Titre

Anonyme

Anonyme

Résumé Avec le développement du Web, de nombreuses informations se retrouvent enfouies dans différents serveurs et l'utilisation de moteurs de recherche n'offre qu'une vision générale ou trop précise. Par exemple, rechercher les experts en fouille de texte en France via un moteur n'est pas une tâche aisée. De la même manière, une requête sur un chercheur spécifique va permettre de savoir qu'il est intéressé par la thématique de la *fouille de textes* mais qu'en est-il des autres membres de son laboratoire ? Quelles sont les compétences générales du laboratoire ? Quels sont les autres chercheurs de ce domaine ? Dans cette démonstration, nous proposons une approche implantée et et expérimentée permettant de cartographier les différentes compétences des chercheurs des laboratoires.

1 Introduction

Cartographier les informations contenues dans un site Web offre de nombreuses perspectives : indexation de données, regroupement de compétences similaires, recherche d'informations spécifiques, etc. Dans cet article nous nous intéressons plus particulièrement à la cartographie de sites web dans un contexte de caractérisation de compétences d'expert [1]. L'objectif principal de cette démonstration est de montrer comment, à partir des données disponibles sur Internet, il est possible de caractériser le mieux possible les compétences des différents chercheurs de laboratoire. De manière à illustrer notre problématique, considérons l'exemple suivant : dans le cadre d'un de ses projets, une entreprise souhaite collaborer avec des spécialistes de fouilles de données. Traditionnellement, l'entreprise va rechercher via un moteur de recherche quelles sont les personnes qui possèdent de telles compétences. Pour cela, elle effectuera différentes requêtes du type [2] : "*fouille de données*", "*fouille de données +expert*", "*fouille de données +expert +Paris*", etc. Malheureusement les résultats retournés par les moteurs sont trop généraux et ne permettent pas de retrouver les véritables experts du domaine qui correspondent à ce que souhaite l'entreprise. Dans cette démonstration nous montrons, d'une part, comment retrouver les différents experts du domaine avec peu de connaissances préalables et d'autre part, comment mettre en exergue les

1. Les recherches présentées dans cet article sont menées conjointement par la société Expernova (http ://www.expernova.com/fr/) et des équipes de recherche du LIRMM (http ://www.lirmm.fr).

2. Les requêtes présentées utilisent la syntaxe de Google mais peuvent tout à fait être adaptées à d'autres moteurs.

mots-clés les plus discriminants à différents niveaux (chercheurs, laboratoire, etc).

L'article est organisé de la manière suivante. Dans la section 2, nous présentons brièvement comment retrouver les différents experts d'un domaine donné et présentons les différentes stratégies utilisées. La section 3 présente de quelle manière ces mesures ont été utilisées dans notre contexte et de quelles manière elles ont été validées. Le système développé est présenté en section 4. Enfin, nous concluons en présentant les travaux que nous menons actuellement.

2 Comment trouver des experts ?

Même si de nombreuses données sont disponibles sur Internet, il est difficile de trouver aux travers des documents des connaissances sur les différents experts. Une approche intuitive est sans doute de rechercher le nom d'un expert via un moteur de recherche. Cependant le problème principal est que malheureusement le nom de cet expert n'est pas forcément connu à l'avance. Dans le cadre de notre prototype, notre objectif est d'extraire via différentes sources d'information les experts potentiels. Pour cela nous utilisons différentes sources disponibles comme :

- *Archives ouvertes.* Le développement des archives ouvertes (e.g. HAL [3]) offre de nouvelles perspectives dans la mesure où il devient possible de connaître pour un chercheur les différentes personnes avec qui il a travaillé mais également de nombreuses informations sur le laboratoire, les dates de publications, le contenu du document.
- *Les conférences.* En analysant les différentes conférences, de nombreuses informations peuvent être extraites. Par exemple, dans le cas d'un conférencier invité il est possible d'extraire, via les résumés, de nombreuses informations sur sa thématique de prédilection. En outre la liste des articles présentées offre de nouvelles opportunités pour rechercher de nouveaux experts pour lesquels nous connaissons déjà les thématiques principales de recherche (i.e. la thématique de la conférence).
- *Le nom des laboratoires.* Issue des sources précédentes ou de différentes sources, nous utilisons cette information pour extraire les différents chercheurs des laboratoires et nous focalisons principalement sur leur propre page web pour récupérer des informations utiles pour mieux caractériser leur compétence.
- *Le nom des chercheurs.* En utilisant des ressources variées, il devient possible de recueillir, de manière automatique, des informations sur les différents chercheurs (e.g. laboratoire, page personnelle, ...). Via ces informations, une analyse fine des données offre de nouvelles opportunités pour enrichir les connaissances et compétences d'un expert.
- ...

L'analyse de toutes ces sources d'information nécessite, bien entendu, de nombreux prétraitements dans la mesure où les données sont disponibles de manière très hétérogènes.

3. http ://hal.archives-ouvertes.fr/

Pour les différents tâches énumérées ci-dessus, nous avons utilisé de nombreuses heuristiques mais également différents patrons obtenus soit par expertise soit par apprentissage à partir d'un ensemble étiqueté de pages web. A l'issue de ces différents traitements, nous disposons pour un chercheur, une équipe ou un laboratoire un ensemble de mots clés caractérisant les différentes compétences.

