L'Agrégation en Recherche d'Information

Une revue critique des principaux modèles théoriques de Recherche d'Information

Mohamed FARAH, Daniel VANDERPOOTEN

Laboratoire LAMSADE
Place du Maréchal de Lattre de Tassigny
75775 Paris Cedex 16

RÉSUMÉ. *Ce papier donne une nouvelle présentation des modèles standards de la recherche d'information, décrits selon deux dimensions. La première porte sur les sources d'évidence qu'utilisent les modèles et la manière dont ils les agrègent pour mesurer l'importance ou le poids d'un terme dans un document. La seconde concerne la manière dont ces poids sont agrégés pour calculer un score de pertinence. Les mécanismes d'agrégation utilisés dans les deux cas sont alors explicités et critiqués motivant le recours à une nouvelle famille de méthodes basées sur de nouveaux mécanismes d'agrégation plus adaptés.*

ABSTRACT. *This paper gives a new presentation of well established Information Retrieval models which are herein described according to two dimensions. The first one concerns the sources of evidence used by each model and the way they are combined to measure the importance or the weight of each query term in a document. The second dimension relates to the way these weights are aggregated to calculate a score of relevance. Aggregation mechanisms used in both cases are highlighted and criticized, which leads to the use of a novel family of retrieval approaches based on more suited aggregation mechanisms.*

MOTS-CLÉS : *Modèles de Recherche d'Information, Agrégation, Pertinence, Analyse Multicritère.*

KEYWORDS : *Information Retrieval Models, Aggregation, Relevance, Multiple Criteria Analysis.*

CORIA 2007, pp. 125–136
Saint-Étienne, 28–30 mars 2007

1. Introduction

La recherche d'information (RI) concerne la représentation, le stockage et l'accès à des documents (texte, image, son ou vidéo) d'une collection. De manière générale, le processus de RI fait intervenir trois entités : un utilisateur, une collection D de n_d documents et un système de RI. L'utilisateur, ayant un besoin en information, interroge le système de RI en lui soumettant une requête q, souvent sous la forme d'une liste de mot-clés, et attend de recevoir comme réponse une liste de documents de la collection répondant potentiellement à cette requête, qu'on appelle documents pertinents. Les systèmes de RI se distinguent par la façon d'interpréter et représenter les documents et les requêtes, mais aussi par la définition exacte de la notion de *pertinence* qui permet de sélectionner et classer les documents soumis à l'utilisateur.

Ce papier vise à donner une nouvelle présentation des modèles de RI traduisant différentes stratégies de mise en correspondance des documents et des requêtes, en montrant que l'agrégation est au coeur de ces modèles. Après avoir défini formellement un modèle théorique de RI (section 2), nous présentons les principaux modèles de RI en distinguant la phase de construction des sources de pertinence (section 3) et la phase de construction du rangement final (section 4). La première phase vise à construire un *modèle de description* formalisant les différentes sources de pertinence. La deuxième phase définit une stratégie d'agrégation de ces sources de pertinence s'intégrant dans un *modèle de rangement*. Les principales caractéristiques et limites des mécanismes d'agrégation utilisés lors de ces deux phases sont par la suite présentées (section 5). Une nouvelle famille de méthodes de RI est alors présentée (section 6) palliant les défauts des approches classiques. Une conclusion termine ce papier.

2. Modèles de RI

Dans la littérature en RI, deux aspects principaux ont un impact direct sur la définition de la notion de pertinence et par conséquent sur l'efficacité des systèmes de RI. Il s'agit d'une part de la définition de l'ensemble des facteurs à considérer pour définir la pertinence. Ces facteurs sont appelés *sources de pertinence*. Il s'agit d'autre part de la manière de combiner ces sources de pertinence afin de bâtir un rangement des documents. La plupart des modèles classiques en RI mesurent la pertinence d'un document vis-à-vis d'une requête par un score agrégeant les différentes sources de pertinence retenues. Celles-ci sont combinées à *deux niveaux d'agrégation*, donnant lieu dans un premier temps au calcul des *poids des termes* de la requête, et dans un deuxième temps au calcul d'un score qui sert de base au rangement final des documents.

Formellement, nous donnons la définition suivante de tout modèle théorique de RI.

