Système de recherche d'information à base d'inclusion graduelle

L. Ughetto[1]— **O. Pivert**[2]— **V. Claveau**[3]— **P. Bosc**[2]

1 *IRISA - Université Rennes 2 - Campus de Beaulieu, F-35042 Rennes cedex, France*
2 *IRISA - ENSSAT - BP 80518, F-22305 Lannion, France*
3 *IRISA - CNRS - Campus de Beaulieu, F-35042 Rennes cedex, France*
{laurent.ughetto,vincent.claveau}@irisa.fr
{pivert,bosc}@enssat.fr

RÉSUMÉ. *Cet article étudie, d'un point de vue expérimental, l'apport des inclusions graduelles issues de la théorie des ensembles flous pour la modélisation d'un système de recherche d'information (SRI), comme l'ont proposé de manière théorique (Bosc* et al., *2008b). Documents et requêtes sont représentés par des ensembles flous, appariés par des opérateurs flous, dont le choix est crucial pour obtenir un système adapté à la RI. S'ils sont bien choisis, le SRI flou obtenu est proche des SRI classiques et obtient des résultats aussi bons, en conservant l'avantage de son cadre théorique fort. À l'inverse, l'examen d'opérateurs inadaptées à la RI souligne les propriétés requises par ce SRI flou. Enfin, nous montrons la valeur ajoutée de ce modèle flou, qui permet d'envisager des extensions du modèle très naturelles. Un exemple simple montre comment utiliser une base de liens morphologiques entre mots dans ce cadre.*

ABSTRACT. *This paper investigates, from an experimental point of view, the use of graded inclusions, from fuzzy sets theory, to model an information retrieval system (IRS) as theoretically proposed by (Bosc* et al., *2008b). Documents and queries are represented by fuzzy sets, which are paired with fuzzy operators. It is shown that the fuzzy logic settings are crucial in order to obtain a system suited for IR. With appropriate settings, it is possible to mimic classical systems, thus yielding results rivaling those of state-of-the-art systems, while preserving its strong theoretical framework. Conversely, examining operators providing poor results sheds light on the necessary properties of such a system. Last, the added-value of using this model is shown by considering possible extensions; in particular, a short experiment shows how one can make the most of morphological links in fuzzy IRSs.*

MOTS-CLÉS : *modèles de SRI, logique floue, inclusion graduelle, implication floue, expressivité*

KEYWORDS: *IRS models, fuzzy logic, graded inclusion, fuzzy implication, query expressiveness*

1. Introduction

Recherche d'information (RI) et bases de données (BD) partagent l'objectif de fournir aux utilisateurs les informations qu'ils demandent. Cependant, il est bien connu que les approches classiques d'interrogation utilisées en BD ne sont pas utilisables en RI. Tout d'abord, elles n'ont pas la flexibilité requise en RI pour réaliser un appariement approximatif entre les mots des requêtes et les documents. Ensuite, elle proposent rarement un moyen de classer les résultats fournis. Cependant, de récentes études sur les BD floues et l'interrogation flexible des BD ont apporté de nouveaux mécanismes théoriques d'interrogation plus adaptés à la RI, comme par exemple (Bosc *et al.,* 2008b).

L'objectif principal de notre étude est de montrer la validité expérimentale du cadre théorique proposé dans (Bosc *et al.,* 2008b), et de construire un SRI flou, à base d'inclusion graduelle. Dans ce modèle, documents et requêtes sont représentés par des ensembles flous, appariés par une inclusion graduelle, c'est-à-dire en utilisant des opérateurs comme les implications floues et les T-normes. Les principaux paramètres du modèle sont l'implication floue qui calcule le degré d'inclusion, la T-norme d'agrégation des scores, et les pondérations de termes dans les documents et les requêtes. Pour définir les poids de façon automatique, des schémas de pondération classiques sont utilisés. Des expérimentations, dont seulement quelques résultats sont détaillés en Section 5.3, ont été menées en faisant varier ces nombreux paramètres. Avec un bon choix d'opérateurs flous et de pondérations, des résultats positifs montrent que le SRI flou proposé obtient des résultats comparables aux systèmes de type OKAPI. Les résultats négatifs, quant-à eux, permettent de définir des propriétés que le SRI flou doit vérifier pour être performant. Elles montrent ainsi comment le cadre théorique de (Bosc *et al.,* 2008b) peut être adapté à la RI. Certaines de ces propriétés sont d'ailleurs bien connues en RI.

Enfin, cette étude montre que le modèle de SRI flou proposé est susceptible d'offrir de meilleures interactions avec l'utilisateur, et lui permet d'écrire des requêtes plus expressives ou de pondérer facilement les termes de sa requête pour exprimer des préférences, ou encore inclure des informations négatives.

Tout d'abord, la section 2 compare notre travail à des approches existantes utilisant la logique floue (LF) en RI. Ensuite, la section 3 présente brièvement le contexte théorique des inclusions graduelles (le lecteur peut se reporter à (Bosc *et al.,* 2008b) pour plus de détails). L'implémentation et les résultats expérimentaux du SRI flou sont ensuite détaillés et commentés dans les sections 4 et 5, en reprenant principalement les résultats obtenus dans (Bosc *et al.,* 2009). Enfin, diverses extensions possibles du modèle sont proposées en section 6.

2. Travaux connexes

Certains aspects de la théorie des ensembles flous (ou de la logique floue) ont été utilisés dans les modèles de RI depuis le début des années 1980 (Buell, 1982, inter

alia). C'est assez naturel dans la mesure où le modèle de RI Booléen a rapidement été étendu à des modèles utilisant des degrés, et que la LF est une extension de la logique Booléenne utilisant des degrés (de vérité). Plusieurs travaux ont introduit de la LF dans des modèles de RI, avec des objectifs bien différents. Cela a été fait par exemple pour gérer l'incertitude dans la représentation des termes (Kraft *et al.,* 2006), améliorer le classement des documents (Boughanem *et al.,* 2005), ou accroître l'expressivité du langage de requête... D'autres ont étendu le modèle classique de RI pour prendre en compte des situations particulières, comme l'utilisation d'échelles de pondération ordinales des termes (Herrera-Viedma, 2001), ou utiliser à la fois des mesures de possibilité et de nécessité pour pondérer les termes (Brini *et al.,* 2005). Mais la plupart de ces travaux n'utilisent la LF que de façon ponctuelle, pour traiter un problème particulier, alors que notre approche propose un cadre théorique complet.

