RICSH : Recherche d'information contextuelle par segmentation thématique de documents

Rachid Aknouche, Omar Boussaid, Fadila Bentayeb

Laboratoire ERIC, Université Lumière Lyon2
5 avenue Pierre Mendès-France, 69676 Bron Cedex, France

{Rachid.Aknouche, Omar.Boussaid, Fadila.Bentayeb }@univ-lyon2.fr

Résumé. Le but principal des systèmes de recherche d'informations (SRI) classiques est de retrouver dans un corpus de documents l'information considérée comme pertinente pour une requête utilisateur. Cette pertinence est souvent liée à la fréquence d'apparition des termes dans le texte par rapport au corpus sans tenir compte du contexte de la recherche. Partant de ce constat, nous proposons dans cet article une approche pour la recherche d'information contextuelle par segmentation thématique de documents (RICSH). Cette approche s'appuie sur la méthode de pondération *tf-idf* que nous avons adaptée dans notre cas pour indexer le corpus. Cette adaptation se situe au niveau de l'importance du terme et de son pouvoir de discrimination par rapport aux fragments de textes et non au corpus. Ces fragments sont obtenus grâce à un processus d'identification des unités thématiques les plus pertinentes pour chaque document.

1 Introduction

Les systèmes de recherche d'informations classiques utilisent des méthodes de pondération et des mesures de similarités pour retrouver des textes pertinents par rapport à une requête utilisateur. La pondération *tf-idf*[1] est l'une des techniques les plus utilisées dans ces systèmes. Elle permet d'évaluer l'importance d'un terme dans un document par rapport à une collection ou à un corpus. Cependant cette formule, telle qu'elle est souvent présentée dans la littérature, ne tient pas compte du contexte de la recherche d'information (RI). On entend par contexte l'endroit d'apparition des termes recherchés dans un document. Il s'agit notamment des fragments de texte qui sont représentés sous forme de sections et de paragraphes dans un document. Par contre, une recherche pertinente devrait considérer ces éléments lors de la phase de formulation de la requête, dans le processus d'appariement et/ou dans la phase de classement des résultats. Les approches classiques de la RI utilisent les premières techniques, dites traditionnelles, qui sont basées sur la représentation linéaire des documents. D'après (Zargayouna, 2004), ces techniques procèdent à des requêtes plates (recherche par mots clés) et ignorent, par conséquent, la

1. désigne un ensemble de schémas de pondération de termes. *tf* signifie *(Term Frequency)* qui désigne le nombre d'occurrence du terme dans le document et *idf (Inverted Document Frequency)* qui est la valeur inverse du nombre de documents dans lesquels le terme est présent ou le pouvoir de discrimination de ce terme.

structure du document. De leurs côtés, (Pinel-Sauvagnat et Boughanem, 2005) stipulent que la granularité des réponses renvoyées aux utilisateurs est restreinte au document tout entier. Or, les documents ne sont plus considérés en tant qu'entités atomiques, mais comme des agrégats d'objets en corrélation qui peuvent être recherchés séparément (Chiaramella, 2001). D'autres approches plus complexes sont apparues pour pallier aux problèmes liés à la recherche sémantique de l'information. Elles sont organisés autour d'une source de connaissances qui est souvent modélisée au moyen d'une ontologie pour approximer les concepts de la requête.

Malheureusement, ces approches ne sont pas adaptées aux documents non structurés. De plus, elles ne tiennent pas compte de la pertinence liée au contexte. La prise en charge de ces éléments dans le cadre d'une nouvelle approche de la RI a constitué, d'ailleurs, notre première motivation pour proposer *RICSH* comme solution pour une **r**echerche d'**i**nformation **c**ontextuelle par **s**egmentation t**h**ématique de documents.