Définition 1 (Modèle de RI) *Un modèle de RI est un quadruplet ($\mathcal{D}$, $\mathcal{F}$, $\mathcal{E}$, $\mathcal{P}$) où :*

– $\mathcal{D}$ est l'ensemble de documents à évaluer et à ranger selon leur degré de pertinence à une requête utilisateur,

– $\mathcal{F}$ désigne la famille des critères, éventuellement conflictuels, à partir desquels les documents seront évalués. Cette famille représente les sources de pertinence,

– $\mathcal{E}$ est l'ensemble des évaluations des des documents selon chacun des critères, c'est-à-dire l'ensemble des vecteurs de performances, un vecteur par document, et

– $\mathcal{P}$ désigne la procédure de rangement qui exploite les évaluations des documents afin de construire une liste triée de documents potentiellement pertinents.

Nous distinguons deux types de modèles de RI correspondant aux deux niveaux d'agrégation : les *modèles de description* $(\mathcal{D},\ \mathcal{F},\ \mathcal{E},\ *)$ qui visent à construire $\mathcal{E}$ à partir de $\mathcal{D}$ et $\mathcal{F}$, et les *modèles de rangement* $(*,\ *,\ *,\ \mathcal{P})$ qui définissent $\mathcal{P}$.

3. Modèles de description

Nous distinguons deux types de modèles de description selon que les sources de pertinence sont issues du contenu des documents (modèles de description textuels) ou de la structure des référencements (modèles de description structurels).

3.1. *Modèles de description textuels*

Pour ces modèles de description, les sources de pertinence correspondent souvent aux différents termes pouvant caractériser les liens sémantiques entre les documents et la requête. À chaque terme est associé un *poids* qui mesure la contribution de ce terme dans la définition de la pertinence. Les différents modèles de description textuels se distinguent essentiellement par la façon de calculer ces poids. Il s'agit essentiellement des modèles algébriques, probabilistes et ensemblistes (Baeza-Yates *et al.,* 1999).

3.1.1. *Modèle vectoriel*

Dans ce modèle (Salton, 1968) les facteurs qui sont utilisés pour le calcul des poids des termes sont essentiellement liés à la forme d'occurrence des termes dans chaque document et dans toute la collection, ainsi qu'à des caractéristiques intrinsèques des documents. Parmi les facteurs les plus souvent utilisés, figurent :

– le nombre d'occurrences d'un terme t_h dans un document d_i, appelé fréquence locale ($tf_{i,h}$).

– la rareté du terme t_h dans la collection D. Elle est souvent mesurée sur la base du nombre d'occurrences du terme t_h dans la collection, appelé fréquence globale (df_h) Parmi les diverses formulations de cette rareté, notée idf_h, nous avons celle de (Sparck Jones, 1972) : $idf_h = \log \frac{n_d}{df_h} + 1$.

– la longueur du document (dl_i).

Ces facteurs sont combinés pour calculer les poids $w_{i,h}$ de chaque terme t_h dans le document d_i. (Salton *et al.,* 1983) proposent la combinaison *tf.idf* suivante mais d'autres combinaisons existent (Savoy *et al.,* 2001) :

$$w_{i,h} = tf_{i,h} * \log \frac{n_d - df_h}{df_h}$$

3.1.2. *Modèle probabiliste standard*

La plupart des modèles probabilistes actuels sont fondés sur la formulation de (Robertson *et al.,* 1976) connue sous le nom BIR ('*Binary Independence Retrieval*').

Ce modèle suppose que des jugements de pertinence sont fournis, notamment, suite à une interaction avec l'utilisateur. La pertinence est alors définie sur la base de la distribution des termes de la requête dans les documents pertinents et les documents non pertinents. (Robertson *et al.,* 1976) proposent la formule générale suivante pour le calcul des poids des termes :

$$w_h = \log \frac{p_h * (1 - q_h)}{q_h * (1 - p_h)}$$

où p_h est la probabilité que le terme t_h soit présent dans un document pertinent, et q_h est la probabilité que le terme t_h soit présent dans un document *non* pertinent.

D'autres variantes du modèle probabiliste BIR ont été proposées. Par exemple, (Robertson *et al.,* 1994) ont développé le modèle Okapi BM25 dans lequel le calcul des poids des termes intègre des aspects relatifs à la fréquence locale des termes, leur rareté et la longueur des documents :

$$w_{i,h} = log\left(\frac{n_d - df_h + 0,5}{df_h + 0,5}\right) * \frac{(k_1 + 1) * tf_{i,h}}{k_1 * \left((1 - b) + b * \frac{dl_i}{\overline{dl}}\right) + tf_{i,h}}$$

où $\overline{dl} = moy_{d_i \in D}(dl_i)$ et k_1, b sont des paramètres qui dépendent de la collection ainsi que du type des requêtes.

