Un moteur de recherche d'événements pour explorer la presse numérique ou historique

Guillaume Bernard, Thomas Blot

Un moteur de recherche d'événements pour explorer la presse numérique ou historique

Guillaume Bernard [1], **Thomas Blot** [2]

1. *Laboratoire L3i*
   Université de La Rochelle, France
   guillaume.bernard@univ-lr.fr
2. *Université de Bordeaux, France*
   thomas.blot@u-bordeaux.fr

RÉSUMÉ. *La presse écrite rapporte tous les jours des événements. Faciliter l'exploration de cette presse est particulièrement pertinent pour les scientifiques en humanités numériques qui en font leur matériau principal. Alors que les approches par clustering sont le sujet majeur de la littérature scientifique, nous présenterons à la conférence un moteur de recherche d'événements contraint par des critères de sobriété énergétique et à la latence faible.*

ABSTRACT. *The written press reports events on a daily basis. Facilitating the exploration of this press is particularly relevant for digital humanities scientists who use it as their primary material. While text-based approaches are the main focus of the scientific literature, we will present at the conference a low-latency, energy-efficient event search engine.*

MOTS-CLÉS : *Ingénierie des données liées, Événements, Fouille dans les données massives*

KEYWORDS: *Linked Data, Events, Data mining*

1. Introduction

La presse rapporte au quotidien les événements qui surviennent dans le monde. Les articles, une fois publiés, ancrent les événements dans le temps et l'espace pour former une chronologie qui s'achève à la disparition médiatique de l'événement. Établir cette chronologie *a posteriori*, même en se focalisant uniquement sur la presse écrite, est un problème ouvert. Nous ajoutons la contrainte d'exiger une réponse rapide de quelques secondes, qui ne nécessite pas d'apprentissage. Ces contraintes s'inscrivent dans le cadre de la sobriété numérique qui est devenue un enjeu sociétal majeur. Nous émettons l'hypothèse que se passer d'apprentissage réduit les coûts de calcul et les temps de traitement. Pourtant, répondre à la question *comment regrouper toutes les connaissances, articles et documents relatifs à un événement passé ?* est d'un grand intérêt pour la recherche en Sciences Humaines et Sociales ou le journalisme. Rechercher les

articles liés à un événement *A* pour établir une chronologie des éléments rapportés par les médias à ce sujet est en effet une tâche cruciale pour ces disciplines.

Afin de répondre à cette problématique et permettre la création de corpus fiables pour ces publics variés, nous modélisons des *histoires* sur les événements passés. En section 2, nous introduirons la modélisation des événements puis, en section 3, nous présenterons notre moteur de recherche d'événements.

2. Modéliser les chronologies d'événements

Le suivi d'événements est exploré par divers projets depuis les années 2000. Ils utilisent des algorithmes de *machine learning* qui reposent sur un pré-traitement du texte, généralement un encodage *TF-IDF* (Miranda *et al*., 2018) ou dense (Linger, Hajaiej, 2020). La date et le texte des articles sont pris en compte pour créer des *clusters* d'articles qui décrivent le même événement.

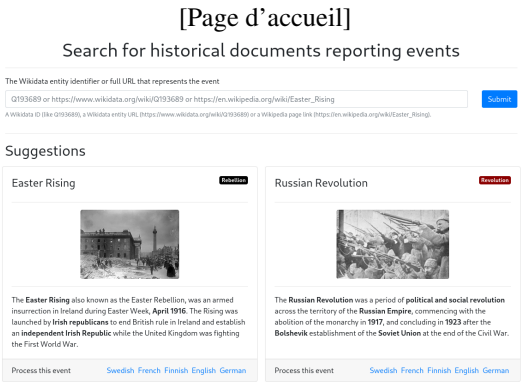
Dans l'analyse d'événements passés, certaines propriétés sont connues d'avance : date, lieu de l'événement ainsi que les entités impliquées. Ces connaissances *a priori* existent dans des bases ouvertes comme Wikidata ou Wikipédia. (Bernard *et al*., 2021) ont développé un outil, `wikivents`, qui analyse ces bases pour extraire ces informations : date, lieu et entités nommées concernées (personnes, lieux, organismes). (Wang *et al*., 2022) ont proposé une architecture d'indexation pour requêter des articles de presse et extraire des informations pour une tâche de *question answering*. Nous suggérons d'associer ces deux stratégies pour : 1. extraire les connaissances sur les événements ; 2. utiliser ces connaissances pour créer des requêtes.