Parmi les travaux qui utilisent de la logique floue dans le mécanisme d'appariement entre requêtes et documents, on peut noter les travaux récents de Herrera-Vielma et al. (Herrera-Viedma *et al.,* 2007) ou Oussalah et al. (Oussalah *et al.,* 2008). Ces derniers proposent aussi d'utiliser une implication floue pour calculer la similarité entre un document et une requête. Toutefois, leur modèle calcule $D \rightarrow Q$, comme cela se fait souvent dans les approches logiques de la RI (voir (Lalmas, 1998)), alors que l'implication est utilisée dans l'autre sens dans notre modèle, pour des raisons détaillées dans la section suivante.

Notre modèle pourrait aussi être comparé à des travaux comme ceux de Salton et al. (Salton *et al.,* 1983), dans la mesure où c'est aussi une extension du modèle Booléen, qui se situe entre ce modèle Booléen et l'approche algébrique (VSM).

3. RI et division de relations

Les systèmes de recherche d'information (SRI) sont fondés sur des modèles caractérisés par 3 aspects principaux : la représentation des documents, le langage de requête et le mécanisme d'appariement. Cette section montre que le SRI flou proposé ici est une généralisation du modèle Booléen sur ces 3 aspects. Les sous-sections détaillent successivement l'approche Booléenne et la façon dont elle est liée à la division de relations, puis comment l'extension de la division aux relation floues (ou graduelles) est liée à une approche de RI graduelle, et enfin les fondements théoriques de notre SRI flou.

3.1. *Division de relations et approche Booléenne en RI*

Dans le modèle de données relationnel, un univers est modélisé par un ensemble de relations manipulées par les opérateurs de l'algèbre relationnelle. Parmi ces opérateurs, la division de la relation $C(A, X)$ par $Q(A)$, notée $C[A \div A]Q$, où A est l'ensemble des attributs communs à C et Q, détermine les valeurs sur X reliées dans

C à toutes les valeurs de A présentes dans Q. Cette opération peut être définie de plusieurs façons équivalentes :

$$x \in C[A \div A]Q \quad \Leftrightarrow \quad \forall a \in Q,\ (x,a) \in C\ , \qquad [1]$$

$$x \in C[A \div A]Q \quad \Leftrightarrow \quad Q \subseteq \Omega^{-1}(x) \quad \text{où} \quad \Omega^{-1}(x) = \{a|(x,a) \in C\}\ . \qquad [2]$$

Dans le modèle Booléen de RI, un document d est un ensemble de termes, et une requête peut être représentée par un ensemble P de termes désirés (ou positifs) et éventuellement un ensemble N de termes à exclure (ou négatifs). Un document d est pertinent s'il contient tous les termes positifs ($P \subseteq d$), et aucun des termes négatifs ($d \cap N = \emptyset$). Les opération ensemblistes jouent donc un rôle central dans de tels SRI.

Supposons que chaque document d est représenté par un ensemble de termes $d = \{t_1, \ldots, t_m\}$, où $t_i \in T$, l'ensemble des termes d'indexation, et une requête q contient seulement des termes positifs $P = \{t'_1, \ldots, t'_n\}$ où $t'_i \in T$. L'ensemble des documents de la collection peut être représenté par une relation normalisée (C_N) dans laquelle un document de m termes est représenté par m n-uplets $\langle d, t_1\rangle, \ldots, \langle d, t_m\rangle$, et la requête par une relation unaire P. La réponse à la requête est alors le résultat de la division de C_N par P.

L'approche Booléenne, à l'origine des SRI, est donc clairement liée aux mécanismes d'interrogation des BD, et en particulier la division de relations. Toutefois, elle a rapidement montré ses limites et n'est plus utilisée en RI depuis longtemps. Parmi les raisons, cette approche ne permet pas de représenter et d'utiliser l'importance relative des termes d'indexation des documents ou des requêtes.

3.2. *Extension graduelle de l'approche booléenne, et division de relations floues*

La plupart des extensions du modèle booléen ont pris en compte la notion d'importance relative des termes par des mécanismes de pondération.

Ce mécanisme est naturel en logique floue. Il consiste à représenter un document par un sous-ensemble flou des termes d'indexation T (Buell, 1982). Chaque terme t appartient à un document d de la collection C à un certain degré $\mu_C(d,t)$, qui représente son degré de représentativité (Waller *et al.,* 1979, Buell *et al.,* 1981). En théorie des sous-ensembles flous, la fonction μ est la fonction d'appartenance d'un élément à un ensemble, et la valeur $\mu_E(x)$ (dans l'intervalle unité) représente le degré d'appartenance de l'élément x à l'ensemble E. De même, une requête q peut aussi être un sous-ensemble flou de T, ou une requête plus complexe, structurée avec des opérateurs logiques flous (ET, OU, NON) (Bookstein, 1980). La pondération des termes des requêtes $\mu_q(t)$ a posé le problème de l'interprétation des poids attribués. La sémantique de ces poids $\mu_q(t)$ est discutée plus bas.

Ensuite, on retrouve les deux étapes classiques des SRI. Tout d'abord, l'utilisation d'une fonction d'appariement qui calcule des scores individuels $S_q(d,t)$ pour chaque terme t d'une requête q et chaque document d. Ensuite, l'utilisation d'une fonction d'agrégation des scores individuels $S_q(d,t)$, $t \in q$, pour obtenir un score global $S_q(d)$ pour chaque document, qui évalue le degré de satisfaction du document pour la requête. Ce degré permet un classement des documents jugés pertinents pour la requête. Dans un SRI flou, on peut utiliser des fonctions d'appariement et d'agrégation floues, donc à valeur dans l'intervalle unité.

