Analyse en dependance et classification de requetes en langue naturelle, application a la recommandation de livres

Anaïs Ollagnier, Sébastien Fournier, Patrice Bellot

Analyse en dépendance et classification de requêtes en langue naturelle, application à la recommandation de livres

Anaïs Ollagnier[*,**] **— Sébastien Fournier**[*] **— Patrice Bellot**[*,**]

[*] *Aix-Marseille Université, CNRS, LSIS UMR 7296, 13397 Marseille, France*
[**] *Aix-Marseille Université, CNRS, CLEO OpenEdition UMS 3287, 13451 Marseille, France*
anais.ollagnier, sebastien.fournier, patrice.bellot@lsis.org

RÉSUMÉ. *Cet article propose une approche de recommandation fondée sur l'analyse de requêtes. En particulier, nous nous focalisons sur des requêtes dans lesquelles l'utilisateur cherche des similitudes entre des livres, des auteurs ou encore des collections. Notre approche consiste, dans un premier temps, à repérer ces requêtes grâce à un procédé de classification automatique supervisée. Dans un deuxième temps, un analyseur en dépendance est employé afin de dégager les liens syntaxico-sémantiques présents au sein de ce type de requêtes. Une fois ces dépendances extraites, plusieurs stratégies d'expansion de ces requêtes sont mises en place afin de les exploiter comme entrées d'un index puis,* via *l'utilisation d'un modèle DFR : InL2. Nos expérimentations montrent que ces pistes sont un pas supplémentaire vers la compréhension des besoins utilisateurs exprimés au sein des requêtes longues et détaillées.*

ABSTRACT. *This article proposes an approach to recommendation oriented analysis queries. In particular, we focus on queries where the user is looking for similarities between books, authors or collections. Our approach is, firstly, to identify these requests through an automatic supervised classification method. Secondly, a dependency analyzer is used to identify the syntactic and semantic links present in this type of requests. Once these dependencies extracted several expansion strategies of these requests are put in place to exploit them as inputs an index and then, by a model DFR: InL2. Our experiments show that these achievements are a further step towards the understanding of user requirements expressed in the lengthy and detailed queries.*

MOTS-CLÉS : *système de recommandation de lectures, classification supervisée de requêtes, analyse en dépendance, expansion de requêtes.*

KEYWORDS: *reading recommendation system, supervised classification queries, dependency analysis, query expansions.*

1. Introduction

Cet article propose une approche de recommandation fondée sur l'analyse de requêtes. En particulier, nous nous focalisons sur des requêtes dans lesquelles l'utilisateur cherche des similitudes entre des livres, des auteurs ou encore un ensemble d'ouvrages présentant une unité que l'on retrouve sous la forme de collection. Par convenance, nous nommerons ce type de requêtes des requêtes *analogues* tout au long de cet article. Nous travaillons sur des requêtes de recherche de livres longues et détaillées exprimant le besoin informationnel de l'utilisateur. Ces travaux sont affiliés aux études menées sur l'analyse de requêtes verbeuses que l'on retrouve dans des domaines tels que les systèmes de questions-réponses ou encore les moteurs de recherche axés sur des entités. L'intérêt grandissant pour la recherche d'information (RI) au sein des requêtes verbeuses en langage naturel s'explique par l'objet de nombreuses nouvelles applications qui est passé de requêtes constituées de mots-clés à des requêtes longues et détaillées (Manish et Bendersky, 2015). Au cours de la dernière décennie, cette problématique a été largement explorée *via* la mise en place de procédés tels que la réduction, la pondération ou encore l'expansion afin d'obtenir une représentation plus efficace de ce type de requêtes. Cependant, la majorité de ces travaux mettent de côté la sémantique de la phrase en exploitant uniquement les mots. Nous souhaitons conserver, au sein de notre approche, la structure de la phrase *via* l'utilisation d'un procédé issu du traitement automatique des langues (TAL) : un analyseur en dépendance. Notre approche consiste, tout d'abord, à repérer les requêtes analogues grâce à un processus de classification automatique supervisée suivant une taxonomie propre aux besoins exprimés dans les requêtes de recherche de livres. Ensuite, une fois ces requêtes analogues repérées, nous utilisons un analyseur en dépendance en vue de dégager les liens syntaxico-sémantiques entre les mots. Dans un troisième temps, une fois ces dépendances extraites, plusieurs stratégies d'expansion de ces requêtes sont mises en place afin de les exploiter comme entrées d'un index puis, *via* l'utilisation d'un modèle de « déviation par rapport à l'aléatoire »[1] (DFR) nommé InL2. Notre contribution porte sur une meilleure préservation de la sémantique de la phrase *via* l'utilisation d'un analyseur en dépendance sur des requêtes analogues. À des fins expérimentales, nous nous focalisons sur ces requêtes mais il est évident que cette approche est généralisable. Nous croyons que des traitements spécifiques doivent être et peuvent être mis en place selon le type de la requête et ainsi dépasser les limites des approches « sacs de mots », afin d'aboutir à une amélioration de la recommandation.

Cet article est structuré comme suit : dans la section 2 nous effectuons un état de l'art. La section 3 présente notre cadre applicatif. La section 4 décrit l'architecture du modèle de recherche de livres. La section 5 détaille l'approche fondée vers la génération d'un modèle de recherche de livres s'appuyant sur la compréhension des besoins utilisateurs formulés au sein de requêtes analogues. Dans la section 6 nous exposons les résultats obtenus, à la fois, sur la détection des requêtes analogues par une approche de classification automatique supervisée et sur les différentes stratégies

1. *Divergence From Randomness* : déviation par rapport à l'aléatoire.

d'expansion de ces requêtes. Enfin, les conclusions sont présentées dans la section 7.

2. État de l'art sur la représentation des requêtes longues et détaillées et sur les modèles de livres

2.1. *Représentation des requêtes longues et détaillées*

La représentation « sacs de mots » est largement utilisée en recherche d'information. Avec cette représentation, la requête initiale est transformée en un ensemble de mots pondérés. Certaines extensions à cette représentation introduisent des syntagmes (Metzler et Croft, 2005), des mots latents (Metzler et Croft, 2007) ou des concepts (Albitar *et al.*, 2014). Un problème majeur persiste : la non-correspondance entre les termes[2]. Ce problème est induit par des phénomènes tels que la synonymie ou la polysémie (Furnas *et al.*, 1987). Diverses approches ont été proposées pour améliorer la représentation de la requête par reformulation dont l'expansion de requête. Elle consiste à étendre la requête d'origine avec d'autres mots qui représentent mieux l'intention réelle de l'utilisateur. L'expansion de requête peut être manuelle, automatique ou interactive. Pour chacune de ces expansions, plusieurs sources sont nécessaires. Il en existe deux types : les résultats de recherche et les bases de connaissances lexicales. Dans le cas de bases de connaissances lexicales dépendantes des données, nous avons à faire à des algorithmes de modification des mots (suppression de suffixes, recherche de similarités entre les mots, de regroupements, etc.). Dans le cas de base de connaissances lexicales indépendantes, se sont des thésaurus, dictionnaires ou lexiques spécifiques au domaine qui sont employés. Parmi les techniques d'expansion observées nous pouvons citer des requêtes fondées sur des logs (Cui *et al.*, 2002), des modèles de marche aléatoire[3] (Collins-Thompson et Callan, 2005), des données de structures auxiliaires (Billerbeck et Zobel, 2006), utilisant des thèmes et des locations (Huang *et al.*, 2007) ou encore des analyses lexicales et sémantiques (Wasilewski, 2011).

2.2. *Modèle de recherche de livres*

Concernant les modèles de recherche de livres, des campagnes d'évaluation telles que *INEX Social Book Search Track*[4] (INEX SBS) (Hall *et al.*, 2014) permettent de s'orienter plus particulièrement sur des modèles de recherche de livres. INEX SBS a été introduit en 2010.
INEX SBS se découpe en deux sous-tâches : la première dédiée à la suggestion et la seconde relative à l'interactivité. Les travaux présentés lors de ces campagnes exploitent trois types d'informations : les informations sociales, les informations descriptives et

2. *Mismatch term* : non-correspondance entre les termes.
3. *Random walk models* : modèles de marche aléatoire.
4. `http://social-book-search.humanities.uva.nl/#/overview`

les informations sur le contenu textuel. Les informations sociales englobent tous les avis émis par les utilisateurs à propos d'un livre (commentaires, évaluations, etc.). Les informations descriptives renvoient aux métadonnées contenant des informations sur un livre telles que l'auteur, le prix, le nombre de pages, etc. Et les informations sur le contenu textuel correspondent au contenu même du livre dans son intégralité ou non. Toutes ces caractéristiques donnent lieu à de nombreuses combinaisons propres aux systèmes de RI dédiés à la recherche de livres.