3 Quid des mesures

L'étape précédente permet d'obtenir de nombreux mots clés caractérisant les activités d'un chercheur ou d'un laboratoire. Cependant ces mots peuvent s'avérer trop généralistes et il est indispensable d'affiner cette caractérisation. Pour reprendre l'exemple de l'introduction, dire qu'une personne dans un laboratoire de fouille de données effectue ses recherches dans ce domaine n'est pas utile. Il est préférable de le caractériser au regard des techniques spécifiques qu'il met en œuvre dans ses travaux. Différentes mesures comme le TF-IDF et OKAPI ([1]) ont été adaptées à notre problématique afin de sélectionner les mots clés les plus discriminants à partir d'un ensemble de documents et selon le niveau de recherche ([2]) .

Des expérimentations menées à partir de 960 pages Web représentant 11 thématiques (par exemple, *chercheurs*, *collaborations*, *prix*, etc.) ont permis de classifier les pages en utilisant des algorithmes de classification supervisée (Naive Bayes, SVM, etc). Nos expérimentations ont montré que ces algorithmes fournissent les meilleurs résultats (F-mesure de l'ordre de 0.60%) en sélectionnant un nombre de mots clés assez réduit (quelques centaines). Ceci confirme donc la qualité des mesures de sélection des mots-clés qui seront utilisées par notre prototype décrit dans la section suivante.

4 Le prototype

Dans le cadre de cette démonstration, nous montrons de quelle manière les différentes informations sont extraites pour caractériser les informations liées aux experts. Le système proposé montre comment sélectionner un expert en fonction de ses compétences. Si celle-ci sont assez floues pour l'utilisateur, les mécanismes mis en place lui permettent d'affiner sa requête par rapport aux informations extraites des sources. La figure 1 illustre de quelle manière l'outil développé peut être utilisé pour mettre en exergue les mots clés caractérisant un laboratoire. Il est alors possible, via le système développé, d'offrir différentes granularités d'analyse (laboratoire, chercheur). Cette flexibilité offre un atout indéniable pour le décideur qui peut, outre rechercher un expert, obtenir une véritable cartographie des spécialités développées.

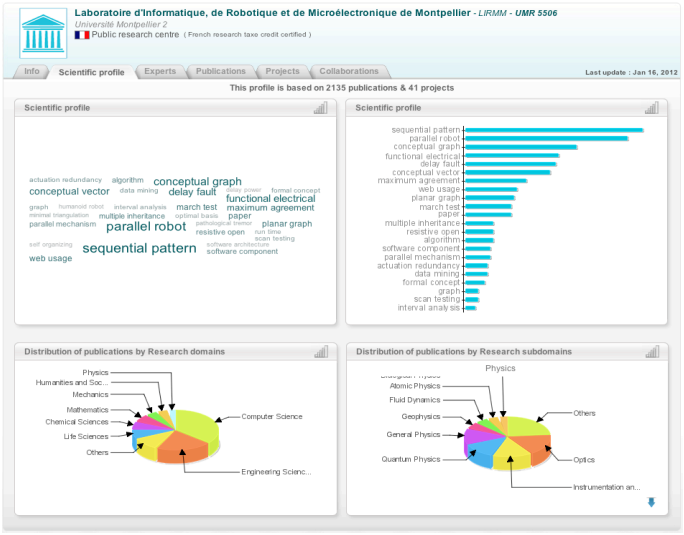

Figure 1. Un exemple d'informations extraites à partir d'un laboratoire de recherche

5 Conclusions et perspectives

Dans cet article, nous avons présenté une approche permettant de caractériser les différentes connaissances ou compétences d'un expert. Actuellement le prototype développé a été utilisé dans différents domaines scientifiques (e.g. biologie, mécanique, physique, informatique, pour différentes régions (e.g. Europe, Asie, Etats Unis, etc.) et les analyses réalisées par des experts du domaine ont montré la pertinence de la proposition. Nos travaux actuels s'intéressent à une meilleure définition des compétences en extrayant des informations plus fines associées à l'expert. Cette phase nécessite d'intégrer de nombreuses sources de données (e.g. brevet, projets, diplômes, ..) de manière à définir un véritable curriculum vitæ des différents chercheurs. En outre, de manière à avoir l'information la plus à jour possible, nous étudions comment prendre en compte l'historique, la chronologie et les différentes informations temporelles qui peuvent être extraites des sources pour compléter ou mettre à jour les informations.

Références

1. V. Claveau. Vectorisation, okapi et calcul de similarité pour le tal : pour oublier enfin le tf-idf. In *Proceedings of the Joint Conference JEP-TALN-RECITAL*, pages 85–98, 2012.
2. S. Bringay, A. Laurent, P. Poncelet, M. Roche, and M. Teisseire. Bien cube, les données textuelles peuvent s'agréger ! In *Actes de EGC (Extraction et gestion des connaissances)*, pages 585–596, 2010.