Cette extension graduelle peut sembler ad hoc. Toutefois, comme précédemment, on peut obtenir la réponse à une requête q en effectuant la division de 2 relations, mais des relations floues, c'est-à-dire dont les n-uplets sont pondérés : celle des documents $C(d,t)$ et celle de la requête $Q(t)$. Pour cela, on généralise l'expression (2) au cas des relations floues, en remplaçant l'inclusion classique par un opérateur d'inclusion graduel g :

$$C[T \div T]Q(d) = g(Q \subseteq \Omega^{-1}(d)) \ , \qquad [3]$$

où $\Omega^{-1}(d)$ est un ensemble flou de termes défini par : $\Omega^{-1}(d) = \{\mu/t | \mu/(d,t) \in C\}$. Dans la notation μ/t, μ est le degré d'appartenance de l'élément t. Cette notation sert souvent à décrire en extension les sous-ensembles flous définis sur des domaines discrets.

La sémantique de la division ainsi obtenue dépend à la fois de l'opérateur d'inclusion et de la sémantique des poids associés aux n-uplets dans les relations C et Q (Bosc *et al.,* 1997). Une façon de modéliser l'inclusion graduelle $g(Q \subseteq \Omega^{-1}(d))$ consiste à utiliser une implication floue (notée $\rightarrow$ dans la suite), ce qui conduit à la formule :

$$g(Q \subseteq \Omega^{-1}(d)) = \min_{t \in Q} \left(\mu_Q(t) \rightarrow \mu_C(d,t)\right) \ . \qquad [4]$$

Dans cette formule, on retrouve la fonction d'appariement (l'implication), et la fonction d'agrégation (le $\min$).

3.3. *Inclusions graduelles et implications floues*

Dans la formule (4), l'opérateur d'appariement est une implication. Selon la nature de l'implication utilisée (R- ou S-implication), on obtient une interprétation différente des degrés d'inclusion, pour une sémantique différente des poids des termes dans les requêtes. On peut trouver dans (Bosc *et al.,* 2008b) un exemple de calculs réalisés avec diverses implications, qui illustre bien ces différences.

R-implication et notion de seuil. Une première approche consiste à interpréter le degré $\mu_Q(t)$ d'un terme t dans une requête Q comme un seuil à atteindre. On est alors totalement satisfait d'un document dès que ce seuil $\mu_Q(t)$ est atteint pour chaque terme t de la requête Q. Lorsque le seuil n'est pas atteint, le document reçoit une pénalité.

Ce comportement est obtenu avec une implication résiduée (ou R-implication), notée $\rightarrow_R$ et définie par (Fodor *et al.,* 1999) :

$$p \rightarrow_R q \ = \ \sup \{u \in [0,1] | \top(p,u) \leq q\} \ , \qquad [5]$$

où $\top$ est une norme triangulaire (i.e. une conjonction floue). Toute R-implication peut aussi s'écrire sous la forme :

$$p \rightarrow_R q = 1 \text{ si } p \leq q, \ f(p,q) \text{ sinon}, \qquad [6]$$

où $f(p,q)$ exprime une satisfaction partielle (une valeur < 1) lorsque l'antécédent p n'est pas atteint par la conclusion q. L'élément minimal de cette classe d'opérateur est connue sous le nom d'implication de Gödel :

$$p \rightarrow_{Gd} q = 1 \text{ si } p \leq q, \ q \text{ sinon},$$

obtenue en choisissant la plus grande T-norme $\top(a,b) = min(a,b)$ dans la formule (5). Parmi les R-implications très utilisées, on peut noter les implication de Goguen et de Lukasiewicz, obtenues respectivement avec le produit $\top(a,b) = a \cdot b$ ou la T-norme de Lukasiewicz $\max(a+b-1,0)$) :

$$p \rightarrow_{Gg} q = 1 \text{ si } p \leq q, \ q/p \text{ sinon},$$
$$p \rightarrow_{Lu} q = 1 \text{ si } p \leq q, \ 1-p+q \text{ sinon}.$$

Avec des R-implications, on voit clairement sur la formule (6) que $\mu_Q(t)$ doit être interprété comme un seuil, puisqu'on obtient une satisfaction maximale (le degré 1) dès que $\mu_C(d,t)$ atteint $\mu_Q(t)$.

S-implication et notion d'importance. On peut aussi voir $\mu_Q(t)$ comme un degré d'importance du terme t dans la requête, par rapport à l'information recherchée. Le cadre logique des implications conduit alors à imposer un degré de satisfaction garantie pour un document lorsque l'importance du terme recherché t est inférieure à 1. En effet, lorsque $\mu_Q(t) < 1$, le terme n'est pas totalement requis et peut donc être absent dans une certaine mesure. La satisfaction totale nécessite que $\mu_C(d,t) = 1$ pour chaque valeur t de Q quelle que soit son importance. Un document n'est pas du tout satisfaisant ($\mu_{C[T \div T]Q}(d) = 0$) seulement lorsque, pour au moins un terme de la requête, $\mu_Q(t) = 1$ (le terme est d'importance maximale) et $\mu_C(d,t) = 0$ (le terme n'est pas du tout représentatif du document). Ce comportement peut être modélisé en utilisant une S-implication (Fodor *et al.,* 1999) notée $\rightarrow_S$:

$$p \rightarrow_S q \ = \ \perp(1-p,q) = 1 - \top(p,1-q) \ , \qquad [7]$$

où $\perp$ est une conorme triangulaire (ou T-conorme, ou S-norme).

Comme dans le cas des R-implications, il existe une infinité de S-implications, en fonction de la norme génératrice choisie. La plus utilisée est l'implication de Kleene-Dienes définie par :

C	t_1	t_2	t_3	t_4
d_1	1	0.9	1	0.2
d_2	0.7	0.6	0.3	0.8

	t_1	t_2	t_3	t_4
q	1	0.4	0	0.6
q'	0.6	0.6	0.3	0.5

	sémantique du poids des requêtes	implication	d_1	d_2
q	importance	Kleene-Dienes	0.4	0.6
		Reichenbach	0.52	0.76
q'	seuil de satisfaction	Gödel	0.2	1
		Goguen	0.4	1
		Lukasiewicz	0.7	1

Tableau 1. *En haut à gauche : relation floue C représentant la collection — En haut à droite : chaque ligne est une relation floue Q représentant une requête — En bas : résultats de la division selon l'implication choisie.*

$$p \rightarrow_{KD} q \;=\; \max(1-p, q) \;\; .$$

C'est l'élément minimal dans la formule (7), obtenu avec la plus petite conorme $\perp$, i.e., le maximum. Lorsqu'on choisit pour $\perp$ la somme probabiliste, on obtient l'implication de Reichenbach :

$$p \rightarrow_{Rb} q \;=\; 1 - p + p \cdot q \;\; .$$

L'implication de Lukasiewicz, vue plus haut, est aussi une S-implication générée par $\perp(a, b) = \min(a + b, 1)$.