La majorité des pistes abordées se fonde sur le principe de correspondance de mots en faisant varier le nombre de champs à considérer tout en jouant sur les poids des paramètres des mots ou groupes de mots. Des techniques de reclassement, qui consistent à classer de nouveau les résultats initiaux fournis par les modèles de recherche, ainsi que des techniques d'enrichissement des résultats fournis par le modèle de recherche sont également utilisées.

La grande majorité des techniques de reclassement présentées consiste à établir un score social (fondé sur les métadonnées générées par les utilisateurs) qui est ensuite combiné linéairement au modèle de recherche. Parmi les scores présentés, Bonnefoy *et al.* (2012) émettent l'hypothèse que plus un livre a de commentaires, et si ses notes sont généralement bonnes, plus il doit être un très bon livre. Benkoussas et Bellot (2013) effectuent une pondération entre l'évaluation de chaque commentaire et les votes attribués par les utilisateurs sur l'utilité de ce même commentaire (*helpful vote*). D'autres travaux tentent également d'effectuer un reclassement mais cette fois *via* des méthodes d'apprentissage de rang[5], ce sont par ailleurs ces travaux qui ont obtenu les meilleures performances en 2014 (Zhang *et al.*, 2014). En 2015, c'est également une technique de reclassement fondée sur des méthodes d'apprentissage de rang qui a obtenu les meilleurs résultats (Gäde *et al.*, 2015). L'originalité de cette approche est de tenir compte des balises de livres qui sont peu utilisées comme le prix et la longueur du livre. Concernant l'enrichissement des résultats, Benkoussas *et al.* (2015) présentent un modèle de RI traditionnel dont les résultats sont enrichis par une analyse effectuée par des graphes.

Certains chercheurs se sont intéressés plus spécifiquement aux besoins exprimés par les utilisateurs en implémentant une base de connaissances. Wu *et al.* (2014) utilisent un jeu d'amorces extraites de requêtes dans lesquelles l'utilisateur cherche des livres similaires à ceux énoncés dans sa requête. Un filtre a été mis en place afin de ne pas laisser passer les livres non nécessaires lors de la recommandation de ce type de requêtes.

3. Cadre applicatif

Afin d'évaluer les modèles de recherche réalisés, nous utilisons les données fournies dans le cadre des campagnes d'évaluation INEX SBS menées par CLEF Initia-

5. *Learning to rank* : apprentissage de rang. L'apprentissage de rang est l'application de techniques d'apprentissage dans la construction de modèles de classement pour les systèmes de recherche d'information (Li, 2011).

tive[6]. Pour notre travail, nous nous orientons vers la tâche de suggestions, qui consiste à établir une liste des livres les plus pertinents en fonction d'une requête émise par un utilisateur. Les données utilisées au cours de nos expérimentations sont celles fournies en 2014. Elles se décomposent en un jeu de 680 requêtes longues et détaillées[7] exprimées en langue naturelle posées par les utilisateurs de LibraryThing[8] (LT), dont un exemple est présenté par la figure 1. Toutes les requêtes fournies sont en anglais. Pour chaque requête nous disposons de cinq champs : <title>, <query>, <narrative>, <group>, <catalog>. Les deux derniers champs cités correspondent respectivement à la communauté dans laquelle la requête est adressée et au catalogue personnel de l'utilisateur qui a écrit la requête. Il est important de préciser que les requêtes formulées sont destinées à d'autres êtres humains et non à un moteur de recherche, ce qui engendre des requêtes longues et détaillées aux structures complexes.

```
<topic id="1584">
  <title>Great alternative history books?</title>
  <mediated_query>alternate history and alternative histories</mediated_query>
  <group>Time Travel, Alternate Histories and Parallel Worlds</group>
  <narrative>    I love alternative histories - two great ones I've enjoyed are
Robert Harris's  Fatherland  and Kim Stanley Robinson's  Years of Rice and Salt .
Any other recommendations?  John     </narrative>
```

Figure 1. *Exemple de requête d'INEX SBS 2014*

La collection de livres est constituée de 2,8 millions de descriptions de livres extraites d'Amazon[9], elle est composée de 64 champs XML (un exemple est présenté figure 2). Parmi ces champs, nous distinguons :

– les métadonnées : <book>, <isbn>, <title>, <authorid>, etc.

– les informations sociales : <reviews>, <summary>, <tags>, <rating>, etc.

Un ensemble de 93 976 profils anonymes d'utilisateurs est également fourni *via* LibraryThing. Ces profils contiennent chacun le catalogue personnel de l'utilisateur. Ce catalogue se compose de l'ensemble des commentaires effectués sur un livre avec son évaluation et éventuellement un jeu de balises relatif à ses thématiques.
L'étude des travaux effectués au cours des campagne INEX SBS nous permet de constater que très peu de travaux ne se penchent sur une interprétation sémantique du contenu de la requête. Bien que Wu *et al.* (2014), se soient intéressés à la compréhension des besoins des utilisateurs, leurs travaux ne s'orientent toutefois pas vers une interprétation sémantique de la requête avec une réelle interprétation des besoins exprimés par les utilisateurs.
Ces dernières années, les travaux vont plus vers une compréhension des profils utilisateurs, ainsi que des métadonnées générées par ces derniers, sans cependant tenir

6. http://www.clef-initiative.eu/ (*Conference and Labs of the Evaluation Forum*)
7. http://inex.mmci.uni-saarland.de/protected/books/inex2014sbs.topics.xml.gz
8. http://www.librarything.fr/
9. http://www.amazon.fr/

<book><isbn>0001360000</isbn><title>Mog's Kittens</title><ean>9780001360006</ean>
<binding>Board book</binding><label>HarperCollins UK</label><listprice>$6.99
</listprice><manufacturer>HarperCollins UK</manufacturer><publisher>HarperCollins UK
</publisher><readinglevel>Ages 4-8</readinglevel><releasedate/><publicationdate>
1994-09-01</publicationdate><studio>HarperCollins UK</studio><edition/><dewey/>
<numberofpages>16</numberofpages><dimensions><height>55</height><width>461</width>
<length>469</length><weight>22</weight></dimensions><reviews><date>2007-11-27</date>
<summary>Cute Book</summary><content>cute board book for the cat lover or animal lover
author of "Hobo Finds A Home"</content><rating>5</rating><totalvotes>0</totalvotes>
<helpfulvotes>0</helpfulvotes></review></reviews>

Figure 2. *Exemple d'un livre de la collection d'INEX SBS 2014*

compte de manière approfondie des besoins d'informations exprimés au travers des requêtes. Nous pensons donc que l'exploitation du TAL peut s'avérer bénéfique et permettre de mieux comprendre les besoins exprimés au sein des requêtes et ainsi améliorer la recommandation. Préserver la structure de la phrase peut être une étape vers l'amélioration de la compréhension des besoins utilisateurs. Au vu des données disponibles *via* INEX SBS, nous proposons un modèle qui se veut une hybridation des systèmes de RI et des systèmes de recommandation classiques. En effet, notre modèle exploite à la fois des informations, issues des commentaires et des évaluations émis par les utilisateurs, ainsi que l'analyse de requêtes longues et détaillées exprimant le besoin informationnel de l'utilisateur. Ces requêtes longues et détaillées se retrouvent bien loin de la suite de mots-clés que l'on utilise en RI et sont porteuses d'éléments relatifs aux informations exploitées par les systèmes de recommandation comme les goûts ou encore les centres d'intérêt des utilisateurs qui peuvent, par ailleurs, s'apparenter à des profils utilisateurs.
Dans la section suivante, nous présentons l'architecture générale du système de recherche de livres que nous employons.