2 Approche RICSH

L'approche RICSH comprend deux phases pour le processus de recherche d'information dans une collection de documents hétérogènes et chacune des phases est constituée de plusieurs étapes de traitement. La première phase est celle du prétraitement du corpus qui consiste, dans un premier temps, à extraire toutes les unités textuelles (mots) contenues dans ces documents, élaguer les mots fonctionnels [2] pour ensuite repérer les unités thématiques de chaque document. La deuxième phase, quant à elle, concerne le processus d'exécution de la recherche qui est basé sur les mécanismes de représentation des documents, sur les techniques utilisées pour la comparaison des documents par rapport à une requête utilisateur, et enfin sur la classification des résultats obtenus en des classes distinctes. Nous illustrons notre approche par une étude de cas organisée autour d'une collection de documents hétérogènes représentant des Curriculum Vitæ (CV) de personnes de différents niveaux d'instruction.

2.1 Phase de prétraitement

C'est une phase de préparation et de nettoyage des données texte. Elle consiste d'abord à parser les documents pour en extraire les unités textuelles. Ensuite, celles-ci sont prétraitées selon une source de connaissances et enfin passées de leur forme linéaire à une représentation hiérarchique grâce au processus d'identification et d'extraction des unités thématiques des documents. La source de connaissance que nous avons utilisé pour traiter la langue anglaise est la base de données lexicale Wordnet [3]. Par contre, pour ce qui des mots français, nous avons conçu un thésaurus que nous avons alimenté à partir d'un dictionnaire. Nous avons également utilisé lors de l'étape de stemmatisation (steming) l'algorithme de Porter pour l'anglais et l'algorithme proposé dans le projet *CLEF* pour le français (Rizoiu et al., 2010). Cette phase est composée de quatre étapes : (1) Extraction des entités textuelles, (2) Filtrage des mots fonctionnels, (3) stemmatisation et (4) Identification et extraction des unités thématiques. Ces étapes sont détaillées davantage dans le rapport technique.

2. ce sont les mots vide de sens ou *stopwords* en anglais qui n'intéressent pas la recherche d'information.

3. WordNet (Miller, 1995) est une base de données lexicale développée depuis 1985 par des linguistes du laboratoire des sciences cognitives de l'université de Princeton. C'est un réseau sémantique de la langue anglaise, qui se fonde sur une théorie psychologique du langage. La première version diffusée remonte à juin 1991.

Identification et extraction des unités thématiques Il s'agira de faire collaborer des outils de TAL (Traitement Automatique de la Langue) traditionnellement utilisés en RI avec la prise en compte des indicateurs lexicaux définis au préalable pour le découpage d'un document en unités thématiques. Une unité thématique désigne les fragments de nature textuelle d'un document faisant référence à un seul thème. Notre démarche s'appuie sur un thésaurus qui englobe l'ensemble des indicateurs lexicaux jugés importants pour ce processus de découpage. Ils sont crées, dans le cadre de cette étude, d'une façon manuelle, par contre pour élargir la couverture des unités thématiques dans un document, nous avons rajouté les synonymes et les différentes formes lexicales que peuvent avoir ces unités dans le corpus. Les fragments de texte obtenus sont ensuite regroupés par unité thématique. La recherche d'information liée au contexte d'apparition des termes s'effectuera, par contre, par rapport à ces fragments. Cette façon de faire permettrait un gain en temps d'exécution et une amélioration dans la pertinence des résultats.

2.2 Phase de représentation et de classification des documents

Indexation du corpus Les *stemmes* obtenus lors de la phase de prétraitement constitueraient des éléments essentiels pour indexer le corpus. Ils sont représentés comme des vecteurs dans un espace vectoriel à *n* dimensions où *n* est le nombre total de *stemmes*. La matrice générée dans notre étude de cas permet de ressortir une relation "*stemme-fragments*". En effet, les lignes de cette matrice représentent les fragments par unité thématique et les colonnes désignent les *stemmes* extraits de ces fragments. Chaque valeur de la matrice détermine la pertinence du terme par rapport au fragment de texte. Elle est obtenue par l'adaptation de la formule *tf-idf*.