Enfin, on peut remarquer que la formule (4) correspond à la division de relations classiques lorsque les termes ne sont pas pondérés, puisque les implications floues généralisent l'implication classique (en particulier elles conservent $1 \rightarrow 0 = 0$ et $1 \rightarrow 1 = 1$).

Effet d'absorption. L'approche logique proposée est de type conjonctif, et produit un *effet d'absorption*. En effet, l'opérateur de division, et en particulier l'agrégation par le $\min$ dans (4), donne comme résultat global le plus petit des degrés d'implication entre un terme de la requête et le document. Pour éviter cet effet néfaste, la formule (4) est relâchée par l'utilisation d'une autre T-norme que le $\min$ dans notre modèle de RI.

Exemple. La table 1 donne les relations floues représentant une collection de deux documents d_1 et d_2, deux requêtes q et q', et les résultats, en fonction de la sémantique choisie. Un effet de seuil apparaît clairement avec q' et d_2. Cet exemple est tiré de (Bosc *et al.,* 2008b).

4. Implémentation et caractéristiques du SRI

Notre SRI implémente l'approche floue décrite dans la section 3. Ainsi, le score d'un document d, pour une requête q est calculé de la façon suivante :

$$S(d,q) = \top_{t \in q}(w_q(t) \rightarrow w_d(t)) \ , \qquad [8]$$

où t est un terme de la requête, $w_q(t)$ son poids dans la requête, $w_d(t)$ (qu'on peut aussi noter $w_C(d,t)$) son poids dans le document, $\rightarrow$ l'implication floue correspondant à l'inclusion graduelle choisie, et $\top$ la T-norme d'agrégation. On voit dans la formule (8) que de nombreux paramètres peuvent être ajustés : le poids des termes dans le document ou la requête, et les opérateurs d'implication et d'agrégation.

Opérateur d'agrégation. Lorsque $\top$ est l'opérateur $\min$, le score $S(d,q)$ obtenu par le document d est le degré d'appartenance de d au quotient de la division floue de la collection par la requête q. Comme $\min$ est la plus grande T-norme, elle fournit le plus grand score $S(d,q)$. Ce score correspond au degré d'inclusion dans d du terme $t \in q$ le moins inclus dans d. Cette vision correspond à une approche BD classique, dans laquelle chaque terme de la requête doit se retrouver dans les n-uplets du résultat. C'est le degré d'inclusion du *terme le plus faible* qui donne la mesure de pertinence du document. Cette approche ne donne pas de bons résultats en RI.

En RI, un document pertinent ne contient pas toujours tous les termes de la requête. Dans la plupart des modèles vectoriels, lorsqu'un terme recherché est absent d'un document, il ne modifie pas le score du document ; il est neutre, ce qui correspond bien à l'agrégation des scores individuels des termes par une somme. Par contre, un terme très représentatif (fréquent dans le document et rare dans la collection) augmente beaucoup le score. Du point de vue de la RI, les *meilleurs termes* sont donc plus importants que les *plus mauvais*. De plus, pour pouvoir classer les documents, leur score final doit tenir compte de tous les scores individuels des termes, alors que l'agrégation par le $\min$ ne conserve que le poids d'un seul terme (le plus mauvais). C'est pour cette raison que l'équation (4) a été relâchée en (8) qui reste une mesure d'inclusion, et que de nombreuses T-normes ont été testées, comme : min, Drastic, Einstein, Lukasiewicz, Product, ou des T-normes paramétrées comme : Dubois et Prade, Hamacher, Yager...

Opérateur d'inclusion graduelle. Comme on l'a vu section 3, deux classes d'opérateurs ont été testées : les R- et les S-implications. Les plus représentatives (et utilisées) de chaque classe ont été choisies pour la première série de tests. Parmi les R-implications : Gödel, Goguen, Lukasiewicz. Parmi les S-implications : Kleene-Dienes, Lukasiewicz, Reichenbach, Willmott. Le lecteur peut trouver leur définition dans de nombreux articles, par exemple (Fodor *et al.,* 1999).

Poids des termes dans les documents. Dans le contexte de la division de relations floues, les poids doivent avoir une sémantique claire (importance, seuil, préférence...). Le schéma de pondération classique d'OKAPI-BM25 a été choisi car il véhicule la notion d'importance relative des termes. Toutefois, comme ils sont ma-

nipulés par des opérateurs flous, les poids doivent être ramenées à l'intervalle unité $[0, 1]$. Les poids d'OKAPI-BM25 ($w_{BM25}(t, d)$) ont donc été normalisés et bornés.

Poids des termes dans les requêtes. Comme pour les poids des termes dans les documents ($w_d(t)$), les poids des termes dans les requêtes ($w_q(t)$) doivent avoir une sémantique claire, de façon à pouvoir être comparés à ceux des documents. C'est d'une importance particulière avec des R-implications, pour lesquelles $w_q(t)$ est un degré de satisfaction à atteindre par $w_d(t)$. Pour l'instant, la seule façon d'obtenir de tels poids dans les requêtes est de les fixer manuellement. Cependant, ils auraient été attribués de façon subjective, et n'auraient pas permis une comparaison équitable avec d'autres SRI. C'est pourquoi un mécanisme de pondération classique et automatique a été choisi, aux dépends de la sémantique, pour ces premiers tests dont l'objectif est la validation de notre modèle de RI. Les poids choisis dépendent de la fréquence des termes dans les requêtes, et sont normalisés et bornés.