3. Architecture du système

Notre architecture se base sur trois éléments : les données de presse à explorer, un outil de description d'événements, ici `wikivents` et une infrastructure pour indexer les documents. Pour l'indexation, nous déployons *ElasticSearch*, comme (Wang *et al*., 2022). Nous extrayons, pour chaque document, son titre et son texte ainsi que les lemmes, les termes et les entités nommées dudit texte. Ces éléments sont utilisés en complément du texte brut pour effectuer les recherches. En entrée du système, un identifiant d'événement sur Wikidata est fourni par l'utilisateur, comme montré en figure 1. Il ou elle peut choisir un événement prédéterminé ou indiquer l'identifiant d'un événement qu'il a repéré sur Wikidata. Les articles étant rédigés dans une langue unique, il convient de sélectionner également une langue cible. `wikivents` extrait les entités, dates ou noms de lieux relatifs à cet événement en explorant à la fois Wikidata et les résumés des pages Wikipédia associées aux événements. La représentation ainsi obtenue est considérée comme agnostique à la langue, car elle ne se base que sur des identifiants de concepts dans la base Wikidata. Par exemple, le lieu « Saint Petersbourg » est identifié par *Q656*. Une seconde étape permet de convertir cette représentation dans une langue cible. Nous utilisons l'ensemble des noms dans une langue donnée utilisés pour décrire les entités sur Wikidata. Une fois la description

des événements ainsi obtenue, elle est transformée par un second outil qui forge une requête pour *ElasticSearch*. La requête est adaptée à la recherche de documents dans une unique langue. Le travail soumis dans cette démonstration se positionne donc à l'interface entre `wikivents` et *ElasticSearch*, tel que présenté en figure 1.

[Page d'accueil]



[Processus général]

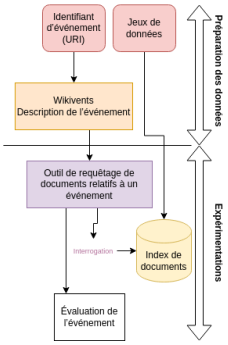


FIGURE 1. *Page d'accueil et schéma décrivant le fonctionnement du système.*

Une fois l'événement analysé et la requête construite, celle-ci est envoyée à *ElasticSearch* qui à son tour recherche les documents et les affiche. Ce sont les articles de presse qui font mention de l'événement. Une fois ces résultats obtenus, l'utilisateur dispose des articles correspondants (figure 2). Pour chaque événement, un court résumé est affiché et l'ensemble des entités détectées - qui ont contribué à la recherche - sont listées. Elles sont dites participantes : elles décrivent l'événement. Les documents connus et rapportant l'événement sont listés et l'image source, si disponible, est affichée. Ici, le texte est extrait d'images par reconnaissance optique de caractères. Les données présentées sont des articles de presse en français publiés par la Bibliothèque Nationale de France et distribués *via* la plate-forme du projet NewsEye [1].

4. Résultats et conclusion

Nous avons comparé notre moteur de recherche à des techniques de *clustering* qui remplissent les mêmes objectifs (Miranda *et al.*, 2018). Nous avons utilisé le jeu de données de ces derniers, composé d'articles en anglais, allemand et espagnol (). Chaque article est associé à un identifiant d'événement qui nous permet d'identifier tous les articles de presse qui évoquent un événement commun. Nous avons comparé la *F1* du *clustering* par *machine learning* avec la *F1* de notre moteur de recherche. Les

1. https://www.newseye.eu

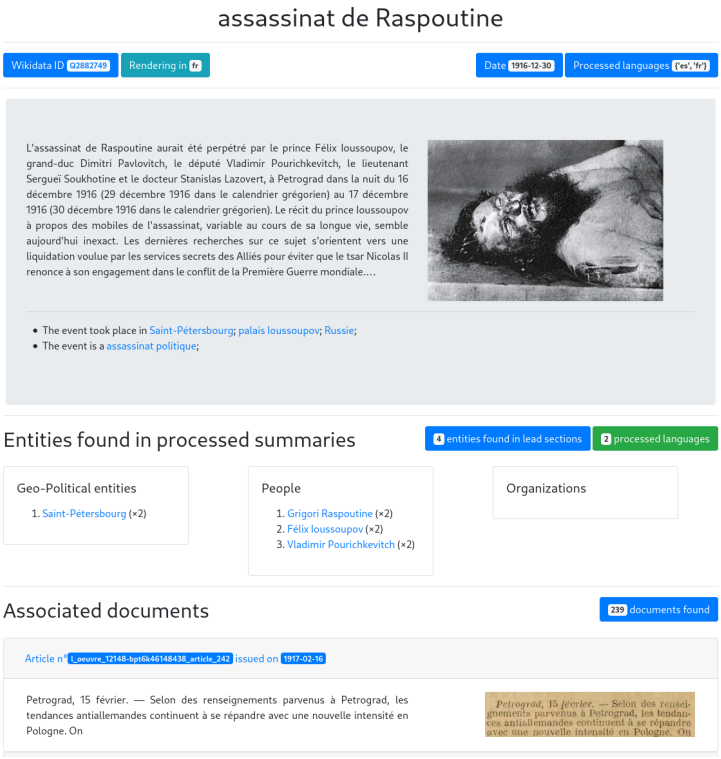


FIGURE 2. *Recherche des documents évoquant l'événement. Un résumé de l'événement est présenté et les articles trouvés sont listés plus bas.*

