Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Conception d'un système de détection d'intention pour un moteur de recherche sur Internet

Estelle Maudet Christophe Servan
Qwant Research, 7 Rue spontini, 75116 Paris, France
intial.lastname@qwant.com

RÉSUMÉ

Dans les moteurs de recherche sur Internet, l'une des tâches les plus importantes vise à identifier l'intention de l'utilisateur. Cet article présente notre étude pour proposer un nouveau système de détection d'intention pour le moteur de recherche sur Internet Qwant. Des logs de clic au système de détection d'intention, l'ensemble du processus est expliqué, y compris les contraintes industrielles qui ont dû être prises en compte. Une analyse manuelle des données groupées a d'abord été appliquée sur les journaux afin de mieux comprendre les objectifs de l'utilisateur et de choisir les catégories d'intention pertinentes. Lorsque la recherche satisfait aux contraintes industrielles, il faut faire des choix architecturaux et faire des concessions. Cet article explique les contraintes et les résultats obtenus pour ce nouveau système en ligne.

ABSTRACT

Designing a User Intention Detection system for a Web Search Engine

In web search engines, one of the most important tasks aims to identify the user's intention. This paper presents our study to propose a new intention detection system for the Qwant web search engine. From the click logs to the detection server, the entire process is explained, including the industrial constraints that had to be taken into account. A manual analysis of clustered data was first applied on the logs to better understand the user's goals and choose relevant intent categories. When research meets industrial constraints, some architectural choices and concessions have to be made. This paper explains the constraints and the results obtained for this new online system.

MOTS-CLÉS : Détection d'intention, Classification, humain-dans-la-boucle, extraction d'information, Recherche industrielle.

KEYWORDS: Intention Detection, Classification Task, Human-in-the-loop Clustering, Information Retrieval, Industrial research.

1 Introduction

Une intention est un but derrière une action spécifique ou un ensemble d'actions. Lorsqu'un utilisateur fait une requête sur un moteur de recherche, il a généralement un but spécifique qui peut être identifié et classé. Dans ce contexte, la détection d'intention est une caractéristique clé d'un moteur de recherche sur Internet.

Au fur et à mesure que la recherche sur les objectifs de l'utilisateur progresse, il est nécessaire de caractériser plus précisément les requêtes afin de mieux répondre aux besoins de l'utilisateur (Baeza-Yates *et al.*, 2006). Au début, des systèmes fondés sur des règles ont été utilisés pour identifier des tentatives particulières. Des études plus récentes ont favorisé l'utilisation de solutions d'apprentissage

Volume 4 : Démonstrations et résumés d'articles internationaux, pages 50–52.
HAL : hal-02768512.

automatique en utilisant des logs de clics et des données Internet supplémentaires comme Wikipedia (Gabrilovich *et al.*, 2009; Ren *et al.*, 2014; Hashemi *et al.*, 2016). Depuis lors, la tâche de détection d'intention peut être considérée comme une tâche de classification. Plusieurs études ont été menées au cours des dernières années avec succès (Kim, 2014; Lai *et al.*, 2015; Zhang *et al.*, 2015; Conneau *et al.*, 2017). Toutefois, le principal inconvénient de ces approches est le non-respect des contraintes industrielles.

Ce travail a eu lieu dans un contexte industriel, le moteur de recherche Qwant. Elle implique des contraintes spécifiques en raison de la spécificité du moteur de recherche ainsi que des impératifs de production. Le premier défi industriel est le passage à l'échelle du système de détection. Comme les utilisateurs n'aiment pas attendre pour avoir leur réponse, la détection de l'intention doit avoir un faible impact sur la latence globale du moteurs de recherche. Le deuxième défi industriel était l'absence de métadonnées disponibles pour contextualiser la requête. En fait, Qwant est un moteur de recherche sur Internet axé sur la confidentialité, ce qui signifie qu'aucune information personnelle n'est stockée sur aucun serveur ni collectée (pas d'historique, pas d'adresse IP, pas de session). Seuls les logs de clics peuvent être utilisés dans le modèle d'intention. Ces logs se composent d'une liste de tuples avec une requête faite par un utilisateur anonyme et l'URL qui a été cliquée.

