# Différentes interprétations d'un modèle de RI à base d'inclusion graduelle

Laurent Ughetto, Vincent Claveau, Rima Harastani

# Différentes interprétations d'un modèle de RI à base d'inclusion graduelle

**L. Ughetto**[1]— **V. Claveau**[2]— **R. Harastani**[3]

1 *IRISA - Université Rennes 2 - Campus de Beaulieu, F-35042 Rennes cedex, France*
2 *IRISA - CNRS - Campus de Beaulieu, F-35042 Rennes cedex, France*
3 *LINA - Université de Nantes, F-44322 Nantes cedex France*
laurent.ughetto@irisa.fr vincent.claveau@irisa.fr
rima.harastani@univ-nantes.fr

RÉSUMÉ. *Récemment, un modèle théorique de RI à base d'inclusion graduelle a été proposé (Bosc* et al., *2008b). Dans ce modèle, dérivé de la division de relations floues, l'inclusion graduelle d'une requête dans un document est modélisée par une implication floue. Dans des travaux précédents, nous avons montré que ce modèle pouvait être interprété comme un modèle vectoriel sous certaines conditions. Dans cet article, nous proposons d'explorer d'autres interprétations possibles offertes par la modélisation à base d'inclusion graduelle. Nous montrons notamment qu'il est possible d'interpréter notre système flou comme un système de RI à base de modèles de langues, et nous revenons sur les liens entre le modèle flou et les modèles logiques de RI. Plus généralement, nous essayons de clarifier les liens existants entre ces différents modèles, vus sous l'angle de notre SRI flou.*

ABSTRACT. *Recently, a theoretical fuzzy IR system, based on gradual inclusion measures, has been proposed (Bosc* et al., *2008b). In this model, derived from the division of fuzzy relations, the gradual inclusion of a query in a document is modeled by a fuzzy implication. In previous papers, we have shown that, under some assumptions, this model can be seen as a Vector Space Model. This paper also studies other itnerpretations of our fuzzy IR models based on gradual inclusions. It is shown that the fuzzy models can be interpreted as language models for IR. The links with logical models to IR are also recalled. More generaly, this paper discusses the links between these models, shown from the angle of our fuzzy models.*

MOTS-CLÉS : *modèles de RI, logique floue, modèles de langue*

KEYWORDS: *IR models, fuzzy logic, language models*

## 1. Introduction

Les communautés de la recherche d'information (RI) et des bases de données (BD) partagent le même but de fournir aux utilisateurs les informations qu'ils demandent. Cependant, il est bien conu que les approches classiques d'interrogation utilisées en BD ne sont pas utilisables en RI : elles n'ont pas la flexibilité requise pour réaliser un appariement approximatif entre les termes des documents et des requêtes ; de plus elles offrent rarement un moyen de classer les résultats. Néanmoins, de récentes études sur l'interrogation flexible des BD ont défini de nouveaux mécanismes d'interrogation plus adaptés à la RI, et suite aux travaux de (Bosc *et al.,* 2008b, Bosc *et al.,* 2008a) sur la division de relations floues, des modèles de RI flous, à base d'inclusion graduelle, ont été définis et validés expérimentalement (Bosc *et al.,* 2009a, Bosc *et al.,* 2009b).

Les inclusions graduelles considérées, extensions floues de l'inclusion ensembliste classiques, sont fondées sur l'implication, ou sur la cardinalité. Par construction, ou d'après leur formule de score, de nombreuses similitudes ont été remarquées entre nos modèles flous et certains modèles classiques de RI. Par exemple, le modèle flou à base d'implication a été obtenu dans (Bosc *et al.,* 2008b) par le biais des opérateurs de division en BD, mais il peut aussi être obtenu directement comme une extension du modèle booléen de RI (Pasi, 1999). D'autres liens ont déjà été mentionnés, et seront rappelés et détaillés, comme celui des modèles flous avec les modèles vectoriels (Bosc *et al.,* 2008b) ou avec les modèles logiques de RI (Ughetto *et al.,* 2010). Cet article montre aussi que les modèles flous peuvent être interprétés comme des modèles de langue. Inversement, ces liens entre modèles montrent que les formules de score de nombreux modèles classiques de RI peuvent être réinterprétés comme des mesures d'inclusion graduelle.

Après avoir rappelé les principes des modèles flous dans la section 2, l'objectif de la section 3 est de détailler ces similitudes, et de montrer que les modèles flous peuvent être interprétés selon divers modèles.

## 2. Modèle de RI flou à base d'inclusion graduelle

Si on considère que les documents et les requêtes sont des ensembles de termes, l'inclusion peut être vue comme un modèle de RI simple : un document est pertinent si et seulement si il contient tous les termes de la requête.

L'un des tout premiers modèles de RI, le modèle Booléen, est fondé sur ce modèle d'inclusion, et mis en œuvre par la logique Booléenne (pour permettre l'écriture de requêtes plus générales). Dans ce modèle, un document est un ensemble de termes. Une requête est une formule logique, composée de termes liés par les opérateurs ET, OU et NON, et qui peut donc être écrite sous forme normale disjonctive. Dans ces conditions, un document est pertinent si et seulement si, pour au moins une des clauses conjonctives de la requête, les termes non-négativés doivent être présents dans le document, les autres doivent en être absents.

Cependant, la requête est souvent plus simplement considérée comme un « sac de termes », assimilables à des multi-ensembles dont tous les mots sont requis, ou encore à une formule dans laquelle les termes sont liées uniquement par l'opérateur ET. La requête ne contient alors ni mots négatifs (pas de NON) ni alternative (pas de OU). Dans ce cas, le modèle Booléen correspond au modèle d'inclusion simple.