4. Architecture du système de recommandation de livres

La figure 3 représente l'architecture générale du système de recherche des livres dans lequel nos travaux s'inscrivent. Nous distinguons deux processus, d'un côté le traitement des requêtes longues et détaillées et de l'autre, le traitement de la collection de livres. Au niveau de la phase 1, les fonctions de représentation sont employées à la fois sur les requêtes et sur la collection. Nos travaux s'intègrent durant cette phase et plus particulièrement du côté de la requête. Nous implémentons notre approche *via* un prétraitement des requêtes qui consiste en une approche de classification automatique supervisée, suivant une taxonomie propre aux besoins exprimés dans les requêtes de recherche de livres. Afin de générer les modèles de classification, nous avons eu recours à deux outils Weka[10] et SVMLight[11]. Nous utilisons deux classes qui se nomment respectivement analogue et non analogue. Une fois cette classification effectuée, nous établissons plusieurs stratégies de représentation des requêtes analogues qui sont intégrées dans les fonctions de représentation des requêtes. Pour

10. `http://www.cs.waikato.ac.nz/ml/weka/`
11. `http://svmlight.joachims.org/`

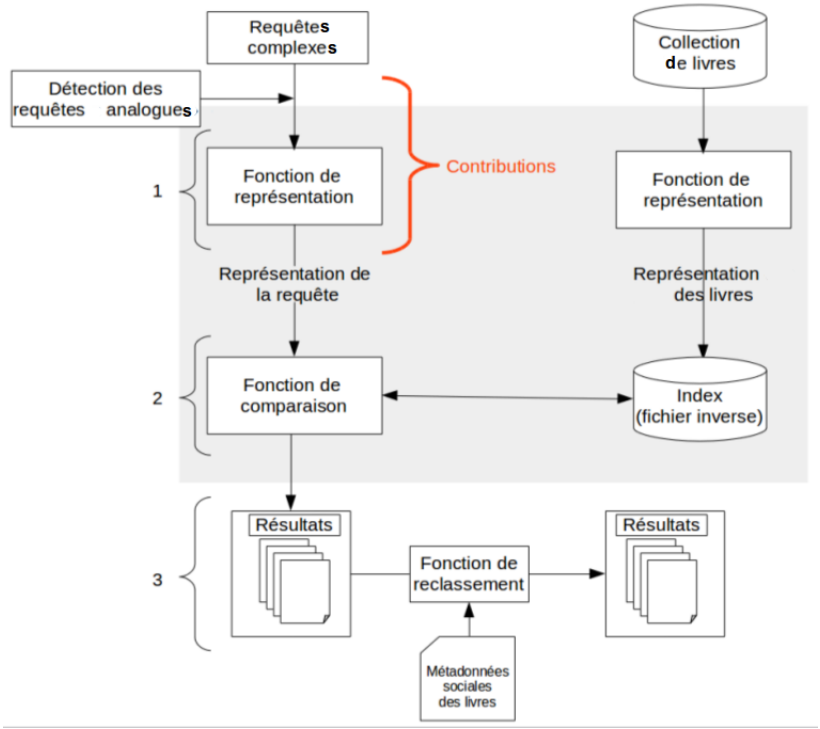

Figure 3. *Architecture du système de recommandation de livres*

effectuer ces différentes représentations nous utilisons un outil nommé Stanford Dependencies[12] (Marneffe *et al.*, 2006). Cet outil nous permet d'obtenir pour chaque requête une analyse en dépendance sous forme de bigrammes de mots associés au nom de la dépendance correspondante. Les informations supplémentaires apportées par ces bigrammes prennent la forme d'une nouvelle balise qui est utilisée lors de la fonction de représentation. La figure 4 présente un exemple d'une requête comprenant l'ajout de l'analyse en dépendance.

Du côté de la collection, la figure 5 présente l'exemple d'un livre présent dans l'index. Nous avons choisi d'indexer tous les champs de chaque livre et de les représenter sous la forme de sacs de mots.

Lors de la phase 2, le modèle de RI InL2 est utilisé *via* le logiciel Terrier[13]. Nous choisissons ce modèle car lors de la campagne INEX SBS 2014 nous avons obtenu la deuxième position (Benkoussas *et al.*, 2014). InL2 est un modèle DFR qui nous permet de calculer un score de pertinence estimé pour chaque livre de la collection selon

12. `http://nlp.stanford.edu/software/stanford-dependencies.shtml`
13. `http://terrier.org/`

```
<topic>
<nb>1584
<query>alternate history and alternative histories
<title>Great alternative history books?
<group>Time Travel, Alternate Histories and Parallel Worlds
<narrative>    I love alternative histories - two great ones I've enjoyed are Robert Harris's
Fatherland  and Kim Stanley Robinson's  Years of Rice and Salt .  Any other recommendations? John
<narrative_analyze> I love alternative histories two ones great ones ones enjoyed ones Fatherland
I enjoyed 've enjoyed are Fatherland Robert Harris Harris Fatherland Kim Robinson Stanley Robinson
Robinson Years Any recommendations other recommendations
</topic>
<topic>
```

Figure 4. *Exemple de représentation d'une requête analogue utilisée lors de la fonction de représentation*

```
<book>
<isbn>1871034000</isbn>
<text>1871034000 Medicine in Early Mediaeval England 9781871034004 Paperback Manchester Centre
for Anglo-Saxon Studies Manchester Centre for Anglo-Saxon Studies Manchester Centre for
Anglo-Saxon Studies 1989-05-01 Manchester Centre for Anglo-Saxon Studies 40 31 551 795 18 Marilyn
Deegan Editor D.G. Scragg Editor History Medical Subjects Books Refinements Binding (binding)
Paperback Format (feature_browse-bin) Printed Books</text>
```

Figure 5. *Exemple de représentation d'un livre utilisée lors de la fonction de représentation*

la requête posée. Les modèles DFR considèrent que, dans un document donné, plus la fréquence d'un mot s'écarte de sa fréquence d'apparition moyenne dans les documents de la collection, plus ce mot est représentatif du document considéré[14] (Robertson *et al.*, 1980). Dans InL2, L renvoie à la succession de Laplace (Wilson, 1927) pour la première normalisation et le 2 à la normalisation de fréquence des termes. Le poids de chaque mot est calculé de la manière suivante :

$$w_d(t,d) = \frac{1}{tf+1}\left(tf.log_2\frac{N+1}{n_t+0,5}\right)$$

où tf est la fréquence du terme t dans le document d. N le nombre de documents dans une collection D et n_t la fréquence du document.
Lors de la phase 3, une fois les résultats obtenus *via* InL2, nous effectuons un nouveau classement des documents *via* la mise en place d'un score social (score d'ordonnancement). Ce score d'ordonnancement est lié à une information précise qui est, dans notre cas, les votes (évaluations) émis par les utilisateurs. En d'autres termes, pour chaque document extrait pour une requête donnée nous calculons un score à partir des informations sociales générées par les utilisateurs telles que les votes et les commentaires. Ce score se fonde sur l'idée que plus un livre a de critiques et de bonnes évaluations plus il est intéressant.

$$Score\_d'ordonnancement(D) = \log(\#commentaires(D)) \times \frac{\sum_{r\in R_D} r}{\#commentaires(D)}$$

14. « *The more the divergence of the within-document term-frequency from its frequency within the collection, the more the information carried by the word t in the document d.* »

$\#commentaires(D)$ est le nombre de commentaires attribués à D, R_D renvoie à l'ensemble des commentaires de D. Après cela, nous effectuons une pondération entre les scores fournis par InL2 et le score social pour chacun des documents. Comme nous le soulignions précédemment, nos contributions, dans cet article, sont présentes au niveau de la phase 1 et plus particulièrement du côté de la requête.
La section suivante présente plus en détail les travaux réalisés afin d'améliorer la compréhension des besoins utilisateurs et plus particulièrement au sein des requêtes analogues.

5. Classification supervisée et analyse des requêtes longues et détaillées

L'utilisation d'une classification des requêtes peut permettre de mieux caractériser les besoins et ainsi d'adapter le traitement employé selon le type de requêtes. Cette section présente, dans un premier temps, la taxonomie propre aux besoins exprimés dans les requêtes de recherche de livres que nous avons mise en place, et dans un second temps, les différents types d'indexation de requêtes que nous avons effectués.

5.1. *Classification de requêtes longues et détaillées par approche supervisée pour la recommandation de lecture*

Des recherches ont porté sur la mise en place de taxonomies permettant d'identifier les besoins utilisateurs (Broder, 2002), (Rose et Levinson, 2004). Or, ces taxonomies très orientées sur l'analyse des requêtes Web s'appliquent difficilement à l'analyse de requêtes de recherche de livres. En effet, nos travaux sont orientés sur des requêtes très spécifiques dont l'intention est purement informationnelle, c'est-à-dire que l'utilisateur exprime un intérêt pour obtenir une information qui, dans notre cas, est exclusivement relative à la recherche de livres. Cependant, inspirés par ces recherches, nous avons transposé ces mêmes réflexions dans l'étude des requêtes de recherche de livres et nous avons pu clairement distinguer deux classes de requêtes analogues et non analogues. Comme nous le soulignions précédemment, les requêtes analogues englobent toutes les requêtes dans lesquelles l'utilisateur exprime un besoin visant à rechercher des similitudes entre des livres, des auteurs ou encore des collections. À l'opposé, les requêtes non analogues expriment des besoins variés. Parmi les besoins exprimés au sein des requêtes non analogues, nous avons établi trois sous-classes :

– *orientée* : une requête dont les termes viennent particulariser la recherche ;

– *non orientée* : une recherche plus générique sur une thématique ;

– *spécifique* : une recherche sur un livre particulier dont le titre est inconnu.