5. Résultats expérimentaux

Les expérimentation ont été menées en faisant varier les différents paramètres du SRI. Les résultats ont été comparés à ceux du modèle OKAPI, avec des paramètres (k_1 et b) identiques à ceux utilisés dans Lemur. Cette section montre à la fois de bons résultats, qui valident le modèle proposé, et de mauvais résultats, accompagnés d'explications sur les causes probables. Par manque de place, seuls quelques résultats jugés caractéristiques sont détaillés.

5.1. *Collections de documents*

Le SRI proposé a été testé sur 3 collections. La première, nommée ELDA, est une petite collection en français, contenant 3499 documents (des questions/réponses de la commission européenne), et un ensemble de 19 requêtes. La deuxième est le collection INIST, qui contient 163.308 documents (des résumés d'articles de diverses disciplines scientifiques) et 30 requêtes. Ces deux collections proviennent de la campagne d'évaluation de RI Amaryllis. La troisième, TIPSTER, est une collection de TREC-3 contenant 173.252 articles du Wall Street Journal et 50 requêtes.

Pour toutes les collections, les documents et requêtes ont été lemmatisées. Les requêtes sont composées de plusieurs champs : titre, corps, description et concepts associés. Dans les expérimentations, seuls le titre et les concepts associés ont été utilisés (sauf pour TIPSTER qui ne contient pas de concepts associés).

5.2. *Propriétés qui conduisent à de mauvais résultats*

Les effets d'absorption et de seuil sont les principaux responsables des mauvais résultats. Ils peuvent intervenir à différents niveaux, impliquant l'opérateur d'agrégation, l'implication ou les poids.

Élément absorbant des T-normes. Dans notre modèle, les scores individuels des termes sont agrégés par une conjonction. Les conjonctions ont un élément absorbant (le zéro) qui pose le problème suivant : dès qu'un terme reçoit le score 0, le document aussi, quels que soient les poids des autres termes.

Le score d'un terme est donné par : $w_q(t) \rightarrow w_d(t)$. Avec la plupart des R-implications, ce score vaut 0 dès que $w_d(t) = 0$, i.e. dès que le terme est absent du document. Avec une S-implication, le score vaut $1 - w_q(t)$ dans ce cas, et il n'est nul que lorsque $w_q(t) = 1$, i.e. lorsque le terme est absent du document et a une importance maximale.

Pour contourner ce problème, la stratégie adoptée est la même que dans les modèles de langue, qui utilisent des techniques de lissage : un mot absent d'un document reçoit un score faible prédéfini, strictement positif. Cela signifie qu'un terme, même absent d'un document *peut* être représentatif de ce document (par exemple il peut être synonyme d'un terme du document).

Effet de seuil des R-implications. Avec une R-implication, $w_q(t)$ est le degré minimal attendu pour $\mu_d(t)$ dans les documents totalement pertinents. Dès que $w_d(t) \geq w_q(t)$, ce degré est donc atteint et le score du terme t, obtenu par $w_q(t) \rightarrow w_d(t)$, vaut 1. La mauvaise conséquence est que deux documents ayant des poids différents $w_{d_1}(t) \neq w_{d_2}(t)$, mais tous deux au dessus du seuil $w_q(t)$, obtiennent le même score 1 et ne peuvent donc être classés. Ici encore, ce comportement reflète une approche BD classique, dans laquelle le système doit seulement retrouver tous les n-uplets pertinents, et n'a pas à les classer. En RI, les documents doivent être présentés par ordre de pertinence décroissante, comme cela se fait dans les approches de requêtage flexible des BD. C'est la raison pour laquelle les R-implications conduisent à de mauvais résultats dans le cas général. On pourrait contourner ce problème en choisissant dans les requêtes des poids $w_q(t)$ (les seuils requis) toujours supérieurs aux poids de ces termes dans les documents. Dans ce cas, le seuil n'est jamais atteint, les scores individuels sont toujours strictement inférieurs à 1, et les documents peuvent être classés. Les deux types d'implications donnent alors des résultats proches, aux dépends de la sémantique.

Effet d'absorption des opérateurs de type min. Certains opérateurs d'agrégation ont un effet d'absorption, comme $\min$, $\max \ldots$ Avec cette classe d'opérateurs, seulement un des termes (ou un petit nombre) est pris en compte pour le calcul du score global du document. La mauvaise conséquence est, encore une fois, un mauvais classement des documents, qui conduit à de mauvais résultats. Ce type d'opérateur est donc à éviter.

5.3. *Résultats*

Parmi les nombreuses combinaisons possibles d'opérateurs et poids que nous avons testées, cette section présente seulement quelques jeux de paramètres, représentatifs aussi bien des bons que des mauvais résultats. La table 2 donne les résultats pour l'implication de Reichenbach associée au produit et à la T-norme d'Einstein ($a \top_E b = (a.b)/(2 - a + b - a.b)$), et pour l'implication de Lukasiewicz, associée au Produit et à la T-norme de Lukasiewicz ($a \top_L b = \max(0, a + b - 1)$).

Les résultats sont évalués en terme de précision moyenne non interpolée (MAP), précision moyenne interpolée (IAP), R-précision (Rprec), et précision pour les k premiers documents (Pk). La différence relative (Rdiff) avec OKAPI est aussi indiquée. Les valeurs en gras sont considérées comme statistiquement significatives par un T-test.