Dans cet article, dans un souci de simplifier les écritures, nous considérons seulement ce cas, des requêtes dites « sac de termes ». Cette restriction peut se faire sans perte de généralité sur les considérations faites sur le modèle flou discuté, car les opérateurs OU et NON s'ajoutent à ce modèle naturellement.

Cette section montre comment le modèle à base d'inclusion, étendu à une mesure d'inclusion graduelle permet d'obtenir un modèle de RI efficace.

## 2.1. *Modèles de RI flous*

La plupart des extensions du modèle booléen cherchent à corriger certains défauts bien connus, à savoir :

– l'absence de pondération des termes du document et de la requête, qui ne permet pas de prendre en compte l'importance des termes dans les documents, ou la préférence de l'utilisateur dans la requête ;

– l'absence de pondération du critère de pertinence : un document est pertinent si la requête est vraie, et non pertinent sinon ;

– le fait qu'un document est jugé non pertinent dès qu'il lui manque un terme de la requête (ou dès qu'un terme négativé est présent), même s'il contient tous les autres.

Pour ce faire, ces extensions s'attachent à deux aspects du modèle : la notion d'importance relative des termes, souvent mise en œuvre par des mécanismes de pondération, et le calcul de l'inclusion. C'est aussi ces deux point qui sont attaqués dans les extensions par logique floue.

### 2.1.1. *Pondération*

Tout d'abord, le mécanisme de pondération 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 $t_j \in T$ appartient à un document $d_i$ de la collection $C$ à un certain degré $\mu_C(d_i, t_j) \in [0, 1]$, qui est choisi de façon à refléter le degré de représentativité du terme par rapport au document (Waller *et al.,* 1979, Buell *et al.,* 1981). En théorie des sous-ensembles flous, la fonction $\mu$ 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 $E$. On note aussi :

$$d_i = \{\alpha_1/t_1, \ldots, \alpha_m/t_m\} \ , \qquad [1]$$

où $\{t_1, \ldots, t_m\}$ sont les termes présents dans $d_i$ et $\alpha_j = \mu_C(d_i, t_j)$ le degré d'appartenance du terme $t_j$ dans le document.

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 $\mu_q(t)$ pose le problème de l'interprétation des poids attribués. Ces poids correspondent le plus souvent à une préférence de l'utilisateur, mais peuvent aussi représenter d'autres grandeurs, comme la capacité de discrimination du terme.

#### 2.1.2. *Inclusion*

Dans le modèle classique, si l'on considère comme pertinents les documents qui contiennent tous les mots de la requête, la pertinence d'un document $d_i$ est donnée par la formule d'inclusion ensembliste suivante :

$$q \subseteq d_i \ . \qquad [2]$$

D'un point de vue axiomatique, il y a deux façons classiques de représenter cette inclusion, soit par une formule logique à base d'implication :

$$q \subseteq d_i \quad \Leftrightarrow \quad \forall t \in T,\ (t \in q \Rightarrow t \in d_i) \ , \qquad [3]$$

soit par une contrainte sur la cardinalité des ensembles :

$$q \subseteq d_i \quad \Leftrightarrow \quad \text{card}(q \cap d_i) = \text{card}(q) \ . \qquad [4]$$

Les sections suivantes présentent des modèles de RI flous obtenus par extension floue de ces deux représentations de l'inclusion, qui ne sont plus alors équivalentes.

Dans les modèles flous obtenus, l'inclusion n'est plus binaire, mais graduelle ; on obtient un degré d'inclusion compris entre 0 et 1. L'appariement entre requête et documents est effectué par cette mesure d'inclusion, et le degré d'inclusion ainsi obtenu, que nous assimilons au degré de pertinence, permet d'ordonner les documents.

Comme on le verra plus bas, on retrouve ainsi les deux étapes classiques des SRI. Tout d'abord, l'utilisation d'une fonction d'appariement qui calcule des scores individuels $S_q(d_i, t_j)$ pour chaque terme $t_j$ d'une requête $q$ et chaque document $d_i$. Ensuite, l'utilisation d'une fonction d'agrégation des scores individuels $S_q(d_i, t_j)$, $t_j \in q$, pour obtenir un score global $S_q(d_i)$ pour chaque document $d_i \in C$, 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, ces fonctions d'appariement et d'agrégation sont floues, donc à valeur dans l'intervalle unité.

### 2.2. *Inclusion par implication*

#### 2.2.1. *Travaux sur l'approche par implication*

L'extension floue de la formule 3 dans un modèle de RI a été initialement proposée dans (Pasi, 1999). L'implication matérielle y est remplacée par une implication floue.

Dans cette approche, l'appariement entre un document et une requête se fait au niveau des termes. Il est évalué par le degré d'implication $\mu_q(t) \rightarrow \mu_C(t, d_i)$. Ces degrés individuels sont agrégés par le quantificateur universel dans la notation classique de la formule [3]. En logique floue, ce quantificateur est remplacé par la conjonction floue min (la plus grande des conjonctions floues), selon le principe du minimum de spécificité.

Les degrés d'inclusion des termes sont ainsi agrégés pour obtenir un degré d'inclusion de la requête dans le document selon la formule :

$$\text{Inc}_q(d_i) = \min_{t \in q} \left(\mu_q(t) \rightarrow \mu_{d_i}(t)\right) \ , \qquad [5]$$

Ce degré $\text{Inc}_q(d_i)$ correspond à la notion de pertinence d'un document $d_i$ pour une requête $q_i$, tel qu'il est exprimé dans le modèle Booléen.