2 Collecte, regroupement et tri des données

L'analyse des regroupements (ou grappes) joue un rôle important dans le domaine de l'exploitation des données. Dans cette étude, l'algorithme des K-Moyennes a été utilisé sur les requêtes comme tâche préliminaire pour trouver de l'information sur les modèles. Nous avons obtenu des regroupements fins avec $K = 500$.

Alors que Wikipedia est écrit en langue naturelle avec peu d'erreurs d'orthographe, les requêtes des logs sont très bruitées et très courtes (la longueur moyenne d'une requête est d'environ 3,1 mots). Nous regroupons les requêtes à l'aide de représentation continues de phrases en considérant ces dernières comme un sac de mots (Joulin *et al.*, 2017). Nous avons appris des représentations continues de mots de 300 dimensions sur l'ensemble des données (le Wikipedia français et les logs de clics complets). Utilisant les résultats du regroupement automatique, chaque grappe a été traitée manuellement pour identifier les intentions pertinentes et les regrouper en utilisant les sites Internet les plus pertinents comme *graine*. Nous avons traité de la même façon pour toutes les classes et obtenu une liste de *graines* intéressantes pour chaque intention. (p.ex. : *jeuxvideo.com*, *gamekult.com* et *steampowered.com* pour la classe « Jeux Vidéos »).

Enfin, 100 millions de requêtes ont été regroupées en seize catégories pour créer le modèle de classification d'intention. 2000 requêtes additionnelles ont été annotées manuellement pour l'évaluation.

3 Performances & conclusion

Le modèle obtenu a un score de 83,60 de précision en considérant les 16 classes du corpus de test. Afin de répondre aux contraintes industrielles, le modèle a été chargé à l' aide d' un serveur REST en C++ *Pistache* [1]. The temps de réponse moyen est de 0,45 ms, ce qui représente environs 2 200 requêtes par seconde. Le code de l'API est disponible sur GitHub [2]. Enfin, le déploiement a permis une amélioration mesurée du taux de clic sur le *Shopping* de plus de 35%.

1. *Pistache* (pistache.io)
2. https://github.com/QwantResearch/text-classifier/

Références

BAEZA-YATES R., CALDERÓN-BENAVIDES L. & GONZÁLEZ-CARO C. (2006). The intention behind web queries. In *International Symposium on String Processing and Information Retrieval*, p. 98–109 : Springer.

CONNEAU A., SCHWENK H., BARRAULT L. & LECUN Y. (2017). Very Deep Convolutional Networks for Text Classification. In *Proceedings of the 15th Conference of the {E}uropean Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, p. 1107–1116, Valencia, Spain : Association for Computational Linguistics.

GABRILOVICH E., BRODER A., FONTOURA M., JOSHI A., JOSIFOVSKI V., RIEDEL L. & ZHANG T. (2009). Classifying search queries using the web as a source of knowledge. *ACM Transactions on the Web (TWEB)*, **3**, 5.

HASHEMI H. B., ASIAEE A. & KRAFT R. (2016). Query intent detection using convolutional neural networks. In *International Conference on Web Search and Data Mining, Workshop on Query Understanding*.

JOULIN A., GRAVE E., BOJANOWSKI P. & MIKOLOV T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, p. 427–431.

KIM Y. (2014). Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing ({EMNLP})*, p. 1746–1751, Doha, Qatar : Association for Computational Linguistics. DOI : 10.3115/v1/D14-1181.

LAI S., XU L., LIU K. & ZHAO J. (2015). Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

REN X., WANG Y., YU X., YAN J., CHEN Z. & HAN J. (2014). Heterogeneous graph-based intent learning with queries, web pages and wikipedia concepts. In *Proceedings of the 7th ACM international conference on Web search and data mining*, p. 23–32 : ACM.

ZHANG X., ZHAO J. & LECUN Y. (2015). Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, p. 649–657.