WIB : un navigateur intégré pour Wikipédia destiné à l'évaluation participative de modèles de pertinence

Christophe Brouard*, Jean-Pierre Chevallet**, Téo Orthlieb*, Habib Slim*

Université Grenoble Alpes, LIG UMR 5217/Equipes AMA* et MRIM**, France
Christophe.Brouard@univ-grenoble-alpes.fr
http://ama.liglab.fr/ brouard/

Résumé. Nous présentons ici l'application WIB (pour Wikipedia Integrated Browser) qui permet de naviguer dans les documents Wikipédia en même temps que dans les termes contenus dans ces documents et les catégories auxquelles ils appartiennent. Selon le(s) type(s) d'item(s) considéré(s) en entrée et en sortie (termes, documents ou catégories), la tâche résolue par l'application varie (recherche d'information, extension de requête, extraction de mots clefs, classification automatique,...) mais il s'agit toujours de sélectionner les items pertinents vis-à-vis de la requête en s'appuyant sur un modèle de pertinence. Cette application est un moyen d'expérimenter en ligne différents modèles de pertinence. Toutes les actions des utilisateurs sont enregistrées et stockées dans une base de données en vue d'analyses comparatives ultérieures. Une première version de l'application est déjà en ligne (http://echo.imag.fr/apps/echopedia/).

1 Introduction

Au-delà des traditionnels systèmes de recherche d'information qui permettent de sélectionner des documents à partir de mots clefs, on peut s'intéresser plus globalement aux modèles de pertinence (Huang et Soergel, 2013), (Brouard et Nie, 2004) permettant de mettre en relation différentes entités. La recherche de tels modèles qui suppose l'élicitation de la relation liant la requête et l'information pertinente reste un sujet de recherche actuel et la définition de méthodes d'évaluation pour ces modèles est un prérequis. L'évaluation traditionnelle repose sur le paradigme de Cranfield (Harman, 2013) dans lequel un ensemble figé de documents et de jugements de pertinence réalisés par des experts pour des requêtes données servent de base de comparaison aux réponses des systèmes. L'avantage du paradigme de Cranfield est de permettre un certain contrôle. Les experts sont censés fournir des jugements fiables et le nombre de documents étant limité, il est possible d'établir des mesures de rappel (en considérant l'ensemble des réponses pertinentes existant dans le corpus). Néanmoins ce mode d'évaluation suppose la mise à disposition de ressources humaines importantes (beaucoup de jugements de pertinence à produire) et par conséquent ne peut être mis en oeuvre que sur des corpus de taille relativement limité. Ces ensembles de documents, requêtes et jugements de pertinence associés sont par ailleurs fabriqués de façon artificielle. Considérant l'existence de banques de connaissances accessibles en ligne, la recherche de solutions s'appuyant sur une évaluation participative des systèmes de recherche d'information se développe (Kazaï, 2018). Une approche

consiste à rendre accessible un système d'interrogation et de recueillir in situ les actions des utilisateurs en vue d'analyses ultérieures. L'intérêt est notamment de pouvoir construire des corpus de taille beaucoup plus importante et directement liés à une véritable utilisation. L'application WIB se situe dans cette approche. Elle permet l'utilisation en aveugle de différents modèles de pertinence en vue de leur comparaison. Dans la suite, nous donnons quelques éléments d'ingénierie sur la prise en main des bases d'articles Wikipédia. Nous continuons en présentant l'application et les différentes tâches qu'elle permet de résoudre. Nous donnons dans la dernière partie quelques éléments architecturaux de l'application qui garantissent une intégration aisée de nouveaux modèles de pertinence. Nous concluons sur le potentiel partage de l'application avec la communauté scientifique.