Indépendamment, cette approche a été à nouveau proposée en 2008 dans (Bosc *et al.,* 2008b). Les auteurs on travaillé plusieurs années sur la division de relations floues, dans le cadre des bases de données floues (voir par exemple (Bosc *et al.,* 1997)). Ayant remarqué le lien entre le modèle Booléen de RI et la division de relations, ils ont supposé que la division de relations floues pouvait conduire à un bon modèle de RI, étendant le modèle Booléen. Le modèle proposé correspond exactement à la formule 5. Une autre contribution de (Bosc *et al.,* 2008b) est l'utilisation d'un modèle d'inclusion tolérante (e.g. par relaxation du quantificateur) qui permet la prise en compte d'exceptions dans le calcul du degré d'inclusion.

Pour être complets, notons aussi que le lien entre l'extension floue de la formule [3] et la division de relations floues a été très brièvement mentionné dans (Baziz *et al.,* 2007).

Toutefois, tous ces travaux étaient purement théoriques, et ce type de modèle de RI flou n'avait pas été validé de façon expérimentale jusqu'aux travaux rapportés dans (Bosc *et al.,* 2009a).

#### 2.2.2. *Choix des opérateurs dans le modèle par implication*

L'extension du modèle classique obtenu dans la formule 5 laisse beaucoup de problèmes ouverts. Tout d'abord, on peut choisir parmi une large variété d'implications floues (R-implications ou S-implications), alors que la conjonction min semble imposée par la théorie. Et on a pu montrer dans (Bosc *et al.,* 2009a) que le choix de ces opérateurs est primordial pour l'obtention d'un système de RI performant. Ainsi, le choix de l'implication, parmi la famille des R-implications ou des S-implications, détermine la sémantique des poids des termes dans la requête, $\mu_q(t)$. Nous rappelons ci-dessous quelques propriétés importantes notés par les différents auteurs ayant utilisé ce modèle flou concernant ces opérateurs.

**Propriété d'absorption de l'opérateur min.** La conjonction min dans la formule 5 est requise par l'application principe de minimum de spécificité dans l'approche de (Pasi, 1999), et nécessaire pour que le résultat de la division de relations floues ait la propriété d'un quotient dans (Bosc *et al.,* 2008b).

Toutefois cet opérateur présente la mauvaise propriété d'être absorbant. De plus, dans la formule, il a un rôle d'agrégation des scores d'inclusion de chaque terme. Avec le min, le score global d'un document est alors donné par un seul terme (celui qui obtient le plus petit score individuel), alors que les scores des autres termes n'interviennent pas. A contrario, il est établi que les systèmes de RI performants de l'état-de-l'art sont ceux qui tiennent compte de tous les termes, selon différentes formules de compensation. Et de fait, les expériences de (Bosc *et al.,* 2008b) ont montré que les opérateurs qui ont une composante absorbante comme le min ($\min(a.b)$) ou la somme bornée ($\max(0, a+b-1)$) conduisent à de mauvaises performances, alors que les t-normes qui possèdent une composante de type produit, comme le produit ($a.b$), ou la t-norme d'Einstein ($a.b/(2-a-b+a.b)$) donnent les meilleurs résultats.

Ainsi, pour obtenir un système de RI fonctionnel, l'opérateur min doit être remplacé par une autre t-norme (conjonction floue). Pour obtenir un modèle flou performant, la formule représentant ce modèle devient donc :

$$\text{Inc}_q(d_i) = \top_{t \in q} \left(\mu_q(t) \rightarrow \mu_{d_i}(t)\right) \ . \qquad [6]$$

Ce remplacement pourrait apparaître comme une entorse à la théorie, toutefois, la formule [6] reste une mesure d'inclusion graduelle (dont le degré obtenu n'est pas maximal), ce qui est suffisant pour assurer la validité de ce modèle.

**Seuil et R-implications.** Dans le cas des R-implications (Fodor *et al.,* 1999), notées $\rightarrow_R$, le degré $\mu_q(t)$ est vu comme un seuil. La satisfaction complète est obtenue dès que $\mu_C(d,t)$ atteint ce seuil pour tous les termes $t$ de $q$. Quand ce seuil n'est pas atteint, une pénalité est appliquée. Toute R-implication peut d'ailleurs être réécrite de la façon suivante :

$$a \rightarrow_R b = 1 \text{ si } a \leq b, \quad f(a,b) \text{ sinon}, \qquad [7]$$

où $f(a,b)$ exprime une satisfaction partielle (inférieure à 1) quand l'antécédent $a$ n'est pas atteint par la conclusion $b$. L'interprétation en terme de seuil est claire dans (7), où l'implication vaut 1 dès que $\mu_C(d,t)$ atteint $\mu_q(t)$.

Cependant, une fois encore, cet effet de seuil est pénalisant dans un moteur de RI qui doit classer des documents selon leur pertinence, et pas seulement déterminer s'ils sont pertinents ou non. En effet, lorsque le seuil est atteint, la R-implication donne 1. Avec ces opérateurs, il est donc impossible de classer 2 documents qui contiennent les termes de la requête aux degrés requis, même si dans un document ces termes ont des poids plus forts que dans l'autre, montrant qu'ils décrivent mieux le document, et que celui-ci est plus pertinent.

Ce point a été vérifié expérimentalement : les expériences de (Bosc *et al.,* 2008b) ont montré que les R-implications ne donnent de bons résultats que lorsque les poids sont choisis de telle façon que le seuil requis n'est jamais atteint.

**Importance et S-implications.** L'autre grande famille d'implications floues est celle des S-implications. Dans la seconde interprétation, $\mu_q(t)$ définit l'importance du terme $t$ (et donc le degré $\mu_C(d,t)$ est modulé par cette importance). Dans ce cadre logique,

la notion sous-jacente est celle d'une satisfaction garantie (à un degré $> 0$) lorsque l'importance n'est pas totale : lorsque $\mu_q(t) < 1$, le terme $t$ n'est pas totalement important, et peut être absent jusqu'à un certain point.