Adaptation de la mesure *tf-idf* pour l'indexation de corpus Pour considérer les unités thématiques lors de la phase d'indexation du corpus, nous calculons la fréquence d'occurrences $TF_{(t,frag_i)}$ d'un terme t par rapport au fragment i de l'unité thématique (U_i). Cependant, la formule *tf-idf* adaptée à ce besoin combinera les deux critères : l'importance du terme dans un fragment de texte et son pouvoir de discrimination par rapport à l'ensemble des fragments. Ainsi, un terme d'une valeur de *tf-idf* adaptée élevée est considéré important dans le fragment, et il doit peu apparaitre dans les autres fragments. La valeur de fréquence d'un terme est :

$$\boldsymbol{TF}_{(t,frag_i)} = \frac{\boldsymbol{F_{t,frag_i}}}{\boldsymbol{N_{frag_i}}}$$

$F_{t,frag_i}$: désigne la fréquence d'apparition du terme t dans le fragment $frag_i$
N_{frag_i} : le nombre total des termes existants dans le fragment $frag_i$
Le calcul de la valeur de discrimination d'un terme par rapport aux fragments est :

$$\boldsymbol{IDF}_{(t,U_i)} = \boldsymbol{log}\,(\frac{\boldsymbol{N_{U_i}}}{\boldsymbol{n_{t,U_i}}})$$

N_{U_i} : le nombre total de fragments dans l'unité thématique i
n_{t,U_i} : le nombre de fragments dans l'unité thématique i contenant le terme t

Ainsi, nous obtenons la formule suivante pour chaque terme contenu dans le fragment :

$$\boldsymbol{TF_{(t,frag_i)}} \cdot \boldsymbol{IDF_{(t,U_i)}} = \frac{\boldsymbol{F_{t,frag_i}}}{\boldsymbol{N_{frag_i}}} \cdot \boldsymbol{log}\,(\frac{\boldsymbol{N_{U_i}}}{\boldsymbol{n_{t,U_i}}})$$

Cette phase d'indexation permettrait ainsi de représenter dans un espace vectoriel à *n* dimensions la pondération des *stemmes* par rapport aux fragments de l'unité thématique. Pour chaque stemme *t* contenu dans le fragment *i* de l'unité thématique *j*, désignée par $frag_i^{U_j}$, on calcule sa valeur de pondération $TF_{(t,frag_i)}.IDF_{(t,U_j)}$ présentée dans les tableaux suivants par $V_{(t_i,frag_i)}$.

Unité thématique : U$_1$				
Fragments	t_1	t_2	...	t_n
$frag_1^{U_1}$	$v_{1,1}$	$v_{1,2}$	...	$v_{1,n}$
:	:	:	...	:
$frag_n^{U_1}$	$v_{n,1}$	$v_{n,2}$	...	$v_{n,n}$

TAB. 1: Indexation de l'unité thématique U_1

Unité thématique : U$_n$				
Fragments	t'_1	t'_2	...	t'_n
$frag_1^{U_n}$	$v'_{1,1}$	$v'_{1,2}$	...	$v'_{1,n}$
:	:	:	...	:
$frag_n^{U_n}$	$v'_{n,1}$	$v'_{n,2}$	...	$v'_{n,n}$

TAB. 2: Indexation de l'unité thématique U_n

Reformulation de la requête L'étape de reformulation de la requête consiste à enrichir la requête utilisateur avec des informations liées au contexte de la recherche avant le lancement du processus d'appariement et d'indexation. Les informations du contexte sont fournies grâce : (1) au processus d'expansion de la requête généré en fonction des synonymes en communs des termes de la requête initiale ; (2) au choix de l'utilisateur de l'unité thématique correspondant à ses critères de recherche. Pour traiter la synonymie, nous avons intégré dans *RICSH* un module permettant de retrouver, pour chaque terme de la requête utilisateur, la liste des synonymes lui correspondant dans le dictionnaire et de ne garder que ceux qui sont en communs. Nous utilisons *WordNet* pour retrouver les synonymes des mots en anglais. Par contre pour les mots en français nous avons conçu un thésaurus des synonymes qui regroupe 36200 mots recensés dans le dictionnaire de la langue française. La requête est ensuite prétraitée à son tour, pour qu'elle soit représentée par un vecteur de terms selon la composition de la matrice générée pour le corpus lors de l'indexation. Cette opération permet, ainsi, d'augmenter davantage les chances de retrouver soit les documents qui correspondent le mieux aux mots contenus dans la requête utilisateur, soit ceux qui se rapprochent le plus sémantiquement de ces critères