2 Ingénierie autour de Wikipédia

L'encyclopédie collaborative Wikipédia créée en 2001 a permis le recueil de connaissances sous forme textuelle dans plusieurs centaines de langues sur la base de la participation des internautes. Il s'agit de la plus grande encyclopédie du monde et les sites de cette encyclopédie en ligne sont très visités. Elle contient par exemple actuellement plus de 2 millions d'articles en français. L'intégralité des contenus peut être récupérée selon différents formats à partir de la page https ://dumps.wikimedia.org/. Les articles Wikipédia respectent une structure mélangeant du XML et le format dédié WikiText. Nous avons développé un parseur paramétrable et utilisable via une interface graphique (figure 1) pour nous permettre d'extraire facilement de ce fichier les informations utiles à nos expérimentations. Afin de réaliser de premiers tests, nous avons extrait pour chaque article, son titre, les catégories auxquelles il appartient et les 100 premiers termes de son introduction.

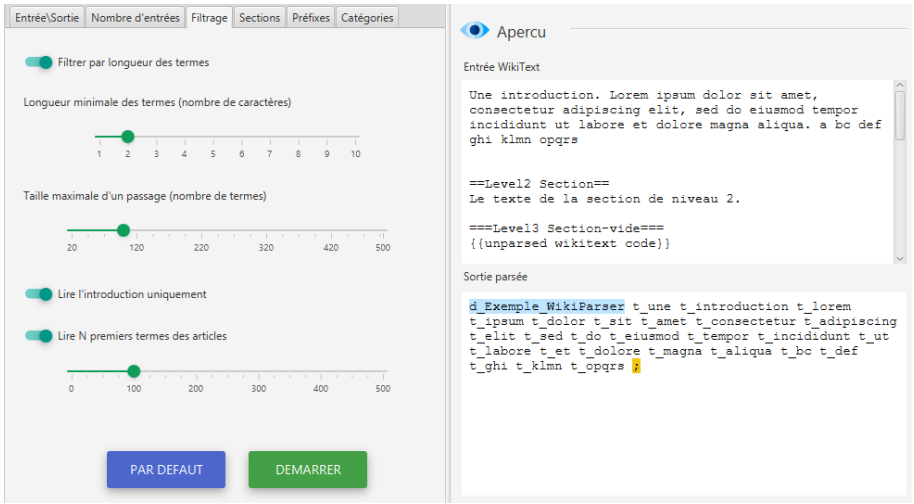

FIG. 1 – *Interface du parseur . Il est possible d'indiquer qu'on ne souhaite garder que l'introduction de chaque article et les 100 premiers termes.*

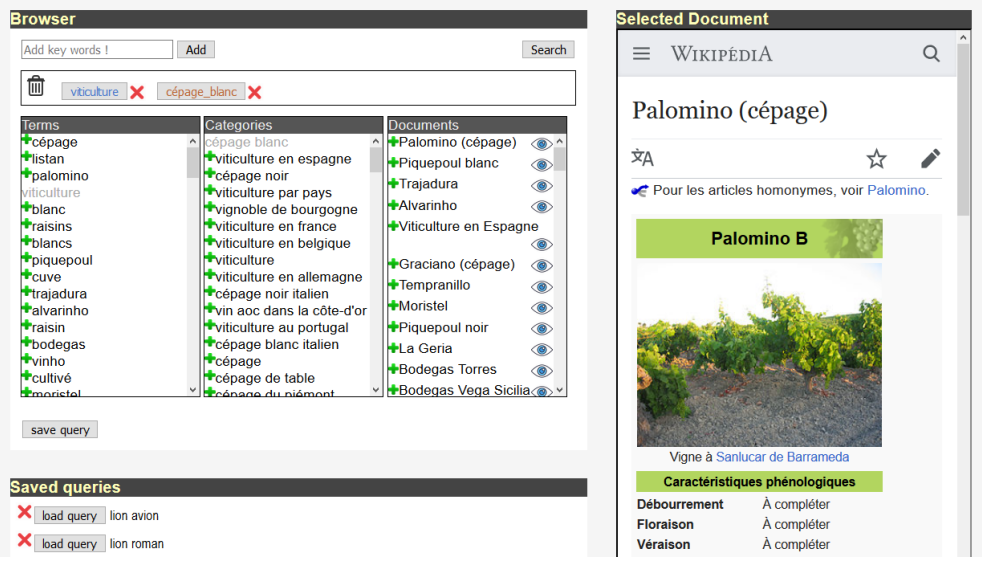

FIG. 2 – *Interface du navigateur WIB. Ici la requête est composée du terme "viticulture" et de la catégorie "cépage blanc". Des termes, catégories et documents sont proposés en réponse.*