Un document est totalement satisfaisant si $\mu_C(d,t) = 1$ pour toutes les valeurs $t$ de $q$ quelle que soit leur importance. Il est totalement insatisfaisant seulement si, pour au moins un terme $t$ dans $q$, on a conjointement $\mu_q(t) = 1$ (le terme est d'importance maximale) et $\mu_C(d,t) = 0$ (il est absent du document). Ce comportement est modélisé par une S-implication (Fodor *et al.,* 1999), notée $\rightarrow_S$, qui peut s'écrire :

$$p \rightarrow_S q \;=\; \perp(1-p, q) \;=\; 1 - \top(p, 1-q) \;, \qquad [8]$$

où $\perp$ est une t-conorme (une disjonction floue).

#### 2.2.3. *Résultats expérimentaux*

Un système basé sur ce modèle flou a été testé sur différentes collections de documents dans (Ughetto *et al.,* 2009b). Il a été paramétré en utilisant une pondération des termes des documents adaptée de BM25 (normalisée pour respecter les propriétés des degrés d'appartenance), et en testant de nombreux opérateurs flous. Il a été montré qu'avec un bon choix de paramètres et d'opérateurs, le système obtient alors des résultats comparables à OKAPI. Ces expérimentations ont aussi permis de déterminer les propriétés que doivent avoir les opérateurs flous pour obtenir un système performant.

### 2.3. *Inclusion par cardinalité*

L'autre vision axiomatique de l'inclusuion présenté dans la sous-section 2.1.2 peut elle aussi être étendue en logique floue. À notre connaissance, cette extension floue de la formule 4 dans un modèle de RI n'a été étudiée que dans le cadre de nos travaux (Bosc *et al.,* 2009b, Ughetto *et al.,* 2009a).

#### 2.3.1. *De l'implication à la cardinalité*

Souvent en RI, un document pertinent ne contient pas tous les termes de la requête. Dans les modèles vectoriels, l'absence d'un terme n'influe pas sur le score d'un document. Cette absence se traduit en fait par l'utilisation de l'élément neutre de la fonction agrégeant les scores terme à terme en un score global (par exemple 0 dans l'addition des scores individuels des modèles vectoriels, ou 1 dans le produit des modèles de langues). De plus, un terme très représentatif (rare dans la collection mais fréquent dans le document) augmente grandement ce score. Du point de vue des modèles vectoriels, auxquels nous avons comparé notre modèle flou dans un premier temps, les termes de score individuel fort sont donc plus importants que ceux de score faible.

Au contraire, notre approche par implication donne plus d'importance aux termes peu présents dans un document. Ce comportement est dû à l'utilisation conjointe d'une implication floue et d'une agrégation s'appuyant sur une T-norme (car le poids maximal 1 est l'élément neutre des T-normes). L'intuition derrière ce fonctionnement, plus

proche du monde des BD que du monde de a RI, est la suivante. Dans les documents ramenés par le système, il est *normal* que les termes soient présents (avec un score fort) et, si on tolère quelques exceptions (termes absents ou peu présents), c'est la pénalité engendrée par ces exceptions qui fait le score. Les termes de faible score ont alors une influence plus forte dans le score du document.

Cette considération nous a menés à une autre approche, plus focalisée sur les termes de la requête présents dans le document : l'approche basée cardinalité. Elle consiste à calculer la cardinalité (floue) de l'intersection entre $q$ et $d_i$, normalisée par la cardinalité (floue) de $q$. Ainsi, par construction, le calcul du score est plus proche de celui des systèmes de RI standard, qui dépend uniquement des termes que le document partage avec la requête.

2.3.2. *Caractéristiques du modèle*

L'inclusion classique basée sur un ratio de cardinalités, donné par la formule [4] est extensible aux ensembles flous :

$$Inc_q(d_i) = \frac{|q \cap d_i|}{|q|} \text{ si } |q| \neq 0,\ 1 \text{ sinon,} \qquad [9]$$

où $|E|$ est la cardinalité (floue) de $E$.

La notion de mesure d'inclusion floue généralisant $Inc$ et basée sur le concept d'entropie floue a été axiomatisée dans (Young, 1996). En utilisant la définition de cardinalité scalaire d'un ensemble flou, introduite dans (De Luca *et al.,* 1972) et souvent appelée cardinalité de Zadeh : $|E| = \sum_{x \in U} \mu_E(x)$, où $U$ est l'univers de $E$, et en utilisant une norme triangulaire $\top$ pour l'intersection, la formule [9] s'écrit :

$$Inc_q(d_i) = \frac{\sum_{x \in U} \top(\mu_q(x),\ \mu_{d_i}(x))}{\sum_{x \in U} \mu_q(x)} \text{ si } \Sigma_{x \in U} \mu_q(x) \neq 0, \quad 1 \text{ sinon.} \qquad [10]$$

Lorsque la requête n'est pas vide (ce qui en principe est toujours le cas), $1/\Sigma_{x \in U} \mu_q(x)$ est une constante strictement positive, que nous notons $k$ dans la suite, qui a pour rôle de normaliser la mesure d'inclusion, de façon à ce qu'elle appartienne à l'intervalle unité $[0, 1]$.

La fonction de score du modèle flou peut donc s'écrire :

$$Inc_q(d_i) = k. \sum_{x \in U} \top(\mu_q(x),\ \mu_{d_i}(x)) \ , \qquad [11]$$

avec $k = 1/\Sigma_{x \in U} \mu_q(x) > 0$.