Phase de comparaison Une fois la matrice construite, la similitude entre fragments et requête peut être calculée selon différentes méthodes. Dans *RICSH*, nous utilisons une métrique à base de cosinus. Cette mesure a l'avantage d'être simple et d'avoir de bonnes performances. Elle permet de calculer l'angle entre deux vecteurs (Aouicha, 2009). La valeur du cosinus est normée (entre 0 et 1). Si le cosinus tend vers 1 alors les deux documents sont proches, sinon s'il tend vers 0 alors ils sont éloignés. Un document est représenté par un vecteur $\vec{d} = (t_1, t_2, ...t_i, ...t_n)$, où $t_i \in [0, 1]$ est le poids d'un terme *i* dans le document. Une requête est également représentée par un vecteur $\vec{q} = (q_1, q_2, ...q_i, ...q_n)$, où $q_i \in [0, 1]$ est le poids du terme *i* dans la requête.

La fonction de correspondance mesure donc la similarité entre le vecteur requête et les vecteurs correspondant aux fragments de texte. Elle est définie comme suit :

$$\mathbf{Cos}(\vec{\mathbf{d}}, \vec{\mathbf{q}}) = \frac{\vec{\mathbf{d}}.\vec{\mathbf{q}}}{\| \vec{\mathbf{d}} \| . \| \vec{\mathbf{q}} \|} = \frac{\sum_{i=1}^{n} \mathbf{t}_i.\mathbf{q}_i}{\sqrt{\sum_{i=1}^{n} \mathbf{t}_i^2 + \sum_{i=1}^{n} \mathbf{q}_i^2}}$$

3 Expérimentations

3.1 Implémentation de l'approche RICSH

Pour évaluer la performance de l'approche RICSH, nous avons implémenté en Java, sous l'environnement Eclipse, des programmes illustrant chaque étape de notre démarche. Notre prototype s'appuie sur JWNL (Java WordNet Library), une API permettant un accès facile au thésaurus WordNet pour retrouver les synonymes d'un mot et sur SAX (Simple API for XML) pour parser un document XML. Nous avons également implémenté sous le SGBDR Oracle une base de données qui héberge : (1) le thésaurus des synonymes de mots ; (2) les unités thématiques générées lors du processus d'identification et d'extraction et (3) les scores de pertinence des résultats. Pour faciliter le processus d'évaluation, nous avons développé une interface permettant à l'utilisateur d'introduire ses requêtes, de combiner ses critères de recherche et enfin d'afficher les résultats selon leur degrés de pertinence par rapport à sa requête.

3.2 Évaluation des performances et résultats

Pour tester notre approche nous avons pris un corpus constitué de 385 documents CVs. Il s'agit des CVs de professeurs, de maîtres de conférences et d'étudiants. Ces CVs existent sous différents format (Pdf, Word et Html). Pour tester la performance et la fiabilité des résultats obtenus, nous avons effectué des expérimentations qui comparent les résultats obtenus à partir d'une recherche d'information classique (par mots clés) avec ceux obtenus d'une recherche considérant, à la fois, le contexte d'apparition des termes dans un fragment de texte et la synonymie des termes de la requete. Pour évaluer la pertinence des résultats, nous avons utilisé les mesures Rappel (R) et Précision (P) habituelles dans l'évaluationn des systèmes de la RI.

$$R = \frac{\textit{Nombre de docs pertinents trouvés}}{\textit{Nombre total de docs pertinents}} \text{ et } P = \frac{\textit{Nombre de docs pertinents trouvés}}{\textit{Nombre total de docs sélectionnés}}$$