3 Un navigateur intégré

Le navigateur WIB (figure 2) permet de définir sa requête avec des termes et/ou des documents et/ou des catégories et présente en retour des termes, des documents et des catégories qu'il est possible d'ajouter ensuite à sa requête pour réitérer le processus. Selon le type des entrées et des sorties observées, on peut déjà identifier neufs types de tâches différentes si on considère des requêtes composées du même type d'item. Par exemple, lorsque la requête ne contient que des termes et qu'on ne considère que les documents proposés en résultat, il s'agit d'une simple tâche de recherche d'information. Toujours avec des termes dans la requête mais dans le cas où le résultat considéré est la liste des catégories cela s'apparente à une tâche de classification automatique et si on considère les termes proposés en résultat cela peut être vu comme une extension de requête. On peut aussi utiliser l'application pour extraire les mots clefs d'un document ou d'une catégorie ou chercher les documents ou les termes correspondant à une ou plusieurs catégories.

4 Un outil de comparaison

En l'état actuel, l'application s'appuie sur le modèle de pertinence ECHO (Brouard, 2012) mais l'objectif est d'intégrer d'autres modèles pour les comparer entre eux. Toutes les actions utilisateurs regroupées par session d'utilisation sont stockées dans une base de données. Ces actions et en particulier les choix d'ajout à la requête peuvent être interprétés comme des retours indiquant une satisfaction de l'utilisateur et pourront être exploitées par la suite pour

comparer les différents modèles. L'application a été conçue avec le souci de permettre une intégration facile d'autres modèles de pertinence. Toutes les interactions de l'application WIB avec le modèle de pertinence sous-jacent se matérialise par un appel AJAX. Pour pouvoir être intégré à l'application un nouveau modèle de pertinence doit se conformer à une API bien définie et être mentionnée dans le fichier de configuration de l'application. Cette API consiste en un simple script php recevant les requêtes composites et retournant les 100 premiers résultats dans des formats JSON.

5 Conclusion et perspectives

Le calcul de certains indices de performances vont être ajoutés à l'application pour pouvoir comparer les différentes instances entre elles. Le code de cette application pourra à terme être mis à la disposition de tous permettant à chacun de créer sa propre instance avec son modèle de pertinence ou ses systèmes dédiés mais quelques efforts sont encore nécessaires pour faciliter le déploiement de nouvelles instances de l'application. L'instance de l'application (s'appuyant sur le modèle ECHO) qui est déjà en ligne : http ://echo.imag.fr/apps/echopedia/ nous permettra d'obtenir un premier retour d'utilisation.

Références

Brouard, C. (2012). Document classification by computing an echo in a very simple neural network. In *IEEE International Conference on Tools with Artificial Intelligence*, pp. 735–741.

Brouard, C. et J.-Y. Nie (2004). Relevance as resonance: a new theoretical perspective and a practical utilization in information filtering. *Information Processing and Management 40*, 1–19.

Harman, D. (2013). Trec-style evaluations. In *Proceedings of the 2012 International Conference on Information Retrieval Meets Information Visualization*, PROMISE'12, pp. 97–115. Springer-Verlag.

Huang, X. et D. Soergel (2013). Relevance: An improved framework for explicating the notion. *JASIST 64(1)*, 18–35.

Kazaï, G. (2018). Challenges in building ir evaluation pipelines. In *Keynote of 40th European Conference on Information Retrieval*.

Summary

Here we present the WIB application (for Wikipedia Integrated Browser) that allows browsing Wikipedia documents. Depending on the type (s) of item (s) considered as input or output (terms, documents or categories), the task resolved by the application varies (information retrieval, query extension, keywords extraction, automatic classification, ...) but the question is always to cope with the selection of the relevant items. This application is a way to experiment online with different models of relevance. This application is currently available on the web at the following url : http://echo.imag.fr/apps/echopedia/.