Notons qu'une *division* de relations floues modélisée par une inclusion basée cardinalité n'est plus une division *stricto sensu* puisqu'en général son résultat n'est pas un quotient (Bosc *et al.,* 2007). Cependant, comme on l'a vu précédemment dans le modèle par inclusion, on s'est inspiré de la formule de la division floue sans chercher à en obtenir une ; la notion d'inclusion graduelle est suffisante pour la construction d'un modèle de RI.

2.3.3. *Résultats expérimentaux*

Pour permettre une comparaison équitable, cette approche a été testée dans les mêmes conditions que le modèle à base d'implication. Elle a été comparée à OKAPI, sur les mêmes collections de documents. Et les pondérations des termes des documents sont aussi adaptées de BM25 (avec un système de normalisation équivalent).

Le seul paramètre libre du modèle est donc la t-norme $\top$ (ET flou) de la formule [11]. De nombreux opérateurs ont été testés, et même si dans ce modèle, la t-norme ne remplit pas le même rôle que dans le modèle à base d'implication, des résultats similaires concernant leur adéquation dans ce modèle se sont fait jour : les opérateur de type min conduisent à de mauvais résultats, alors que ceux ayant une composante produit donnent les meilleurs résultats.

On peut noter que, lorsque la t-norme choisie dans [11] est le produit, et en reprenant les pondérations BM25 (à une normalisation près), la formule obtenue correspond est équivalente à celle du calcul du score dans OKAPI. Il n'est donc pas surprenant d'obtenir des résultats quasiment identiques à ceux d'OKAPI dans ce cas. Avec d'autres *bons* opérateurs, comme la t-norme d'Einstein, les résultats sont comparables. Les autres t-normes ont le plus souvent donné des résultats moins bons.

Pour les détails des expérimentations, le lecteur est invité à consulter (Bosc *et al.,* 2009b) ou (Ughetto *et al.,* 2009a).

## 3. Liens avec les modèles standards

La section précédente a présenté les modèles de RI flous à base d'inclusion graduelle dérivés du modèle du modèle Booléen. Dans cette section, nous montrons que ces modèles peuvent aussi être vu comme des extensions de certains autres modèles de RI classiques.

### 3.1. *Modèle booléen de RI*

Comme on l'a vu dans la section 2.1, le lien entre nos modèles flous et le modèle booléen de RI est direct. Les modèles flous sont construits comme des extensions du modèle booléen de RI.

La partie présentée ici ne considère que des requêtes *sacs de termes* ou, si on considère la requête comme une formule logique, une conjonction de termes. Dans le modèle à base d'implication, cette conjonction, correspond à la t-norme $\top$ dans la formule [6]. La formule peut facilement être étendue pour prendre en compte une disjonction (par exemple la t-conorme duale de $\top$) et une négation ($1 - x$).

Dans le modèle à base de cardinalité, la prise en compte des disjonctions et négations peut se faire de façon aussi simple, par des combinaisons d'ensembles ; les ratios

de cardinalités sont calculés sur plusieurs intersections, et combinés par des ET, OU et NON flous.

### 3.2. *Modèle flou à base d'implication et modèles logiques de RI*

#### 3.2.1. *Modèles logiques de RI*

Les modèles logiques de RI ont été l'objet de nombreuses études dans les années 1990. Keith van Rijsbergen a été l'un des premiers à proposer une interprétation logique de la recherche d'information au moyen du concept d'implication d'une requête par un document $d \rightarrow q$, où $\rightarrow$ est le connecteur d'implication formalisé dans la logique considérée (van Rijsbergen, 1986). Plusieurs études ont ensuite analysé le rôle potentiel de la logique comme cadre formel pour définir des modèles de RI. Sebastiani a présenté une très bonne analyse de l'état de l'art et des interprétations différentes proposées dans la littérature (Sebastiani, 1998), de même que Lalmas (Lalmas, 1998).

Dans cette approche, documents et requêtes sont représentés par des formules logiques (d'où le nom du modèle) : le plus souvent souvent une conjonction des termes qu'ils contiennent. Par exemple, si le document $d_i$ est défini par l'ensemble de termes $\{t_1, \ldots, t_n\}$, il sera représenté par la formule $d_i = t_1 \wedge \ldots \wedge t_n$. La requête $q$ est aussi représentée par une formule logique, une conjonction de termes dans les approches de type « sac de termes », ou une formule plus générale (qui peut contenir des conjonctions et des disjonctions), mise sous forme normale.

Pour déterminer si un document $d_i$ est pertinent pour une requête $q$, un modèle logique de RI cherche à vérifier le statut de la formule $d_i \rightarrow q$. Si elle est valide, le document est pertinent. Comme on l'a vu plus haut, cela peut se faire de quatre façons, qui sont équivalentes en logique propositionnelle. Dans d'autres logiques, elles ne sont pas toujours équivalentes, et parfois certaines ne sont pas possibles. Les deux premières proviennent de la théorie des modèles. On peut montrer que :

– $\models d_i \rightarrow q$ : la formule $d_i \rightarrow q$ est valide (i.e., vraie quelle que soit la valeur de vérité des termes $t_j$),

– $d_i \models q$ : la formule $q$ est conséquence logique de $d_i$ (i.e., les interprétations qui vérifient $d_i$ vérifient aussi $q$).

Les deux autres proviennent de la théorie de la preuve. On peut montrer que :

– $\vdash d_i \rightarrow q$ : la formule $d_i \rightarrow q$ est un théorème (démontrable par une méthode de preuve),

– $d_i \vdash q$ : la formule $q$ peut être déduite de la formule $d_i$ (par une méthode de preuve).