L'intégralité du jeu de requêtes des utilisateurs de LibraryThing est classifiée manuellement par trois annotateurs, chaque classe est ensuite choisie suite à un accord interannotateur. Pour la classe *analogue*, nous recensons 300 requêtes de ce type soit 44,2 % du corpus fourni par INEX SBS 2014 et pour la classe non analogue 380 requêtes, soit 55,8 % du corpus. Nous pouvons voir ci-dessous des exemples de requêtes

pour nos deux classes. Dans le cas des requêtes analogues, nous voyons en gras les groupes de mots qui nous paraissent les plus caractéristiques de ce type de requêtes. Ces groupes de mots précèdent souvent l'expression du goût du lecteur ou son désir de recommandation d'autres lectures.

i) Analogue :

- Are we going to pick a new book to group read soon ? I'd like to propose a Brandon Sanderson book since he just joined Green Dragon. **My preference is** Elantris . Does anyone else have **any thoughts or suggestions** ?
- I was **completely enchanted by** the story of Katherine and John of Gaunt. I'm wondering if **anybody could recommend** historical books with **similar quality** of writing and charismatic characters. Thank you !

ii) Non analogue :

i) non orientée :

- What bioethics books do you have in your collection ? Any recommendations ?

ii) orientée :

- Anyone got any suggestions on books dealing with early (pre-wwii) gay movements, especially outside germany and britain ?

iii) spécifique :

- I've been wracking my brain and searching the web for the answer to this. In which Stephanie Plum Novel does Ramirez come back ?

5.2. *Représentation des requêtes analogues*

Suite à la classification établie dans la section 5.1, nous nous penchons plus particulièrement sur la représentation des requêtes analogues. Nous présentons plusieurs stratégies de représentation des requêtes analogues *via* l'utilisation d'un analyseur en dépendance.
Nous supposons que parmi les dépendances trouvées certains types peuvent permettre de caractériser les besoins exprimés. Les dépendances que nous utilisons pour les expansions sont réduites, c'est-à-dire que les dépendances impliquant les prépositions, les propositions conjointes, ainsi que des informations sur les référents de clauses relatives sont réduites pour obtenir des dépendances directes entre les mots. Par exemple, pour les dépendances impliquant la préposition « *in* », nous avons :
$prep(based-7, in-8), pobj(in-8, LA-9) \Longrightarrow prep\_in(based-7, LA-9)$.
Ces réductions s'effectuent grâce à des listes préalablement définies de prépositions, de propositions conjointes, ainsi que sur les référents de clauses relatives. Ces

réductions se présentent sous la forme de bigrammes associés à la dépendance correspondante. Les nombres présents à côté des mots correspondent aux indices (ou index) de chaque terme au sein de la requête.
Nous présentons différentes études menées afin de raffiner la sélection des dépendances à des éléments permettant de caractériser le besoin d'informations exprimé par un utilisateur au sein des requêtes *analogues*. L'objectif de ces études est de nous fournir des bigrammes de mots permettant d'améliorer la représentation des requêtes et ainsi d'élargir les correspondances potentielles avec l'index fondé sur la représentation des livres.

5.2.1. *Étude fréquentielle*

Nous faisons une analyse fréquentielle qui nous permet d'avoir une vision globale des types de dépendance les plus redondants au sein des requêtes analogues. L'hypothèse émise est que, d'une part, certains types de dépendance sont moins porteurs d'informations, et que d'autre part, une redondance au niveau de certains types de dépendance peut être un vecteur d'informations sur les caractéristiques structurelles de ce type de requêtes et nous permettre ainsi de mettre en exergue les besoins informationnels. Nous retenons les dépendances nominales (nn) qui composent 7,29 % des dépendances présentes dans les requêtes analogues sur un total de 23 728 dépendances (3,38 % dans les requêtes non analogues). Les dépendances composées de prépositions sont également retenues car les prépositions sont des mots qui permettent une incidence, c'est-à-dire qu'ils établissent un rapport logique entre les mots. Parmi les prépositions jugées pertinentes pour la compréhension des besoins utilisateurs, nous pouvons citer : les prépositions composées avec *of* (prep_of) (2,63 % des dépendances, requêtes non analogues : 2,15 %), *to* (prep_to) (0,67 % des dépendances, requêtes non analogues : 0,53 %) et *about* (prep_about) (0,45 % des dépendances, requêtes non analogues : 0,36 %). La figure 6 présente l'exemple d'une requête après analyse, avec en gras, les types de dépendance relevés précédemment comme étant pertinents. Nous pouvons observer que l'analyse des requêtes effectuée par Stanford Dependencies nous fournit des bigrammes de mots associés à la dépendance correspondante. Concernant les nombres présents à côté des mots, nous n'utilisons pas cette information, nous la supprimons lors d'un prétraitement effectué sur la requête afin de ne conserver que les bigrammes de mots ainsi que le nom de la dépendance. Cette analyse nous permet d'extraire des bigrammes de mots, pouvant être non consécutifs, correspondant à des liens syntaxico-sémantiques potentiellement représentatifs des besoins d'informations exprimés dans les requêtes analogues que nous intégrerons par une expansion pour chacune des requêtes.

5.2.2. *Étude des modèles de classification*

Avec cette étude, nous souhaitons montrer que l'analyse des modèles de classification permet d'extraire des éléments caractéristiques des requêtes analogues. Nous pensons que les éléments caractéristiques des requêtes analogues peuvent s'assimiler à des amorces référant à l'expression des besoins de l'utilisateur (par exemple, l'ex-

nsubj(enchanted−4, I−1), cop(enchanted−4, was−2), advmod(enchanted−4, completely−3), root(ROOT −0, enchanted−4), det(story−7, the−6), prep_by(enchanted−4, story−7), **prep_of(story-7, Katherine-9)**, **prep_of(story-7, John-11)**, conj_and(Katherine−9, John−11), prep(story−7, of−12), dep(of−12, Gaunt−13), dobj(wondering−17, Gaunt−13), nsubj(wondering−17, I−15), aux(wondering−17, m−16), rcmod(Gaunt−13, wondering−17), mark(recommend−21, if−18), nsubj(recommend−21, anybody−19), aux(recommend−21, could−20), advcl(wondering−17, recommend−21), amod(books−23, historical−22), dobj(recommend−21, books−23), amod(quality−26, similar −25), prep_with(recommend−21, quality−26), **nn(characters-31, writing-28)**, conj_and(writing−28, charismatic−30), **nn(characters-31, charismatic-30)**, **prep_of(quality-26, characters-31)**, ccomp(enchanted−4, T hank−33), dobj(Thank−33, you−34)

Figure 6. *Résultat de l'analyse en dépendance pour la requête :* " I was completely enchanted by the story of Katherine and John of Gaunt. I'm wondering if anybody could recommend historical books with similar quality of writing and charismatic characters. Thank you ! "

pression « *similar to* »). De ce fait, extraire ces informations ainsi que les éléments voisins peut permettre de préciser les informations que l'utilisateur cherche à obtenir.