données et métriques utilisées ont été décrites par (Bernard, 2022) et les références vers les jeux de données y sont partagées.

Nous avons analysé le jeu de données Miranda *et al.* à la recherche des événements qui sont décrits par un nombre élevé d'articles. Nous avons retenu 81 événements, soit plus de 7 000 documents sur les plus de trente mille qui composent le corpus. Les événements sont de différents types : judiciaires, conflits, sportifs ou politiques, pour ne citer que les plus importants. Pour chacun de ces 81 événements, nous avons recherché leur identifiant Wikidata manuellement en parcourant à la fois les articles et l'encyclopédie. Cette association faite, il nous est possible d'évaluer le processus inverse et de vérifier que le moteur de recherche associe adéquatement des identifiants d'événements aux documents qui les décrivent.

Nous évaluons la capacité des algorithmes de *clustering* ou le moteur de recherche à grouper correctement les documents selon qu'ils rapportent les mêmes événements. Nous obtenons des métriques de *clustering* de précision, de rappel et de F-mesure, et

dans chaque langue, puisque le moteur de recherche fonctionne sur un principe monolingue. Puisque la valeur de précision et de rappel dépend du nombre de documents renvoyés par le moteur de recherche (plus le nombre de documents récupéré augmente, plus la précision décroît), nous avons, par analogie, considéré que la F-mesure maximale représente l'expertise du scientifique en humanités numériques. Par exemple, il ou elle sait déterminer grâce à ses connaissances à partir de quel moment les documents renvoyés ne sont plus pertinents au regard de la requête initiale. C'est cette valeur de F-Mesure maximale que nous utilisons pour évaluer un événement unique. La F-Mesure pour une requête (évaluée donc sur 81 événements) est la médiane de cet ensemble de F-Mesures par événement.

À contexte comparable (mêmes données, même langue, mêmes événements), nous obtenons pour le moteur de recherche une F-Mesure de 0,67 en anglais. Elle est de 0,91 pour les méthodes basées sur du *machine learning*. La différence de 0,24 point peut s'expliquer par l'absence de données suffisamment descriptives sur Wikidata et Wikipédia. Cela entraîne une mauvaise représentation des événements (dates, lieux ou participants manquants). Une autre cause possible est la redondance de certains événements ayant lieu annuellement (ex: Roland Garros) et présents dans le jeu de données testé pour lesquels les données Wikidata sont peu fiables ou incomplètes, du fait de cette redondance. Compte tenu de la sobriété de la méthode, les résultats sont cependant encourageants et peuvent être obtenus rapidement, sans investissement important. À ce titre, cette méthode et ces outils peuvent être appliqués à de nombreux types de documents et de nombreuses situations de recherche impliquant la création de corpus.

Remerciements

Les auteurs souhaitent remercier particulièrement Cyrille Suire pour sa relecture attentive ainsi que Cyril Faucher et Antoine Doucet pour les conseils scientifiques qu'ils ont apporté durant l'élaboration de ce projet.

Bibliographie

Bernard G. (2022). *Détection et suivi d'événements dans des documents de presse historique*. Thèse de doctorat non publiée, Université de La Rochelle (ULR), La Rochelle.

Bernard G., Suire C., Faucher C. *et al*. (2021). A Comprehensive Extraction of Relevant Real-World-Event Qualifiers for Semantic Search Engines. In *Proceedings of the 25th International Conference on Theory and Practice of Digital Libraries*, p. 153–164.

Linger M., Hajaiej M. (2020). Batch Clustering for Multilingual News Streaming. In *Proceedings of Text2Story - ECIR.*

Miranda S. a., Znotiņš A., Cohen S. B. *et al*. (2018). Multilingual Clustering of Streaming News. In *2018 Conference on EMNLP*, p. 4535–4544.

Wang J., Jatowt A., Yoshikawa M. (2022). *ArchivalQA: A Large-scale Benchmark Dataset for Open Domain Question Answering over Historical News Collections.*