Le lecteur peut trouver des exemples dans (Lalmas, 1998).

3.2.2. *Deux notations pour un même modèle*

À première vue, les modèles ensemblistes (comme le modèle booléen) et les modèles logiques conduisent à deux modélisations du problème de RI qui semblent opposées. En effet, on obtient $q \rightarrow d_i$ dans le premier cas, et $d_i \rightarrow q$ dans le second.

Cependant, on peut montrer qu'il s'agit simplement d'un jeu d'écriture. Le modèle logique de RI qui utilise la logique des propositions correspond en fait au modèle booléen de RI. Le modèle flou à base d'implication, présenté section 2.2 peut donc être vu comme une généralisation du modèle logique. Si on aboutit à l'implication $q \rightarrow d_i$ dans le premier cas, et $d_i \rightarrow q$ dans l'autre, c'est seulement une conséquence de la formalisation du problème, et le fait que $d_i$ et $q$ ne représentent pas la même chose dans les deux cas.

Dans le cas des requêtes conjonctives, l'équivalence est immédiate. Le modèle booléen dit que le document est pertinent si tous les termes de la requête sont présents dans le document. Le modèle logique dit que dans toute valuation qui vérifie $d$ (i.e. lorsque les termes du document sont considérés comme vrais, quelle que soit la valeur de vérité des autres), la formule $q$ doit être vraie. Supposons que la formule $q$ est la conjonction des termes. Pour que $q$ soit vraie, il faut que tous ses termes soient vrais, et donc qu'ils soient aussi présents dans le document. On retrouve la condition d'inclusion des termes de la requête dans le document. La formulation diffère, mais la condition est identique ! Dans le cas de requêtes plus générales, cette équivalence peut être démontrée formellement.

Les représentations sous la forme $q \rightarrow d$ et $d \rightarrow q$ que l'on retrouve dans la littérature ne sont donc opposées que par des jeux d'écriture ; elles signifient en fait la même chose, sous des formalismes différents. Il serait d'ailleurs plus juste de noter $d \rightarrow q$ pour les modèles logiques où $d$ et $q$ sont des formules représentant respectivement le document entier et la requête entière, et $q(t) \rightarrow d(t)$ (voire $\forall t,\ q(t) \rightarrow d(t)$) pour les modèles ensemblistes, qui travaillent terme à terme.

On peut en conclure que le modèle flou à base d'implication est un modèle logique de RI, utilisant la logique floue.

## 3.3. *Modèles vectoriels*

3.3.1. *Formules de score dans les modèles vectoriels*

Dans les modèles vectoriels, chaque document $d_i$ est représenté par un vecteur donc chaque composante est un terme $t \in T$. La valeur $w_{t,d_i}$ de chaque composante dépend du schéma de pondération adopté. Le poids d'un terme absent est le plus souvent considéré comme nul.

De manière similaire, les requête sont également représentées par des vecteurs, avec un également un schéma de pondération identique ou non à celui des documents. Le score d'un document est alors donné par une mesure de similarité (souvent la me-

sure de cosinus, équivalente à une distance $L_2$ quand les vecteurs sont normalisés) entre le vecteur de la requête et celui du document, ce qui donne, après normalisation par la longueur du document et de la requête :

$$\text{sim}(d_i, q) = \frac{\sum_{t \in q} w_{t,d_i} \cdot w_{t,q}}{\sqrt{\sum_{t \in q} w_{t,d_i}^2} \cdot \sqrt{\sum_{t \in q} w_{t,q}^2}} \ , \qquad [12]$$

où $w_{t,d_i}$ est le poids du terme $t$ dans le document $d_i$, et $w_{t,q}$ le poids du terme $t$ dans la requête $q$. En notant $1/k_{d_i} = \sqrt{\sum_{t \in q} w_{t,d_i}^2}$ la longueur du vecteur document et $1/k_q = \sqrt{\sum_{t \in q} w_{t,q}^2}$ la longueur du vecteur de la requête, la formule devient :

$$\text{sim}(d_i, q) = \sum_{t \in q} k_{d_i}.w_{t,d_i} \cdot k_q.w_{t,q} \ , \qquad [13]$$

qui forme un schéma général des calculs de score dans les modèles vectoriels.

#### 3.3.2. *Généralisation par les modèles flous*

Ces formules de score sont souvent vues comme un calcul de score individuel de chaque terme $t \in T$, obtenu par une fonction d'appariement (le produit) des poids du terme dans le document $w_{t,d_i}$ et dans la requête $w_{t,q}$, suivi d'une agrégation de ces scores (par la somme).

On retrouve ce fonctionnement dans le modèle flou à base d'implication, représenté par la formule [6], où les poids des termes dans la requête et le document sont appariés par une implication floue, puis agrégés par une conjonction floue. Les poids peuvent être calculés de la même façon, ce que nous avons fait en reprenant le modèle de pondération $BM25$ en le normalisant.

Ce lien est encore plus direct avec le modèle flou d'inclusion basé cardinalité représenté par la formule 9. En effet, le fonction d'agrégation est aussi la somme, et la fonction d'appariement est une conjonction floue, que l'on peut choisir comme étant par exemple le produit.

Ainsi, ces modèles flous peuvent également être vus directement comme des généralisations du modèle vectoriel de RI. Ils peuvent ainsi bénéficier des diverses améliorations apportées aux mécanismes de pondération du type tf-idf.

