitement ?

4. Quelles sont les possibilités d'interrogation sur les différentes représentations du contenu de texte ?
    - Comment sont-elles exploitées dans le cadre des applications basées sur le TAL ?
    - Comment adresser l'interrogation au sens « analytics » (requêtes data science avec des opération découverte de communautés, « centrality », découverte de liens) sur les graphes ?
    - Quelles sont les solutions proposées par les SGBD à graphes ? Comment les positionner par rapport à aux solutions du style ML pipelines/DS pipelines ?
    - Est qu'il est pertinent de penser à un langage de requête SPARQL-étendue où l'utilisateur peut inclure une 'fonction' concernant l'analyse de données d'un graphe ? Quels avantages ? Quelle utilisation ? Quels défis ?