3.1.3. *Modèle de langage probabiliste*

Initialement développé par (Ponte *et al.,* 1998), ce modèle construit un modèle de langage M_{d_i} pour chaque document d_i et mesure la probabilité qu'une requête soit produite à partir d'un modèle de langage donné. Cette mesure est complètement définie par la donnée, pour chaque terme t_h, de sa probabilité de provenir d'un modèle de langage donné M_{d_i}, c'est-à-dire $prob(t_h|M_{d_i})$, définissant ainsi les poids des termes. (Ponte *et al.,* 1998) calculent le poids d'un terme comme suit :

$$w_{i,h} = \lambda_h * \left(\frac{tf_{i,h}}{dl_i}\right)^{(1-\hat{r}_{i,h})} * \left(\sum_{d_i : t_h \in d_i} \frac{tf_{i,h}}{df_h * dl_i}\right)^{\hat{r}_{i,h}} + \frac{(1 - \lambda_h) * \sum_{d_i} tf_{i,h}}{\sum_{t_h} \sum_{d_i} tf_{i,h}}$$

où $\hat{r}_{i,h} = \left(\frac{1}{1 + \frac{\sum_{d_i} tf_{i,h}}{df_h}}\right) * \left(\frac{\sum_{d_i} tf_{i,h}}{1 + \frac{\sum_{d_i} tf_{i,h}}{df_h}}\right)^{tf_{i,h}}$ et λ_h est un paramètre propre à t_h.

Des variantes de ce modèle ont été proposées. Par exemple, les travaux de (Liu *et al.,* 2004) enrichissent le modèle de langage d'un document par des informations provenant de documents similaires. (Gao *et al.,* 2005) intègrent la co-occurrence des termes dans le calcul de leurs poids.

3.1.4. *Modèle DIA ('Darmstadt Indexing Approach')*

Selon (Fuhr *et al.,* 1991), le modèle BIR présente un inconvénient majeur, les informations collectées sur la pertinence n'étant applicables que localement, pour une

seule requête. Ces auteurs proposent une autre façon d'indexer qui consiste à travailler avec des *descripteurs de pertinence* ('*relevance descriptions*'). Un descripteur de pertinence est un vecteur $x_{i,h}$ où chaque élément correspond à une mesure de performance du document d_i contenant le terme t_h. Leur modèle se base sur une approche appelée DIA ('*Darmstadt Indexing Approach*') initialement utilisée pour les besoins de l'indexation automatique. Ce modèle probabiliste calcule la probabilité de pertinence d'un document d_i, contenant le terme t_h de la requête, en fonction de la répartition du descripteur de pertinence de ce terme t_h entre les documents pertinents et non pertinents des requêtes précédentes. Cette dernière mesure de probabilité définit les poids $w_{i,h}$. Elle peut être calculée sur la base d'un simple calcul de fréquences. Elle peut aussi être calculée en utilisant des fonctions d'indexation (Fuhr *et al.,* 1991).

3.1.5. *Modèle des ensembles flous*

La représentation des documents par les seuls termes qui les composent est considérée comme une description partielle de leurs contenus (Ogawa *et al.,* 1991). Dans le modèle des ensembles flous, un document peut être représenté par des termes qui ne font pas partie de sa description initiale mais qui ont des liens sémantiques avec elle. Pour cela, on définit la proximité sémantique de deux termes t_h et $t_{h'}$ par le coefficient $c_{h,h'} = \frac{df_{h,h'}}{df_h + df_{h'} - df_{h,h'}}$ où $df_{h,h'}$ est le nombre de documents contenant t_h et $t_{h'}$. Ainsi, le lien sémantique entre un terme t_h et un document d_i est donné par :

$$\mu_{i,h} = \perp_{t_k \in d_i} c_{h,k} \qquad \text{où } \perp \text{ est un opérateur d'agrégation disjonctif flou.}$$

Le poids d'un terme t_h d'un document d_i peut être donné par $w_{i,h} = \mu_{i,h}$. D'autres formules peuvent également être utilisées incluant des facteurs tels que $tf_{i,h}$ et df_h.

3.2. *Modèles de description structurels*

Lorsque les documents d'une collection sont connectés par référencement, on peut faire correspondre à cette collection un graphe $G = (D, L)$ où D représente l'ensemble des documents et L l'ensemble des liens de référencement. Selon la répartition des liens dans ce graphe, chaque document peut paraître plus ou moins important. La matrice d'adjacence (*documents* $\times$ *documents*) de ce graphe est notée $M_{D,D} = (m_{i,i'})_{i,i'=1,\ldots,n_d}$. Nous présentons deux modèles de références.