### 3.4. *Modèles de langue*

#### 3.4.1. *Des modèles de langue à la RI*

Un modèle de langue est une fonction qui attribue une probabilité à un terme ou une suite de termes de la langue, à partir d'un corpus. Le modèle le plus utilisé est le $n$-gramme, qui suppose que la probabilité d'apparition d'un terme ne dépend que des $n-1$ termes qui le précèdent. En RI, c'est ce modèle n-gramme qui souvent

utilisé sous sa forme la plus simple, c'est-à-dire avec des unigrammes. Ce modèle ne tient donc pas compte de l'ordre des mots, ce qui reveint à adopter de nouveau une représentation *sac de mots*. Ainsi, la probabilité pour qu'un terme $t$ soit générée par le document $d_i$ est estimé par la fréquence du terme dans le document, normalisée par la longueur du document :

$$P(t|d_i) = \frac{\text{tf}_{t,d_i}}{\sum_{u \in d_i} \text{tf}_{u,d_i}} \ . \qquad [14]$$

Le score d'un document $d_i$ pour une requête $q$ est la probabilité que le document engendre la requête ; il est donné par le produit des probabilités individuelles des termes de la requête :

$$\text{score}(d_i, q) = \prod_{t \in q} P(t|d_i) \ . \qquad [15]$$

Cependant, lorsqu'un terme de la requête est absent du document, la formule [14] lui attribue une probabilité nulle, et le produit dans [15] conduit à un score nul pour le document. Pour rendre ce score tolérant à l'absence de termes de la requête et pour avoir de meilleures estimations des probabilités, de nombreuses méthodes de lissage ont été proposées ; elles permettent d'attribuer un score non nul dans [14] à tout terme de la collection. Par exemple, dans le modèle de Hiemstra et Kraaij (Hiemstra *et al.,* 1999), la probabilité *lissée* est obtenue par interpolation entre la probabilité que le terme soit engendré par le document, et celle que le terme soit engendré par la collection :

$$P_l(t|d_i) = \lambda . P(t|d_i) + (1 - \lambda) P(t|C) \qquad \lambda \in ]0, 1[ \ . \qquad [16]$$

### 3.4.2. *Modèles de langue et modèles vectoriels*

Dans (Hiemstra *et al.,* 1999), les auteurs font remarquer que la formule de leur modèle de RI à base de modèle de langue peut se réécrire sous une forme équivalente à celle d'un modèle vectoriel. En effet, après lissage, la formule de score [15] devient :

$$\text{score}(d_i, q) = \prod_{t \in q} \left( \lambda . \frac{\text{tf}_{t,d_i}}{\sum_{u \in d_i} \text{tf}_{u,d_i}} + (1 - \lambda) . \frac{\text{df}_t}{\sum_{u \in C} \text{df}_u} \right) \ , \qquad [17]$$

$$\propto \sum_{t \in q} \text{tf}_{t,d_i} . \log \left( 1 + \frac{\text{tf}_{t,d_i}}{\text{df}_t . \sum_{u \in d_i} \text{tf}_{u,d_i}} + \frac{\lambda . \sum_{u \in C} \text{df}_u}{(1 - \lambda)} \right) \qquad [18]$$

À des constantes près, on retrouve un score sous forme d'un modèle vectoriel (où apparaissent même l'équivalent d'un TF et d'un IDF) :

$$\text{sim}(d_i, q) = \sum_{t \in q} w_{t,d_i} \cdot w_{t,q} \ , \qquad [19]$$

où $w_{t,d_i} = \text{tf}_{t,d_i}$ et où $w_{t,d_i} = \log \left( 1 + \frac{\text{tf}_{t,d_i}}{\text{df}_t . \sum_{u \in d_i} \text{tf}_{u,d_i}} + \frac{\lambda . \sum_{u \in C} \text{df}_u}{(1 - \lambda)} \right)$.

3.4.3. *Modèles floues et modèles de langue*

Dans la mesure où les modèles flous sont assimilables à des modèles vectoriels (cf. section 3.3.2) et où certains modèles de langue peuvent être vus comme des modèles vectoriels (cf. section 3.4.2), il est naturel de s'interroger sur les liens entre modèles flous et modèles de langue.

Dans le modèle flou à base d'implication, l'opérateur principal de la formule [6] est une conjonction, comme dans la formule des modèles de langues [15]. On peut noter aussi que les probabilités, comme les degrés d'appartenance sont à valeur dans l'intervalle unité.

Ces propriétés similaires soulèvent d'ailleurs des problèmes similaires. ainsi, le problème des probabilités nulles pour les termes absents se retrouve dans le modèle flou. Il se limite au cas où un terme absent du document reçoit un poids nul, et où ce terme reçoit un poids maximal de 1 dans la requête. Pour éviter cette situation, dans l'implémentation de (Bosc *et al.,* 2009a), les poids ont été normalisés, pour être toujours supérieurs à une faible valeur fixée $\epsilon$. Ce mécanisme peut en fait être considéré comme un lissage par *absolute discounting* (Ney *et al.,* 1994). Le modèle flou basé sur les cardinalités peut aussi se réécrire sous la forme d'un modèle de langue, en appliquant la transformation inverse de celle proposée par Hiemstra, et présentée dans la section précédente. Les différentes versions des modèles flous présentées sont donc là encore très directement assimilables à des modèles de langue.

## 3.5. *Extension directe des modèles de langue*

Récemment, dans (Harastani, 2010), une extension floue directe du modèle de Hiemstra a été étudiée. Dans la formule de score du modèle de Hiemstra :

$$\text{score}(d_i, q) = \prod_{t \in q} \Big(\lambda.P(t|d_i) + (1-\lambda)P(t|C)\Big) \qquad \lambda \in ]0,1[\ . \qquad [20]$$

l'opérateur produit a tout d'abord été remplacé par une conjonction floue :

$$\text{score}(d_i, q) = \top_{t \in q} \Big(\lambda.P(t|d_i) + (1-\lambda)P(t|C)\Big) \qquad \lambda \in ]0,1[\ . \qquad [21]$$

et de nombreux opérateurs ont été testés expérimentalement, dans des conditions similaires aux autres expérimentations mentionnées section 2.2.