5.2.2.1. Étude du modèle généré par un classifieur bayésien multinomial naïf (MNB)

MNB est une version spécialisée de la classification naïve bayésienne qui est conçue pour les documents textes. Un classifieur bayésien naïf est un modèle à caractéristiques statistiquement indépendantes. En termes simples, ce classifieur suppose que l'existence d'une caractéristique pour une classe, est indépendante de l'existence d'autres caractéristiques. Dans le cas de MNB, ce classifieur modélise explicitement le nombre de mots et ajuste les calculs sous-jacents (McCallum et Nigam, 1998).
Afin d'extraire les bigrammes de mots les plus représentatifs des requêtes analogues, nous découpons ces requêtes ainsi que les requêtes non analogues *via* des fenêtres glissantes de deux mots afin d'obtenir des requêtes uniquement constituées de bigrammes. À partir de ces requêtes, nous appliquons le modèle MNB qui nous fournit un score de pertinence pour chaque bigramme de chacune des classes. Le tableau 1 présente une liste non exhaustive des bigrammes les plus représentatifs des requêtes analogues extraits par le modèle MNB. Une fois ces bigrammes extraits, nous les comparons aux

other, suggestions	other, books
I, enjoy	I, loved
just, finished	been, reading
I, finished	ve, read

Tableau 1. *Liste non exhaustive des bigrammes extraits par le modèle MNB*

dépendances produites par Stanford Dependencies afin d'extraire aussi tous les types de dépendance contenant ces bigrammes. Le voisinage direct de ces bigrammes est également extrait. À partir de ces groupes de bigrammes, nous établissons plusieurs représentations des requêtes *via* l'ajout d'une expansion. Pour ce modèle de classi-

fication, l'extraction du bigramme de mots précédant le bigramme courant donne les meilleures performances lors de la recommandation. Le tableau 2 présente un exemple d'expansion fondée sur des bigrammes de mots. Nous avons également, à partir des

Requête	Expansion
I love alternative histories - two great ones I've enjoyed are Robert Harris's Fatherland and Kim Stanley Robinson's Years of Rice and Salt . Any other recommendations ?	Salt Fatherland Any recommendations, Salt Fatherland I enjoyed, ones Fatherland Any recommendations

Tableau 2. *Exemple de requête avec une expansion fondée sur des bigrammes de mots basée sur les résultats du modèle MNB*

dépendances extraites *via* ce modèle de classification, établi des bigrammes fondés sur des catégories syntaxiques. Nous avons, à partir du voisinage du bigramme de catégories syntaxiques courant, établi une liste des bigrammes de catégories syntaxiques les plus récurrents. Une fois ces bigrammes de catégories syntaxiques repérés nous extrayons leur contenu afin de ne garder que les mots qui les composent. Dans ce cas-ci, les meilleures performances obtenues lors de la recommandation sont constatées suite à l'extration des deux bigrammes de catégories syntaxiques précédant et suivant le bigramme de catégories syntaxiques courant. Ce qui donne, par exemple, des bigrammes de catégories syntaxiques du type :

– *parataxis*[15]*, conjonction and, déterminant, modificateur adjectival, dépendant / objet direct, sujet nominal, auxiliaire, modificateur de rapport de clauses / nom, préposition by, marqueur, modificateur adverbial, auxiliaire*.

Le tableau 3 présente un exemple d'expansion fondé sur des bigrammes de mots réalisé à partir de bigrammes de catégories syntaxiques.

Requête	Expansion
I love alternative histories - two great ones I've enjoyed are Robert Harris's Fatherland and Kim Stanley Robinson's Years of Rice and Salt . Any other recommendations ?	've enjoyed ones Fatherland I enjoyed ones enjoyed enjoyed ones

Tableau 3. *Exemple de requête avec une expansion fondée sur des bigrammes de catégories syntaxiques basée sur les résultats du modèle MNB*

15. Coordination des phrases et des clauses sans conjonction de coordination.

5.2.2.2. Étude du modèle généré par C4.5 (J48)

J48 est un algorithme de classification supervisé fondé sur l'algorithme ID3 (Baltié, 2002) auquel il apporte plusieurs améliorations. Cet algorithme se fonde sur une mesure de l'entropie dans l'échantillon d'apprentissage pour construire son modèle.
Tout comme MNB, nous découpons les requêtes analogues et non analogues *via* des fenêtres glissantes de deux mots afin d'obtenir des requêtes uniquement constituées de bigrammes. À partir de ces requêtes, nous appliquons J48 qui construit un arbre de décision. Nous sélectionnons ensuite les branches composées de bigrammes dont le poids est le plus important pour notre classe analogue. Le tableau 4 présente un extrait des bigrammes établis par ce modèle de classification.

to, start	other, suggestion
I, reading	other, by
read, have	anyone, recommended
read, liked	finished, having

Tableau 4. *Liste non exhaustive des bigrammes extraits par le modèle J48*

Tout comme pour MNB, une fois ces bigrammes extraits nous les comparons aux dépendances produites par Stanford Dependencies afin d'extraire tous les types de dépendance contenant ces bigrammes. Nous extrayons ensuite les bigrammes de mots avoisinant le bigramme courant.
Pour ce modèle de classification, l'extraction des bigrammes de mots précédant et suivant le bigramme courant donne les meilleures performances lors de la recommandation. Le tableau 5 présente un exemple d'expansion fondé sur des bigrammes de mots.

Requête	Expansion
British/Irish authors I've read include Susan Cooper, C.S. Lewis and J.R.R. Tolkein . Can anyone recommend something strongly that might fit more or less into this vein ?	British/Irish authors authors read authors include

Tableau 5. *Exemple de requête avec une expansion fondée sur des bigrammes de mots basée sur les résultats du modèle J48*

Nous avons également, à partir des dépendances extraites *via* ce modèle de classification, établi une liste de bigrammes de catégories syntaxiques les plus récurrents. Pour ce modèle de classification, l'extraction des deux bigrammes de catégories syntaxiques précédant et suivant le bigramme courant donne les meilleures performances lors de la recommandation. Ce qui donne, par exemple, des bigrammes de catégories syntaxiques du type :

– nom, nom, préposition by / sujet nominal, auxiliaire, modificateur adverbial / conjonction and, auxiliaire, modificateur adverbial.

Le tableau 6 présente un exemple d'expansion fondé sur des bigrammes de mots réalisé à partir de bigrammes de catégories syntaxiques.

Requête	Expansion
British/Irish authors I've read include Susan Cooper, C.S. Lewis and J.R.R. Tolkein . Can anyone recommend something strongly that might fit more or less into this vein ?	Tolkein recommend Can recommend anyone recommend

Tableau 6. *Exemple de requête avec une expansion fondée sur des bigrammes de catégories syntaxiques basée sur les résultats du modèle J48*

Les résultats obtenus par ces différentes hypothèses sont présentés dans la section suivante.

6. Expérimentations

Dans cette section, nous présentons les mesures d'évaluation. Ensuite, nous détaillons les résultats obtenus par les différentes approches de classification supervisée des requêtes analogues et non analogues. Enfin, nous présentons les différentes indexations effectuées sur les requêtes analogues et l'impact sur la tâche de recommandation.

6.1. *Mesures d'évaluation*

Le tableau 7 présente les mesures d'évaluation utilisées. Concernant la classification on trouve la précision, le rappel ainsi que la F-mesure qui sont les mesures usuelles de la tâche INEX SBS. Pour la classification, supposons une classe i dans laquelle nous devons classer nos requêtes et supposons que le système donne pour cette classe **vp** requêtes vraies positives, **vn** requêtes vraies négatives, **fp** requêtes fausses positives, **fn** requêtes fausses négatives. Concernant l'évaluation des différents modèles de recherche de livres, nous utilisons les mesures suivantes : *Mean Reciprocal Rank* (MRR) et *Mean Average Precision* (MAP).

Afin d'attester de la significativité des résultats entre nos différents modèles de recherche de livres, nous utilisons le test des rangs signé de Wilcoxon[16] (Hull, 1993).

16. Le test de Wilcoxon est un test d'hypothèses statistiques non paramétriques utilisé lorsque l'on compare deux échantillons connexes, échantillons appariés, ou des mesures répétées sur un

Mesures pour la classification

Nom	Formule	Description
Précision	$P = \frac{vp}{vp+fp}$	Proportion de solutions trouvées qui sont pertinentes. Mesure la capacité du système à refuser les solutions non pertinentes.
Rappel	$R = \frac{vp}{vp+fn}$	Proportion des solutions pertinentes qui sont trouvées. Mesure la capacité du système à donner toutes les solutions pertinentes.
F1-mesure	$F = \frac{2PR}{P+R}$	Moyenne harmonique de la précision et du rappel. Mesure la capacité du système à donner toutes les solutions pertinentes et à refuser les autres.

Mesures pour les modèles de recherche de livres

Nom	Formule	Description
Mean Reciprocal Rank (MRR)	$MRR = \frac{1}{\|Q\|}\sum_{i=1}^{\|Q\|}\frac{1}{Rank_i}$	Le MRR est une mesure pour évaluer une liste de réponses possibles à un échantillon de requêtes, ordonné par la probabilité d'exactitude. Le rang inverse d'une réponse est l'inverse du rang de la première bonne réponse. Le rang inverse moyen est la moyenne des rangs réciproques pour un échantillon de requêtes Q.
Mean Average Precision (MAP)	$MAP = \frac{\sum_{q=1}^{A} veP(q)}{Q}$	La MAP correspond à la précision moyenne pour un ensemble de requêtes. En d'autres termes, la MAP est la moyenne des scores moyens de précision pour chaque requête.