3.2.1. *PageRank*

(Brin *et al.,* 1998) ont proposé un modèle de RI, basé sur les travaux en bibliométrie, dans lequel la notoriété ou l'importance d'un document dépend de celle des documents citants. Dans ce modèle, les documents sont considérés comme des états ergodiques d'une chaîne de Markov où la probabilité de passage d'un document d_i à un document $d_{i'}$ est donnée par :

$$p_{i,i'} = \left(p * \frac{m_{i,i'}}{d^+(d_i)}\right) + \left((1-p) * \frac{1}{n_d}\right)$$

où p est un paramètre et $d^+(d_i)$ est le demi-degré extérieur de d_i.

La mesure de l'importance des documents, intitulée '*PageRank*', correspond à la distribution stationnaire de la chaîne de Markov. Elle peut être calculée à l'issue d'un calcul itératif utilisant la formule de mise à jour de l'équation suivante :

$$PageRank_i = \sum_{d_k \in D} p_{k,i} \cdot \ PageRank_k$$

3.2.2. *HITS ('Hypertext-Induced Topic Selection')*

(Kleinberg, 1999) confère à chaque document d_i deux rôles ou statuts : le rôle *élu* ('*authority*') en tant que document pertinent pour une requête et le rôle *électeur* ('*hub*') en tant que document pointant sur un document pertinent.

On suppose alors qu'un 'bon' élu est pointé par plusieurs 'bons' électeurs et vice versa. Pour chaque document d_i, on calcule deux scores : $auth_i$ en tant qu'élu et hub_i en tant qu'électeur. Ces scores sont obtenus en appliquant un algorithme itératif. À l'itération l, ces scores sont actualisés selon les formules suivantes puis normalisés :

$$auth_i^{<l>} = \sum_{d_{i'} \in D} m_{i',i} \cdot hub_{i'}^{<l-1>}; \qquad hub_i^{<l>} = \sum_{d_{i'} \in D} m_{i,i'} \cdot auth_{i'}^{<l>}$$

On montre que $auth$ est le vecteur propre principal de la matrice $M_{D,D}^t M_{D,D}$ et que hub est le vecteur propre principal de la matrice $M_{D,D} M_{D,D}^t$.

4. Modèles de rangement

Pour pouvoir ranger les documents dans un ordre décroissant de pertinence ou de probabilité de pertinence, il est nécessaire de combiner les différentes sources de pertinence qui sont définies dans la phase de description. Les modèles de rangement calculent pour chaque document un score, appelé *rsv* ('*retrieval status value*') et rangent les documents selon l'ordre décroissant de ce score. La majorité des modèles de rangement correspondent à des opérateurs d'agrégation analytiques. Dans cette section, nous passons en revue les opérateurs les plus souvent utilisés dans la littérature en précisant les modèles de description auxquels ils s'appliquent le plus souvent.

4.1. *Opérateur de type somme pondérée*

La façon la plus directe et intuitive pour agréger les poids des termes est d'utiliser l'opérateur de somme pondérée. C'est le cas du modèle probabiliste standard, du modèle de langage probabiliste, ou du modèle DIA. En effet, ces modèles rangent les documents en se basant sur le principe de rangement probabiliste de (Robertson, 1977) qui stipule que la performance des systèmes est meilleure lorsque les documents sont rangés selon un ordre décroissant de leurs probabilités de pertinence. Dans ces modèles, la pertinence d'un document d_i à une requête q est donnée par :

$$rsv(d_i, q) = \sum_{t_h \in q \cap d_i} w_{i,h}$$

4.2. *Mesures de similarités*

Utilisées principalement dans le modèle vectoriel, ces mesures traduisent une proximité algébrique des vecteurs représentatifs des documents et requêtes. Dans ce cas, les composants de ces vecteurs correspondent aux poids des termes qui sont calculés pour les documents ainsi que pour les requêtes :

$$rsv(d_i, q) = \sum_{h=1}^{n_t} w_{i,h} \cdot w_{q,h}$$

où $w_{q,h}$ est le poids du terme t_h dans la requête q et n_t est le nombre de termes d'indexation. Cette formulation suppose implicitement que les corrélations entre les termes sont nulles. D'autres mesures de similarité sont proposées dans (Van Rijsbergen, 1979) et (Salton *et al.,* 1983). Les mesures les plus connues sont la mesure de similarité angulaire ('*cosine similarity*'), la mesure Jaccard et la mesure Dice.