INIST	OKAPI	SRI à base d'inclusion graduelle							
implic.		Reichenbach				Lukasiewicz			
t-norme		Einstein	Rdiff	Product	Rdiff	Lukasiewicz	Rdiff	Product	Rdiff
MAP %	21.75	**23.22**	**(+6.79%)**	**23.13**	**(+6.37%)**	**0.03**	**(-99.85%)**	**23.03**	**(+5.90%)**
IAP %	24.13	**25.60**	**(+6.10%)**	**25.50**	**(+5.70%)**	**0.20**	**(-99.17%)**	25.38	(+5.17%)
Rprec %	25.85	28.20	(+9.09%)	27.94	(+8.08%)	**0.03**	**(-99.90%)**	**28.09**	**(+8.69%)**
P5 %	50.00	45.33	(-9.33%)	49.33	(-1.33%)	**0.00**	**(-100.00%)**	48.00	(-4.00%)
P10 %	42.67	42.67	(0.00%)	42.00	(-1.56%)	**0.00**	**(-100.00%)**	43.67	(+2.34%)
P100 %	17.03	**18.27**	**(+7.24%)**	**18.20**	**(+6.85%)**	**0.03**	**(-99.80%)**	**18.23**	**(+7.05%)**
P500 %	5.39	**5.64**	**(+4.70%)**	**5.61**	**(+4.08%)**	**0.03**	**(-99.38%)**	**5.63**	**(+4.58%)**

ELDA	OKAPI	SRI à base d'inclusion graduelle							
implic.		Reichenbach				Lukasiewicz			
t-norme		Einstein	Rdiff	Product	Rdiff	Lukasiewicz	Rdiff	Product	Rdiff
MAP %	57.14	56.86	(-0.49%)	56.91	(-0.42%)	**1.11**	**(-98.06%)**	56.29	(-1.50%)
IAP %	58.09	57.89	(-0.36%)	57.88	(-0.37%)	**1.98**	**(-96.59%)**	57.38	(-1.23%)
Rprec %	55.33	53.82	(-2.73%)	54.64	(-1.26%)	**0.67**	**(-98.78%)**	**53.03**	**(-4.16%)**
P5 %	77.24	76.55	(-0.89%)	74.48	(-3.57%)	**1.38**	**(-98.21%)**	75.17	(-2.68%)
P10 %	68.28	68.62	(+0.51%)	68.97	(+1.01%)	**0.69**	**(-98.99%)**	67.93	(-0.51%)
P100 %	27.00	26.86	(-0.51%)	26.83	(-0.64%)	**1.00**	**(-96.30%)**	26.83	(-0.64%)
P500 %	6.67	6.66	(-0.10%)	6.67	(+0.00%)	**0.87**	**(-86.97%)**	6.66	(-0.10%)

TIPSTER	OKAPI	SRI à base d'inclusion graduelle							
implic.		Reichenbach				Lukasiewicz			
t-norme		Einstein	Rdiff	Product	Rdiff	Lukasiewicz	Rdiff	Product	Rdiff
MAP %	18.14	18.61	(2.61%)	18.66	(2.87%)	**2.53**	**(-86.08%)**	18.66	(2.87%)
IAP %	20.09	**20.83**	**(3.69%)**	20.90	(4.06%)	**2.70**	**(-86.55%)**	20.90	(4.02%)
Rprec %	22.42	22.85	(1.91%)	23.31	(4.00%)	**3.47**	**(-84.54%)**	23.32	(4.02%)
P5 %	31.60	32.40	(2.53%)	32.80	(3.80%)	**5.60**	**(-82.28%)**	32.80	(3.80%)
P10 %	30.40	32.00	(5.26%)	31.80	(4.61%)	**6.00**	**(-80.26%)**	32.00	(5.26%)
P100 %	17.14	17.14	(0.00%)	17.08	(-0.35%)	**3.64**	**(-78.76%)**	17.06	(-0.47%)
P500 %	7.33	7.37	(0.49%)	7.34	(0.11%)	**0.85**	**(-88.43%)**	7.35	(0.27%)

Tableau 2. *Résultats pour les collections INIST, ELDA et TIPSTER*

Lorsque les différents paramètres sont choisis pour éviter les mauvaises propriétés répertoriées plus haut, et grâce à la pondération BM-25, les résultats de notre SRI sont positifs, et comparable à ceux d'OKAPI (parfois même un peu meilleurs), ce qui se vérifie par l'absence de différences jugées statistiquement significatives.

Opérateurs. Pour l'ensemble des collections, les meilleurs résultats sont obtenus avec l'implication de Reichenbach associée à la T-norme produit ou Einstein. Dans quelques cas, l'implication de Lukasiewicz, et la pseudo-implication de Larsen (le produit), donnent aussi de bons résultats. Les implications paramétrées donnent aussi de bons résultats, mais principalement lorsque leur comportement se rapproche du produit.

Il est intéressant de constater que l'utilisation de T-conormes (des disjonctions) pour l'agrégation des scores, produit souvent des résultats similaires (bien qu'inférieurs de quelques pourcents) à ceux obtenus avec les T-normes associées. Surprenant au premier abord (car OU signifie « un au moins », alors que ET veut dire « tous »), ce résultat s'explique par le fait que ce n'est pas la valeur finale du score qui importe, mais la façon de prendre en compte les scores individuels des termes pour produire le score final du document, qui conditionne le classement. Or, T-norme et T-conorme associées ont souvent un comportement similaire de ce point de vue.

6. Expressivité du modèle à base d'inclusion graduelle

6.1. *Expressivité de la requête*

Un schéma de pondération classique a été utilisé pour valider notre modèle. Toutefois, la fréquence des termes dans les requêtes ne représente pas vraiment l'importance des termes par rapport au besoin d'information de l'utilisateur, en particulier lorsque les requêtes sont des phrases plutôt que seulement des mots-clés. S'il n'est pas possible en général de demander à l'utilisateur de pondérer les termes de ses requêtes par des nombres réels, l'approche graduelle proposée permet cependant de simplifier la pondération manuelle, par exemple en demandant à l'utilisateur de présenter ses termes par ordre d'importance, d'utiliser une échelle d'importance ordinale, ou de les saisir par catégorie d'importance (par exemple en remplissant 3 à 5 cases d'importance décroissante dans un formulaire).