Tableau 7. *Mesures d'évaluation*

6.2. *Expérimentations sur la classification automatique des requêtes*

Dans le cadre de ces expérimentations, nous comparons trois techniques de classification : « Machine à vecteurs de support »[17](SVM), MNB et J48. Pour l'implémen-

seul échantillon afin d'évaluer si leur moyenne de population des rangs est diffèrente. Le résultat de ce test est exprimé par une valeur p. Le procédé est généralement utilisé pour comparer la valeur de p à un seuil préalablement défini (typiquement 5 %). Si la p-valeur est inférieure au seuil, nous rejetons l'hypothèse nulle en faveur de l'hypothèse alternative, et le résultat du test est « statistiquement significatif ». Sinon, si la p-valeur est supérieure au seuil, nous ne rejetons pas l'hypothèse nulle, et nous ne pouvons rien conclure sur les hypothèses formulées.

17. *Support Vector Machine* : machine à vecteurs de support.

tation du SVM, nous utilisons l'outil SVMLight et pour l'implémentation de MNB et J48 l'outil Weka. L'objectif de ces expérimentations est de parvenir à détecter les requêtes analogues afin d'arriver à repérer les caractéristiques structurelles qui les composent.
Concernant les paramétrages effectués, nous établissons pour SVM une liste de mots les plus caractéristiques de chaque classe que nous utilisons comme attributs. Cette liste est réalisée grâce à deux algorithmes : *GainRatioAttribute* (GRA) et *InfoGainAttribute* (IGA). GRA consiste en une modification du gain de l'information qui permet de réduire sa polarisation. IGA permet d'effectuer un ratio d'informations pour acquérir les informations intrinsèques, il est utilisé pour réduire un biais en faveur des attributs à valeurs multiples. Après plusieurs tests, nous choisissons pour GRA d'utiliser une fréquence minimale d'apparition des termes de 3 (Fq3) combinée à une liste comprenant tous les mots *All Words* (AW). Pour IGA, nous avons choisi d'utiliser une fréquence minimale d'apparition des termes de 1 (Fq1) combinée à une liste pour laquelle nous avons supprimé les mots dont le score « Élimination récursive de caractéristiques »[18] (RFE) est égal à 0 (AW-0). Concernant les paramètres utilisés lors de la classification *via* MNB, nous optons pour des descripteurs binaires qui sont associés aux mots du vocabulaire *via* l'utilisation, sans paramètre, de la fonction StringToWordVector de Weka. La fréquence minimale d'un terme est réglée sur 1. Pour J48, nous optons également pour la fonction StringToWordVector. Nous choisissons de convertir tous les mots en minuscules. La fréquence minimale d'un terme est réglée sur 1. Concernant les paramètres internes de J48, nous modulons juste le facteur de confiance qui est utilisé pour configurer la taille de l'arbre de décision à 0,10.
Dans le tableau 8, nous pouvons observer que les meilleures performances sont ob-

Paramètres	Précision	Rappel	F-mesure
SVM IGA-AW-0Fq1	**96,8 %**	71,4 %	82,2 %
SVM GRA-AWFq3	90,5 %	**92,7 %**	**91,6 %**
MNB	79,7 %	79,3 %	79,3 %
J48	71,2 %	71,8 %	71,5 %

Tableau 8. *Évaluations de la classification des requêtes analogues et non analogues*

tenues par SVM, MNB et J48 présentent des résultats beaucoup plus faibles sur chacune des mesures. La meilleure F-mesure ainsi que le meilleur rappel sont obtenus suite à l'utilisation de la combinaison GRA-AWFq3 avec SVM. La meilleure précision, quant à elle, est observée sur la combinaison IGA-AW-0Fq1 avec SVM. Le net écart de performances entre SVM et les autre classifieurs peut s'expliquer par la très grande robustesse du SVM face à des données hétérogènes (Auria et Moro, 2008).

Les résultats obtenus lors de ces évaluations nous permettent d'établir que des caractéristiques structurelles permettent de qualifier les requêtes analogues. Dans la suite

18. *Recursive Feature Elimination* : élimination récursive de caractéristiques.

de nos expérimentations, nous exploitons ces caractéristiques structurelles extraites de l'étude des modèles générés par MNB et J48 afin d'établir plusieurs représentations des requêtes analogues.

6.3. *Expérimentations sur la représentations des requêtes analogues*

Dans cette section, nous présentons les résultats obtenus suite aux différentes modulations de représentation des requêtes analogues effectuées au sein du modèle de recherche de livres. Tous les modèles présentés utilisent le modèle de RI InL2. Le tableau 9 donne une description des caractéristiques des différents modèles utilisés.

Nom	**Description**
Baseline	- Indexation de tous les champs des livres - Indexation de tous les champs des requêtes
FullDep	- Indexation de tous les champs des livres - Indexation de tous les champs des requêtes dont une balise contenant l'ensemble des dépendances générées lors de l'analyse en dépendance
SelectDep	- Indexation de tous les champs des livres - Indexation de tous les champs des requêtes dont une balise contenant les résultats de l'étude fréquentielle
MNB	- Indexation de tous les champs des livres - Indexation de tous les champs des requêtes dont une balise contenant les bigrammes de mots résultant de l'étude du modèle de classification MNB
MNB Pattern	- Indexation de tous les champs des livres - Indexation de tous les champs des requêtes dont une balise contenant les patrons extraits de l'étude du modèle de classification MNB
J48	- Indexation de tous les champs des livres - Indexation de tous les champs des requêtes dont une balise contenant les bigrammes de mots résultant de l'étude du modèle de classification J48
J48 Pattern	- Indexation de tous les champs des livres - Indexation de tous les champs des requêtes dont une balise contenant les patrons extraits de l'étude du modèle de classification J48

Tableau 9. *Description des modèles de recherche de livres*

Le tableau 10 présente les résultats obtenus suite à nos différentes expérimentations. Nous constatons que les meilleurs résultats sont obtenus suite à l'utilisation du modèle J48. Le test de Wilcoxon nous permet de constater la significativité de ses per-

formances par rapport à notre modèle de référence Baseline avec une p-valeur pour la MAP de 0,04919 et une p-valeur pour le Recip_rank de 0,01968. Ces résultats nous confortent dans le fait que l'apport d'informations supplémentaires comme les liens syntaxico-sémantiques permet d'améliorer la pertinence de la recommandation. Le deuxième modèle obtenant des résultats sensiblement supérieurs à notre modèle de référence Baseline est J48 Pattern, le test de Wilcoxon donne pour la MAP une p-valeur de 0,04764 et pour le Recip_rank une p-valeur de 0,04788. Nous pouvons constater que l'exploitation des bigrammes générés par les différentes modulations effectuées sur le modèle de classification J48 est une piste prometteuse. Ce sont ces bigrammes qui permettent de mieux révéler les caractéristiques structurelles des requêtes analogues et ainsi améliorer les performances lors de la recommandation. Concernant les autres modèles, nous pouvons noter des performances sensiblement similaires sauf dans le cas du modèle SelectDep. Les performances peu concluantes de SelectDep peuvent s'expliquer par le choix des types de dépendance comme les noms qui ont pour effet de restreindre la représentativité des requêtes. En effet, les requêtes analogues sont composées de beaucoup de noms de livres et d'auteurs. Dans l'ensemble, ces résultats nous permettent d'établir que l'apport d'informations *via* l'utilisation de liens syntaxico-sémantiques permet d'améliorer sensiblement les performances lors de la recommandation. Bien que, sur l'ensemble des modèles, les résultats présentent des performances sensiblement similaires au modèle de référence Baseline ces résultats restent encourageants. Nous savons que par rapport à des tâches de RI *ad hoc* ou de RI Web qui obtiennent des MAP deux fois supérieures les résultats présentés ne démontrent pas un gain très important. Cependant, nous sommes face à une tâche qui n'est étudiée que depuis quelques années. De plus, l'emploi du TAL n'est que très peu exploité dans ce contexte. Ces premières expérimentations nous encouragent à exploiter le TAL comme moyen d'interprétation des besoins utilisateurs au sein des requêtes longues et détaillées.

