Webpage Information Extraction

December 2019

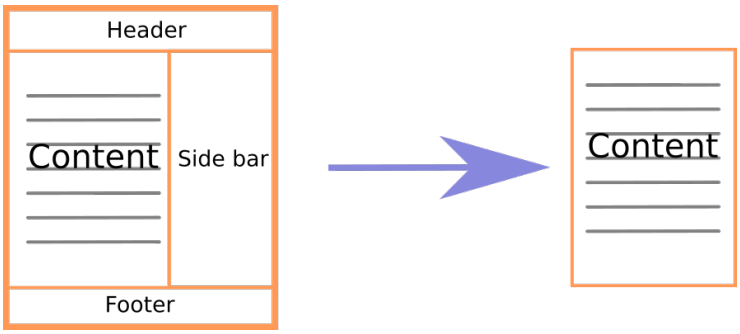

FIGURE 1 – Illustration de la tâche d'extraction d'information d'un document web

1 Contexte

Né d'un audacieux challenge, Qwant est le premier moteur de recherche à développer une véritable alternative sur le marché européen de l'Internet. Ce qui fait notre particularité est de protéger à tout prix la vie privée de nos utilisateurs. De fait, nous ignorons tout de ces derniers et cependant, notre objectif est de les aider à trouver ce qu'ils cherchent sur Internet.

Si vous souhaitez travailler dans un environnent intellectuellement stimulant, si vous pensez que la vie privée compte même si vous n'avez « rien à cacher », si faire avancer les choses pour un internet plus accessible , alors nous sommes sur la même longueur d'onde !

2 Sujet

Ce stage porte sur l'analyse de pages web brutes en vue de leur indexation par le moteur de recherche de Qwant (`www.qwant.com`). Lors de la phase de «crawling» du web [4], les robots d'exploration récupèrent les données brutes de pages web (HTML + CSS + JavaScript). Cependant dans ces documents

toutes les informations ne sont pas pertinentes pour être indexées [1] et peuvent être ignorées comme les formulaires de connexion, les balises de navigations …. En se concentrant sur ces éléments, il est ainsi plus facile d'isoler le cœur du document et d'en extraire sa sémantique.

De nombreuses méthodes existent pour résoudre ce problème en partie [2, 5, 3].

L'objectif ici est de travailler directement sur le contenu brut des pages web, c'est-à-dire sur le HTML et le CSS associé. Cela permet d'éviter de réaliser un rendu complet de la page avec exécution du JavaScript qui serait trop lent pour une indexation à très grande échelle.

Pour réaliser l'extraction des éléments d'intérêt, ce stage s'intéressera aux approches d'apprentissage statistique. En particulier, il s'agira de modéliser le document comme un graphe d'entités HTML altérées par les propriétés CSS correspondantes (c'est-à-dire produire un pseudo-DOM). Il sera ensuite possible d'appliquer des modèles d'apprentissage sur graphe avec ou sans supervision.

Une première tâche consistera à détecter automatiquement les éléments caractéristiques de la page (titre, paragraphes, images, …). En fonction des progrès du stage, nous pourrons par la suite aborder la mise en correspondance des images et des textes (par exemple, identifier la légende d'une image) et la catégorisation automatique des documents.

3 Profil recherché

Nous recherchons pour ce stage un ou une candidate de niveau M1/M2 ou en dernière année d'école d'ingénieur avec une formation en informatique ou en mathématiques appliquées. Le ou la candidate devra démontrer une bonne compréhension de l'écosystème web (connaissances en HTML, CSS et JavaScript), si possible avec une première expérience en développement web. En outre, de bonnes connaissances en apprentissage statistique sont indispensables ainsi qu'une maîtrise d'un langage de programmation tel que Python. Une première expérience en apprentissage automatique sur graphes ou en traitement du langage naturel est un plus.

4 Organisation

Le stage commencera au printemps 2020 pour une durée de 4 à 6 mois. Il sera supervisé par Hicham Randrianarivo (Qwant Research) en collaboration avec Raphaël Fournier-S'niehotta et Nicolas Audebert du CEDRIC (laboratoire d'informatique du CNAM). Le stage se déroulera 4 jours par semaine dans les locaux de Qwant (16^{e}arrondissement de Paris) et pour 1 jour par semaine dans les locaux du CEDRIC (3^{e}arrondissement de Paris).

Références

[1] Search engine indexing. `https://en.wikipedia.org/wiki/Search_engine_indexing`.

[2] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, page 2670–2676, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[3] Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, page 1535–1545, USA, 2011. Association for Computational Linguistics.

[4] Christopher Olston and Marc Najork. Web crawling. *Foundations and Trends® in Information Retrieval*, 4(3):175–246, 2010.

[5] Sunita Sarawagi. Information extraction. *Foundations and Trends® in Databases*, 1(3):261–377, 2008.