Il est ainsi montré qu'avec les bons opérateurs, les résultats obtenus par cette extension floues sont similaires ceux de Hiemstra, ce qui valide cette approche. D'une part, ces expérimentations ont confirmé ce que nous avions appris sur les performances des conjonctions floues en RI, à savoir que les opérateurs avec une composante produit, ou les opérateurs paramétriques dont le comportement se rapproche du produit, donnent les meilleurs résultats, alors que ceux présentant une propriété d'absorption en sont pas souhaitable pour agréger des scores.

Dans un deuxième temps, d'autres opérateurs d'agrégation flous ont été testés, à la place de $\top$ dans la formule [21], comme par exemple des opérateurs de moyenne, avec globalement les mêmes variations autour des résultats du modèle de Hiemstra.

## 4. Conclusion et perspectives

Nous avons rappelé le principe de fonctionnement des modèles de RI flous fondés sur des inclusions graduelles, soit à base d'implication, soit à base de cardinalité. Nous avons pu montrer que ces deux familles de modèles (ou parfois une seule) peuvent être vues comme différents modèles de RI classiques : des modèles booléens étendus, des modèles logiques, des modèles vectoriels, et des modèles de langue. Ces modèles flous se voulant des généralisations des modèles classiques, il y a beaucoup de pistes à explorer pour maintenant considérer d'autres opérateurs que ceux utilisés pour "imiter" les systèmes classiques. Notamment, dans le cas de l'extension floue du modèle de Hiemstra, beaucoup extensions restent à explorer, en particulier sur le mécanisme de lissage par interpolation. En effet, si on s'affranchit des probabilités, pour interpréter les poids par des préférences ou des possibilités, on peut en particulier revoir le mécanisme de pondération entre $P(t|d_i)$ et $P(t|C)$.

En prenant le point de vue opposé, il est intéressant de noter que les fonctions de scores de tous ces modèles, qui ont pourtant des fodnements théoriques variés, se ramènent à des mesures d'inclusion graduelles. Ces liens seront exploités dans de futurs travaux, pour essayer de déterminer comment les caractéristiques et qualités de chaque modèle classique peuvent nous permettre d'obtenir des modèles flous plus efficaces.

## 5. Bibliographie

Baziz M., Boughanem M., Loiseau Y., Prade H., « Fuzzy Logic and Ontology-based Information Retrieval », *in* , P. Wang, , D. Ruan, , E. Kerre (eds), *Studies in Fuzziness and Soft Computing*, vol. 215/2007, Springer, p. 193-218, 2007.

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., « Graded-Inclusion-Based Information retrieval Systems », *Proceedings of the European Conference on Information Retrieval, ECIR'09*, Toulouse, France, p. 321-336, 2009a.

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.

Bosc P., Rocacher D., Pivert O., « Characterizing the result of the division of fuzzy relations », *International Journal of Approximate Reasoning*, vol. 45, p. 511-530, 2007.

Bosc P., Ughetto L., Pivert O., Claveau V., « Implication-Based and Cardinality-Based Inclusions in Information Retrieval », *Proceedings of the IEEE International Conference on Fuzzy Systems (Fuzz-IEEE'09)*, Jeju Island, South Korea, p. 2088-2093, 2009b.

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.

De Luca A., Termini S., « A definition of non-probabilistic entropy in the setting of fuzzy sets theory », *Information and Control*, vol. 17, p. 301-312, 1972.

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.

Harastani R.*, « Information Retrieval : From Language Models to Fuzzy Logic »*, Master's thesis, Université de Rennes 1 - IRISA, Rennes, France, 2010.

Hiemstra D., Kraaij W., « Twenty-One at TREC-7 : ad-hoc and cross-language track », *Proceedings of the 7th Text Retrieval Conference TREC-7, NIST Special Publication 500-242*, p. 227-238, 1999.

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

Ney H., Essen U., Kneser R., « On Structuring Probabilistic Dependencies in Stochastic Language Modelling », *Computer Speech and Language*, vol. 8, p. 1-38, 1994.

Pasi G., « A logical formulation of the Boolean model and of weighted Boolean models », *LUMIS workshop at ECSQARU'99*, Londres, 1999.

Sebastiani F., « On the role of logic in Information Retrieval », *Information Processing and Management*, vol. 34, n$^o$ 1, p. 1-18, 1998.

Ughetto L., Pasi G., Claveau V., Pivert O., Bosc P., « Implication in Information Retrieval Systems », *in* , G. Pasi (ed.), *e-Proceedings of the 9th international conference on Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO'09)*, Paris, France, 2010.

Ughetto L., Pivert O., Claveau V., Bosc P., « Recherche d'information et inclusions graduelles », *Actes des Journées Francophones sur la Logique Floue et ses Applications (LFA'09)*, Annecy, France, p. 125-132, 2009a.

Ughetto L., Pivert O., Claveau V., Bosc P., « SRI à base d'inclusion graduelle », *Actes de la COnférence en Recherche d'Infomations et Applications (CORIA'09)*, Presqu'île de Giens, France, p. 235-250, 2009b.

van Rijsbergen C. J., « A non-classical logic for information retrieval », *The Computer Journal*, vol. 29, n$^o$ 6, p. 481-485, 1986.

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

Young V., « Fuzzy subsethood », *Fuzzy Sets & Systems*, vol. 77, p. 371-384, 1996.