Modèle	**Recip_Rank**	**MAP**
Baseline	0,1587	0,0339
FullDep	0,1481	0,0305
SelectDep	0.1366	0,0292
MNB	0.1549	0,0324
MNB Pattern	0,1510	0,0320
J48	**0.1616**	**0.0374**
J48 Pattern	0,1590	0,0348

Tableau 10. *Évaluations sur les différentes stratégies d'intégration de l'analyse en dépendance dans la représentation des requêtes analogues*

6.3.1. *Analyse qualitative des livres suggérés pour une requête analogue*

Dans cette sous-section, nous proposons l'exemple d'une analyse qualitative des livres suggérés pour une requête analogue par chacun des modèles présentés précédemment. Cette démarche a pour objectif, à plus grande échelle, de voir si certains

modèles fournissent des résultats plus performants selon la structure de certaines requêtes analogues. Cette constatation nous permettrait d'induire l'existence de sous-classes ou de facettes au sein même de la classe des requêtes analogues et ainsi de raffiner nos traitements lors de la phase de représentation des requêtes. Dans le cadre de la campagne INEX SBS 2014, un fichier contenant des livres associés à des valeurs de pertinence pour chaque requête est fourni (Qrels). Le processus de sélection des livres, lors de INEX SBS 2014, se fonde sur les suggestions proposées par les utilisateurs de LibraryThing. Les valeurs de pertinence attribuées pour chaque livre sont calculées en s'appuyant sur un arbre de décision qui regroupe tous les cas de figure possibles. Par exemple, un livre jugé positivement par un utilisateur et dont les caractéristiques répondent aux besoins exprimés au sein de la requête obtient une valeur de pertinence de 6. À l'inverse, un livre ne correspondant pas aux besoins exprimés au sein de la requête obtient une valeur de pertinence égale à 0. Les Qrels 2014 se compose de 8 918 livres associés à des valeurs de pertinence avec une moyenne de 8 livres par requête. La liste qui suit présente un extrait des livres obtenus pour la requête :

– « *I love alternative histories - two great ones I've enjoyed are Robert Harris's Fatherland and Kim Stanley Robinson's Years of Rice and Salt. Any other recommendations ? John* ».

Concernant les deux livres cités dans cette requête, tous les deux font référence à des histoires alternatives, c'est-à-dire, à des récits dans lesquels l'histoire est réécrite à partir de la modification d'un événement du passé. *Fatherland* conte une histoire dans laquelle les nazis ont gagné la guerre et *Years of Rice and Salt* remonte au XIV^{e} siècle siècle et imagine que la peste a tué 99 % de la population. Nous présentons pour chaque modèle les trois premiers livres retournés par le modèle de recherche accompagnés de commentaires permettant d'identifier la nature du livre.

- *Qrels 2014*

  i) **Titre :** *Pavane* **Auteur :** Keith Roberts. **Commentaires :** histoire alternative, fondée sur une histoire ramifiée autour de la mort de la reine Elizabeth et de l'Armada espagnole qui a réussi à conquérir l'Angleterre.

  ii) **Titre :** *The Yiddish Policemen's Union : A Novel* **Auteur :** Michael Chabon. **Commentaires :** À la fois polar, histoire d'amour, hommage aux Noirs des années 1940, et une exploration des mystères de l'exil et la rédemption.

  iii) **Titre :** *Fatherland* **Auteur :** Robert Harris. **Commentaires :** clairement cité dans la requête.

- *Baseline*

  i) **Titre :** *The Years of Rice and Salt* **Auteur :** Kim Stanley Robinson. **Commentaires :** clairement celui cité dans la requête.

  ii) le même livre dans une autre édition.

  iii) le même livre dans une autre édition.

- *FullDep*

  i) **Titre :** *The First Men In : U.S. Paratroopers and the Fight to Save D-Day* **Auteur :** Ed Ruggero. **Commentaires :** histoire d'une dangereuse mission confiée à un parachutiste durant la Seconde Guerre mondiale.

  ii) **Titre :** *Combat Jump : The Young Men Who Led the Assault into Fortress Europe, July 1943* **Auteur :** Ed Ruggero. **Commentaires :** fiction autour de la Seconde Guerre mondiale fondée sur des entrevues avec des anciens combattants de la 82e division aéroportée.

  iii) le même livre dans une autre édition.

- *SelectDep*

  i) **Titre :** *The Years of Rice and Salt* **Auteur :** Kim Stanley Robinson. **Commentaires :** clairement celui cité dans la requête.

  ii) le même livre dans une autre édition.

  iii) le même livre dans une autre édition.

- *MNB*

  i) **Titre :** *The Iron Lance* **Auteur :** Stephen Lawhead. **Commentaires :** trilogie épique du combat d'une noble famille écossaise pour son existence et sa foi au cours de l'âge des grandes croisades.

  ii) **Titre :***La Herejia* **Auteur :** Romain Sardou. **Commentaires :** thriller qui se déroule au Moyen Âge.

  iii) **Titre :** *Russia and the Soviet Union : An Historical Introduction from the Kievan State to the Present* **Auteur :** John M Thompson. **Commentaires :** Une introduction historique de l'État de Kiev à aujourd'hui.

- *MNB Pattern*

  i) **Titre :** *The Iron Lance* **Auteur :** Stephen Lawhead. **Commentaires :** trilogie épique du combat d'une noble famille écossaise pour son existence et sa foi au cours de l'âge des grandes croisades.

ii) **Titre :** *La Herejia* **Auteur :** Romain Sardou. **Commentaires :** thriller qui se déroule au Moyen Âge.

iii) **Titre :** *Summer Lightning* **Auteur :** Judith Richards. **Commentaires :** fiction fondée autour des camps de concentration.

- *J48*

i) **Titre :** *The master of the High Castle* **Auteur :** Philip K. Dick. **Commentaires :** fiction autour de la Seconde Guerre mondiale après la capitulation des alliés.

ii) **Titre :** *Letters Back To Ancient China* **Auteur :** Herbert Rosendorfer. **Commentaires :** mandarin chinois du X^{e} siècle siècle qui se déplace vers le XX^{e} siècle siècle dans sa machine à voyager dans le temps.

iii) **Titre :** *Summer Lightning* **Auteur :** Judith Richards. **Commentaires :** fiction fondée sur des camps de concentration.

- *J48 Pattern*

i) **Titre :** *Inside GHQ : The Allied Occupation of Japan and Its Legacy* **Auteur :** Eiji Takemae. **Commentaires :** compte rendu après l'occupation du Japon donnant un aperçu de l'état de japonnais contemporain.

ii) **Titre :** *Letters Back To Ancient China* **Auteur :** Herbert Rosendorfer. **Commentaires :** mandarin chinois du X^{e} siècle siècle se déplace vers le XX^{e} siècle siècle dans sa machine à voyager dans le temps.

iii) le même livre dans une autre édition

Suite à l'étude de ces résultats, nous pouvons établir la synthèse suivante :

- *Qrels 2014* ne répond que partiellement aux besoins exprimés dans la requête. Seul, le premier livre renvoie à une histoire alternative. Le deuxième est du type roman noir et le dernier fait clairement référence à *Fatherland* qui est déjà cité dans la requête.

- *Baseline* retourne trois fois le même livre, *The Years of Rice and Salt*, mais avec trois ISBN différents. Le titre de ce livre fait partie des livres présents dans la requête.

- *FullDep* renvoie à des livres de fiction dont les thématiques se rapprochent de *Fatherland* cité dans la requête.

- *SelectDep* présente les mêmes résultats que *Baseline*.

- *MNB* retourne deux livres correspondant à la thématique de *The Years of Rice and Salt*. Le troisième est, quand à lui, loin de répondre aux besoins exprimés dans la requête car la thématique et le genre littéraire ne sont pas bons.

- *MNB Pattern* renvoie deux livres de fiction dont les thématiques se rapprochent de *The Years of Rice and Salt* cité dans la requête. Et le troisième rentre dans le cadre

des besoins exprimés dans la requête de par sa thématique et son genre littéraire.

- *J48* fournit deux livres correspondant aux besoins exprimés dans la requête. Seul, le deuxième s'éloigne au niveau thématique et genre littéraire de ce qui est stipulé dans la requête.

- *J48 Pattern* retourne un premier livre correspondant à la thématique de *Fatherland*. Les autres livres ne correspondent pas aux besoins exprimés dans la requête car la thématique et le genre littéraire ne sont pas bons.