Ce modèle flou permet également de prendre en compte très naturellement l'utilisation d'opérateurs flous dans les requêtes. De nombreux travaux se sont intéressés à ces opérateurs, que ce soient des ET/OU flous (Herrera-Viedma *et al.,* 2007), ou des opérateurs moins standard (Mercier *et al.,* 2006). Dans notre SRI, cela peut servir à prendre en compte les concepts associés aux requêtes, en particuliers lorsqu'ils sont composés de plusieurs mots. Par exemple, *« pollution de l'air », « effet de serre »* peuvent donner de meilleurs résultats lorsqu'ils sont exprimés par (*pollution* AND *air*) OR (*effet* AND *serre*), plutôt que par 4 mots indépendants. La richesse des opérateurs de logique floue permet aussi de moduler la sémantique des conjonctions et disjonctions. Par exemple, min/max véhiculent la notion d'indépendance. Dans l'expression :

$$\max(\min(\mu_d(\textit{pollution}), \mu_d(\textit{air})), \min(\mu_d(\textit{effet}), \mu_d(\textit{serre}))) ,$$

la disjonction max signifie que le « meilleur » des concepts associés suffit : il donne le score global. Le min signifie que les 2 termes doivent être présents dans le document, car le score du concept est égal au score du « plus mauvais » terme. D'autres opérateurs, comme le produit ou la somme probabiliste, véhiculent la notion de renforcement. Par exemple, avec la somme probabiliste à la place du max, dans l'expression précédente, plus il y a de concepts associés dans le document, plus le score est élevé.

Le modèle flou permet aussi de modéliser et d'exploiter des parties de requêtes négatives, c'est-à-dire des termes que l'on ne souhaite pas voir apparaître. Cela peut se faire en utilisant l'antidivision (Bosc *et al.,* 2008a), opération duale de la division. L'antidivision de $C(A, X)$ par $Q(A)$, renvoie les éléments x de C tels que $\forall a \in Q$, $(a, x) \notin C$, c'est-à-dire, dans notre modèle, les documents qui ne contiennent pas les termes négatifs.

Si la plupart de ces extensions ne sont pas vraiment originales, dans notre modèle elles peuvent reposer sur une approche bien fondée. Les opérateurs, les poids, et les résultats peuvent alors tirer profit d'une sémantique claire. Cela peut aider à obtenir de meilleurs résultats.

6.2. *Utilisation d'informations lexicales*

L'utilisation d'informations lexicales externes (e.g. des synonymes, des liens morphologiques...) peut aussi se faire naturellement. En RI, ces informations sont souvent utilisées pour étendre des requêtes (Voorhees, 1998, inter alia). Se pose alors le problème de leur pondération par rapport aux termes originaux de la requête et éventuellement de savoir quand arrêter l'enrichissement (doit-on prendre les termes liés aux termes liés...). Ces problèmes sont souvent résolus de manière ad-hoc (Moreau *et al.,* 2007, Voorhees, 1998, inter alia). Dans notre approche floue, ces informations peuvent servir à dilater le dividende de notre division, c'est-à-dire à enrichir le document. Plus formellement, l'inclusion basée sur ce principe peut se définir comme précédemment par :

$$g(Q \subseteq \Omega^{-1}(d)) = \min_{t \in Q} (\mu_Q(t) \rightarrow \mu_{dil(C)}(d, t)) \quad [9]$$

mais avec le degré d'appartenance du terme dans le document défini par :

$$\mu_{dil(C)}(d, t) = \top_{t' \in U} \perp(\mu_C(d, t'), \mu_{\mathrm{rsb}}(t, t')) \quad [10]$$

Cette dernière équation rend compte de l'importance d'un terme dans un document comme dépendant de $\mu_{\mathrm{rsb}}(t, t')$, la proximité du terme et d'un de ses mots liés (synonyme ou autre), et de $\mu_C(d, t')$, le poids de ce mot lié dans le document. Ces deux éléments sont combinés par une T-norme, et la T-conorme permet en quelque sorte de choisir le « meilleur » des mots liés au terme initial, ou de renforcer un terme dont beaucoup de mots liés apparaissent. Un exemple de l'utilisation de la dilatation est détaillé dans (Bosc *et al.,* 2008b).

L'avantage de cette formulation est de combiner simplement avec une T-norme la force du lien entre le mot requête et les mots liés présents dans le document et leurs poids. Les différentes T-normes et T-conormes possibles donnent différentes sémantiques à cette combinaison.

Pour illustrer l'intérêt de ce mécanisme, nous avons mis en place une expérience simple dans laquelle nous exploitons une base de liens morphologiques. Dans cette base construite automatiquement (voir (Moreau *et al.,* 2007) pour les détails de sa construction), un mot comme *pollution* est lié à *pollutions*, *polluant*, *anti-pollution*, *etc.* avec un score représentant la fréquence du lien morphologique. Ce score, une fois ramené dans l'intervalle $[0, 1]$, représente $\mu_{\text{rsb}}(t, t')$. La dilatation du dividende est mise en œuvre avec le produit comme T-norme, et la somme bornée comme T-conorme.

Pour ces expériences, nous utilisons de nouveau la collection INIST, mais sans lemmatisation pour faire ressortir au mieux l'intérêt des liens morphologiques. On se place donc dans le cas où on ne connait pas de techniques pré-existantes de lemmatisation (voir (Moreau *et al.,* 2007) pour une étude de l'influence de la lemmatisation dans ce cadre). Le tableau 3 présente les résultats obtenus avec les mêmes conventions que précédemment. Pour comparaison, nous indiquons les résultats obtenus par OKAPI dans lequel les requêtes sont étendues avec les mêmes mots liés morphologiquement aux termes des requêtes.

	OKAPI avec requêtes étendues	SRI flou avec dilatation du dividende	
MAP	15.36	**17.81**	**(+16.00 %)**
IAP	17.67	**20.13**	**(+13.92 %)**
Rprec	20.13	**23.14**	**(+14.97 %)**
P5	40.67	**42.67**	**(+4.92 %)**
P10	35.67	36.33	(+1.87 %)
P100	14.63	**16.23**	**(+10.93 %)**
P500	5.14	5.19	(+0.91 %)

Tableau 3. *Inclusion d'informations morphologiques par dilatation du dividende*

Les résultats de cette expérience abondent largement dans le sens de l'intérêt de ce mécanisme de dilatation puisque l'on constate une amélioration significative des résultats comparés à l'utilisation brute des liens morphologiques en extension de requête dans OKAPI. Le poids donné à ces variantes a été obtenu naturellement, à l'aide de la formule donnée ci-dessus. Bien entendu, une étude des différentes T-normes et T-conormes utilisables pour la dilatation et leur influence sur les résultats reste cependant à mener.