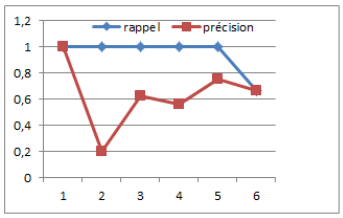

FIG. 1: *Courbes d'une RI classique*

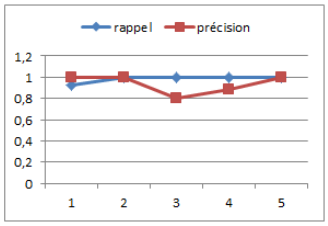

FIG. 2: Courbes RICSH

Les résultats de notre approche, présentées par les courbes de la figure (cf Fig.2), montrent une nette amélioration dans les mesures de rappel et de précision par rapport à ceux des méthodes classiques. Cette amélioration réside particulièrement dans les taux de précision. Elle est obtenue grâce à la nouvelle technique d'indexation du corpus issue de l'adaptation de la mesure *tf-idf* et au processus d'expansion de la requête généré en fonction des synonymes en communs des termes de requêtes. Ces résultats expliquent notamment notre première motivation de considérer le contexte de la recherche dans notre approche RICSH.

4 Conclusion

Nous avons présenté dans cet article une approche de recherche d'information contextuelle par segmentation thématique de documents (RICSH). Elle repose sur huit étapes de traitement regroupées dans deux phases principales : Une de prétraitement du corpus et une de représentation et de classification des documents. Une des particularités de notre approche par rapport aux systèmes de recherche d'informations classiques est de considérer le contexte lié à la recherche d'information. L'originalité de notre démarche est donc double. D'une part, nous définissons des techniques de repérage et d'extraction des unités thématiques des documents. D'autre part, nous proposons une adaptation de la mesure *tf-idf* pour évaluer la pertinence d'apparition des termes dans ces documents. L'intérêt de considérer de telles unités thématiques dans notre prototype a permet de donner des résultats satisfaisants par rapport aux méthodes classiques. Ces résultats sont, d'ailleurs, étayés dans la phase d'expérimentation par des calculs de valeurs de rappel et de précision pour des requêtes utilisateurs.

Dans les travaux futurs, nous comptons approfondir notre approche par l'introduction d'autres métriques telle que la divergence de Kullback-Leibler prenant davantage en compte la notion de contexte dans la RI. Une des perspective consiste à intégrer ces approches d'analyse de textes dans un cadre d'analyse en ligne OLAP (*On-Line Analytical Processing*).

Références

Aouicha, M. (2009). *Une approche algébrique pour la recherche d'information structurée*.

Chiaramella, Y. (2001). Lectures on information retrieval. Chapter Information retrieval and structured documents, pp. 286–309. New York, NY, USA : Springer-Verlag New York, Inc.

Pinel-Sauvagnat, K. et M. Boughanem (2005). A la recherche de noeuds informatifs dans des corpus de documents XML. In *Conférence francophone en Recherche d'Information et Applications (CORIA), Grenoble, 09/03/2005-11/03/2005*, pp. 119–134. IMAG.

Rizoiu, M.-A., J. Velcin, et J.-H. Chauchat (2010). Regrouper les données textuelles et nommer les groupes à l'aide des classes recouvrantes. In *10ème Conférence EGC 2010, Hammamet, Tunisie*, Volume E-19 of *Revue des Nouvelles Technologies de l'Information*, pp. 561–572.

Zargayouna, H. (2004). Contexte et sémantique pour une indexation de documents semi-structurés. In *CORIA*, pp. 161–178.

Summary

The principal goal of traditional Information Retrieval Systems (IRS) is to find relevant information for a user requests in a corpus of documents. This relevance, often relates to frequency appearance terms in the text without taking into account of their contex. We propose in this article an approach for a contextual information retrieval by topic segmentation of documents (*RICSH*). In this approach, we adpat *tf-idf* method to index the corpus. This adaptation focuses on term importance considering the fragment of texts rather than corpus. The fragmentation is obtain by an identification process of most relevant thematic units for each document.