Ce que nous pouvons retenir de cette analyse est que les résultats varient selon le type d'expansion choisi. L'utilisation d'une expansion comme apport d'informations supplémentaires provoque bel et bien un impact non négligeable sur la recommandation. Concernant plus particulièrement les résultats obtenus pour cette requête, il est intéressant de constater que la grande majorité des livres sélectionnés tiennent compte d'au moins un des aspects exprimés dans la requête : le genre littéraire et la thématique. Ces résultats sont également intéressants car ils reflètent deux problématiques que l'on peut rencontrer en RI : la difficulté de se détacher des termes de la requête et le fait que chaque document est jugé indépendamment des autres ce qui provoque la présence de doublons.
Il est important de relever que dans l'évaluation des systèmes de recherche de livres fondés sur des requêtes, il est difficile de fournir une réponse optimale. En effet, le fait de partir d'une requête longue et détaillée exprimée par un utilisateur peut engendrer plusieurs interprétations de cette dernière. Il est donc difficile du côté des annotateurs qui établissent les jugements de pertinence et du côté des systèmes de fournir une réponse optimale face à toutes les contraintes énoncées dans la requête. De plus, dans le cadre plus spécifique des requêtes analogues, il est parfois difficile de juger quels sont les aspects intrinsèques du livre qui font que l'utilisateur recherche une similarité avec ledit ouvrage. Cette analyse nous conforte dans le fait que des sous-classes ou des facettes sont envisageables selon les caractéristiques de certaines requêtes analogues afin d'employer le modèle le plus adapté au type de requêtes analogues rencontrées. Comme nous le soulignions dans la section précédente, nous pouvons voir, par exemple, qu'utiliser la sélection des dépendances a pour effet de restreindre la représentativité de cette requête.

7. Conclusions

Les systèmes de recommandation ont tendance à considérer la formulation des requêtes et le processus de récupération comme une activité axée sur des tâches simples. Toutefois, les intentions cachées derrière les requêtes sont plus complexes et le processus de recherche devrait être orienté vers une caractérisation de ces tâches notamment par la prise en compte des besoins formulés par les utilisateurs. Dans cet article nous avons tenté de nous concentrer sur la compréhension ainsi que sur la représentation des besoins utilisateurs au sein des requêtes de recherche de livres et plus particulièrement au sein des requêtes analogues. Nous avons pu observer que notre méthodologie de classification des requêtes selon la taxonomie que nous avons établie offre des

résultats plus que satisfaisants avec une précision moyenne relevée pour nos deux algorithmes à plus de 90 % (cf : tableau 8). Concernant l'utilisation d'un analyseur en dépendance, nous avons pu constater que son utilisation dans l'analyse des requêtes analogues nous permet de faire un pas vers une meilleure interprétation des besoins exprimés par les utilisateurs. En effet, l'analyse plus fine des livres retournés pour une requête nous a permis de constater que l'apport d'informations supplémentaires par expansion permet d'améliorer la pertinence des résultats. Les tests de Wilcoxon nous ont permis d'observer des degrés de significativité différents sur certaines requêtes. Ce phénomème nous amène à penser que des sous-classes ou des facettes sont envisageables selon les caractéristiques de certaines requêtes analogues afin d'employer le modèle le plus adapté. De plus, l'analyse détaillée des résultats obtenus pour une requête analogue sur chacun de nos modèles nous permet de corroborer cette hypothèse. Dans nos futurs travaux, nous tenterons d'établir les caractéristiques des requêtes analogues en fonction des meilleures performances obtenues selon les modèles employés. À plus long terme, nous envisageons de mettre en place une stratégie de représentation propre au type de requêtes en fonction de la classe prédite lors de la classification automatique. Nous pensons, par exemple, pour les requêtes orientées extraire les termes qui viennent particulariser la requête.

8. Bibliographie

Albitar S., Fournier S., Espinasse B., « An Effective TF/IDF-Based Text-to-Text Semantic Similarity Measure for Text Classification », *Web Information Systems Engineering–WISE 2014*, Springer, p. 105-114, 2014.

Auria L., Moro R. A., « Support vector machines (SVM) as a technique for solvency analysis », *DIW Berlin Discussion Paper*, German Institute for Economic Research, 2008.

Baltié J., « DataMining : ID3 et C4. 5 », Epita SCIA, 2002.

Benkoussas C., Bellot P., « Book recommendation based on social information », *CLEF 2013 Evaluation Labs and Workshop Online Working Notes*, 2013.

Benkoussas C., Hamdan H., Albitar S., Ollagnier A., Bellot P., « Collaborative Filtering for Book Recommendation », *Working Notes for CLEF 2014 Conference*, 2014.

Benkoussas C., Ollagnier A., Bellot P., « Book Recommendation Using Information Retrieval Methods and Graph Analysis », *Working Notes for CLEF 2015 Conference*, CLEF, 2015.

Billerbeck B., Zobel J., « Efficient query expansion with auxiliary data structures », *Information Systems*, vol. 31, nº 7, p. 573-584, 2006.

Bonnefoy L., Deveaud R., Bellot P., « Do Social Information Help Book Search ? », *Workshop INEX.*, p. 109, 2012.

Broder A., « A taxonomy of web search », *ACM Sigir forum*, vol. 36, ACM, p. 3-10, 2002.

Collins-Thompson K., Callan J., « Query expansion using random walk models », *14th ACM conference on Information and knowledge management*, p. 704-711, 2005.

Cui H., Wen J.-R., Nie J.-Y., Ma W.-Y., « Probabilistic query expansion using query logs », *11th ACM conference on Digital libraries*, p. 325-332, 2002.

Furnas G. W., Landauer T. K., Gomez L. M., Dumais S. T., « The vocabulary problem in human-system communication », vol. 30, ACM, p. 964-971, 1987.

Gäde M., Hall M., Huurdeman H., Kamps J., Koolen M., Skov M., Toms E., Walsh D., « Overview of the SBS 2015 Interactive Track », *CLEF 2015 Evaluation Labs and Workshop Online Working Notes*, 2015.

Hall M. M., Huurdemann H., Koolen M., Skov M., Walsh D., « Overview of the INEX 2014 interactive social book search track », *Working Notes for CLEF 2014 Conference*, CLEF, p. 480-493, 2014.

Huang S., Zhao Q., Mitra P., Giles C. L., « Query expansion using topic and location », *7th IEEE International Conference on Data Mining*, p. 619-624, 2007.

Hull D., « Using statistical testing in the evaluation of retrieval experiments », *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, p. 329-338, 1993.

Li H., « A short introduction to learning to rank », *IEICE TRANSACTIONS on Information and Systems*, vol. 94, nº 10, p. 1854-1862, 2011.

Manish G., Bendersky M., « Information Retrieval with Verbose Queries », *Proposal for a Tutorial at SIGIR'15 Conference*, 2015.

Marneffe M.-C. D., MacCartney B., Manning C. D. *et al.*, « Generating typed dependency parses from phrase structure parses », *Proceeding of the 5th edition of the International Conference on Language Ressources and Evaluation*, LREC, p. 98-109, 2006.

McCallum A., Nigam K., « A comparison of event models for naive bayes text classification », *AAAI-98 workshop on learning for text categorization*, Citeseer, p. 41-48, 1998.

Metzler D., Croft B. W., « A Markov random field model for term dependencies », *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, p. 472-479, 2005.

Metzler D., Croft B. W., « Latent concept expansion using markov random fields », *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, p. 311-318, 2007.

Robertson S., van Rijsbergen C. J., Porter M., « Probabilistic Models of Indexing and Searching », *SIGIR*, p. 35-56, 1980.

Rose D. E., Levinson D., « Understanding user goals in web search », *Proceedings of the 13th international conference on World Wide Web*, ACM, p. 13-19, 2004.

Wasilewski P., « Query Expansion by Semantic Modeling of Information Need », *Proceedings of International Workshop CS&P*, 2011.

Wilson E. B., « Probable inference, the law of succession, and statistical inference », *Journal of the American Statistical Association*, vol. 22, nº 158, p. 209-212, 1927.

Wu S.-H., Liao P.-K., Lin H.-W., Hsu L.-J., Xiao W.-L., Chen L.-P., Ku T., Chen G.-D., « Query Type Recognition and Result Filtering in INEX 2014 Social Book Search Track », *CLEF 2014 Evaluation Labs and Workshop Online Working Notes*, 2014.

Zhang B.-W., Yin X.-C., Cui X.-P., Qu J., Geng B., Zhou F., Hao H.-W., « USTB at INEX2014 : Social Book Search Track », *CLEF 2014 Evaluation Labs and Workshop Online Working Notes*, 2014.