7. Conclusion

Le modèle de RI à base d'inclusion graduelle présenté dans cet article semble prometteur. Lorsque les paramètres sont bien choisis (cf. sections 4 et 5.3), on a montré que ce modèle donne des résultats comparables à ceux du marché, tout en apportant un cadre théorique fort. Des propriétés que le modèle doit avoir pour produire de bons résultats ont aussi pu être identifiés lors de cette étude.

Le plus intéressant réside cependant dans les extensions assez naturelles que permet ce modèle. Nous avons montré qu'il était prometteur pour l'utilisation de ressources lexicales externes habituellement utilisées en extension de requêtes. Il doit permettre également de construire et d'exploiter des requêtes beaucoup plus expressives, tout en conservant le cadre théorique et une sémantique claire des pondérations utilisées. En particulier, des procédures simples et intuitive de pondération des requêtes, ou d'enrichissement peuvent être exploitées. Nous serons cependant confrontés au problème de l'évaluation de ces techniques à grande échelle, par manque de collections de RI adaptées.

D'autres travaux concernant ce modèle sont aussi à l'étude. Par exemple, l'utilisation des inclusions tolérantes quantitative et qualitative proposées pour ce modèle (Bosc *et al.,* 2008b), doivent être validées expérimentalement. Ces inclusions tolérantes pourraient permettre de lever certaines restrictions que nous avons relevées dans cet article, quant-à l'emploi de certaines famille de T-normes. Enfin, les différentes expérimentations présentées dans cet article font ressortir le fait que les mécanismes de RI habituels se concentrent principalement sur l'intersection entre requêtes et documents, alors que ceux opérant dans le domaine des BD s'intéressent plutôt à l'inclusion, par la mesure de la quantité de termes des requêtes hors du document. Une définition de l'inclusion entre ensembles, basée cette fois sur la cardinalité, pourrait rapprocher plus encore ces deux philosophies.

8. Bibliographie

Bookstein A., « Fuzzy requests : an approach to weighted Boolean searches », *Journal of the American Society for Information Science*, vol. 31, p. 240-247, 1980.

Bosc P., Claveau V., Pivert O., Ughetto L., « On the use of tolerant graded inclusions in information retrieval », *Proceedings of the European Conference on Information Retrieval, ECIR'09*, Toulouse, France, 2009. à paraître.

Bosc P., Dubois D., Pivert O., Prade H., « Flexible queries in relational databases – The example of the division operator », *Theoretical Computer Science*, vol. 171, p. 281-302, 1997.

Bosc P., Pivert O., « On a Parameterized Antidivision Operator for Database Flexible Querying », *Proceedings of the 19th International Conference on Database and Expert Systems Applications, DEXA'08*, Turin, Italy, p. 652-659, 2008a.

Bosc P., Pivert O., « On the use of tolerant graded inclusions in information retrieval », *Actes de la 5e Conférence en Recherche d'Information et Applications, CORIA'08*, Trégastel, France, p. 321-336, 2008b.

Boughanem M., Loiseau Y., Prade H., « Improving Document Ranking in Information Retrieval Using Ordered Weighted Aggregation and Leximin Refinement », *Proceedings of the European Society for Fuzzy Logic and Technology Conference, EUSFLAT'05*, Barcelona, Spain, p. 1269-1274, 2005.

Brini A., Boughanem M., Dubois D., « A Model for Information Retrieval Based on Possibilistic Networks », *Proceedings of the 12th String Processing and Information Retrieval International Conference, SPIRE'05*, Buenos Aires, Argentina, p. 271-282, 2005.

Buell D., « An analysis of some fuzzy subset applications to information retrieval systems », *Fuzzy Sets & Systems*, vol. 7, p. 35-42, 1982.

Buell D., Kraft D., « Threshold values and Boolean retrieval systems », *Information Processing & Management*, vol. 17, p. 127-136, 1981.

Fodor J., Yager R., *Fundamentals of Fuzzy Sets — The Handbook of Fuzzy Sets Series (D. Dubois and H. Prade eds.)*, Kluwer Academic Publishers, chapter Fuzzy Set-theoretic Operators and Quantifiers. Chap. 1.2, p. 125-193, 1999.

Herrera-Viedma E., « Modeling the retrieval process for an information retrieval system using an ordinal fuzzy linguistic approach », *Journal of the American Society for Information Science and Technology*, vol. 52, p. 460-475, 2001.

Herrera-Viedma E., López-Herrera A., Luque M., Porcel C., « A Fuzzy Linguistic IRS Model Based on a 2-Tuple Fuzzy Linguistic Approach », *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 15, n^o 2, p. 225-250, 2007.

Kraft D. H., Pasi G., Bordogna G., « Vagueness and uncertainty in information retrieval : how can fuzzy sets help ? », *Proceedings of IWRIDL'2006*, p. 1-10, 2006.

Lalmas M., « Logical Models in Information Retrieval : Introduction and overview », *Information Processing & Management*, vol. 34, n^o 1, p. 19-33, 1998.

Mercier A., Imafouo A., Beigbeder M., « Using a Fuzzy Proximity Matching Function », *ENSM-SE at CLEF 2005*, vol. 4022/2006 of *LNCS*, p. 187-193, 2006.

Moreau F., Claveau V., Sébillot P., « Automatic morphological query expansion using analogy-based machine learning », *Proceedings of the European Conference on Information Retrieval, ECIR'07*, Rome, Italie, avril, 2007.

Oussalah M., Khan S., Nefti S., « Personalized information retrieval system in the framework of fuzzy logic », *Expert Systems with Applications*, vol. 35, p. 423-433, 2008.

Salton G., Fox E., Wu H., « Extended Boolean Information Retrieval », *Communications of the ACM*, vol. 26, n^o 12, p. 1022-1036, 1983.

Voorhees E., *C. Fellbaum (ed.), WORDNET : An Electronic Lexical Database*, The MIT Press, chapter Using WORDNET for Text Retrieval, p. 285-303, 1998.

Waller W., Kraft D., « A mathematical model of a weighted Boolean retrieval system », *Information Processing & Management*, vol. 15, p. 235-245, 1979.