4.3. *Opérateurs de la logique booléenne (Modèle booléen)*

Selon (Van Rijsbergen, 1986), la RI peut être assimilée à une *implication logique*. Cette vision de la RI est celle qui est à la base du *modèle booléen*. C'est un modèle élémentaire mais utile en RI car il est simple et intuitif. Dans ce modèle, les documents ainsi que les requêtes sont exprimés selon le formalisme de la logique de Boole. Un document d est alors pertinent pour une requête q lorsqu'il vérifie la formule $d \rightarrow q$. La pertinence est donc assimilée à un prédicat logique : un document est soit pertinent, soit non pertinent, ce qui ne permet aucune discrimination entre les documents pertinents. Par ailleurs, il s'avère difficile de tenir compte de l'incertitude dans ce modèle.

4.4. *Normes L_p (Modèle booléen étendu)*

Cette famille de mesures est principalement utilisée dans le cadre du *modèle booléen étendu* qui a été introduit par (Salton *et al.,* 1983) pour pallier aux inconvénients du modèle booléen standard. Dans ce modèle, selon que la requête q est disjonctive ou conjonctive, l'on utilise respectivement la première ou la deuxième formulation de la norme L_p de Minlowski-Hölder pour mesurer la pertinence d'un document d_i :

$$rsv(d_i, q) = \left(\frac{\sum_{t_h \in q} (w_{i,h})^p}{n_q} \right)^{\frac{1}{p}} ; \quad rsv(d_i, q) = 1 - \left(\frac{\sum_{t_h \in q} (1 - w_{i,h})^p}{n_q} \right)^{\frac{1}{p}}$$

où n_q est le nombre des termes de la requête q.

Les normes L_p les plus connues sont la norme rectilinéaire (L_1), la norme euclidienne (L_2) et la norme de Tchebychev (L_∞).

4.5. *Opérateurs flous*

Ce genre d'opérateurs a été initialement utilisé dans le cadre du *modèle des ensembles flous* dans lequel chaque document d_i est représenté par un sous-ensemble

flou $\{(t_h,\, w_{i,h}),\ h = 1, \ldots, n_t\}$ qui donne pour chaque terme t_h son degré d'appartenance au document d_i. À partir de cette représentation, nous pouvons associer à chaque terme t_h le sous-ensemble flou $\{(d_i,\, w_{i,h}),\ \forall\, d_i \in D\}$ résumant ses liens sémantiques aux documents de la collection D. La réponse à une requête q est obtenue en appliquant les opérateurs flous classiques sur les sous-ensembles flous des différents termes de la requête. Formellement, en considérant la forme normale disjonctive d'une requête q, c'est-à-dire $q = (q_1 \vee q_2 \ldots \vee q_l)$ où q_k est une composante conjonctive, la pertinence d'un document d_i à q est donnée par :

$$rsv(d_i, q) = \perp_{k=1,\ldots l} (\top_{t_h \in q_k} (w_{i,h}))$$

où $\perp$ et $\top$ sont des opérateurs d'agrégation disjonctif et conjonctif flous, respectivement, tels que les opérateurs $\max$ et $\min$ (Zadeh, 1965).

D'autres opérateurs d'agrégation flous peuvent également être utilisés pour combiner les poids des termes. La famille d'opérateurs *OWA* ('*Ordered Weighted Averaging*') de (Yager, 1988) a été utilisée par (Bordogna *et al.,* 1997). De même (Boughanem *et al.,* 2005) utilisent les opérateurs *leximin* et *discrimin* de (Dubois *et al.,* 1997) pour l'agrégation des performances des documents.

4.6. *Autres modèles*

Il existe d'autres modèles de rangement issus des travaux en apprentissage dans le domaine de l'intelligence artificielle. Il s'agit notamment du modèle neuronal et du modèle inférentiel. Dans le modèle neuronal, c'est le niveau d'activation final des noeuds documents qui donne la mesure de pertinence utilisée dans le rangement final (Kwok, 1995). Dans le modèle inférentiel, la pertinence d'un document à une requête correspond au degré de croyance que l'observation du document va satisfaire le besoin de l'utilisateur émanant la requête (Turtle *et al.,* 1990).

5. Caractéristiques des mécanismes d'agrégation de chaque niveau

Les formules utilisées à chaque niveau d'agrégation sont en général des formules analytiques traduisant des logiques d'agrégation de deux types : une logique d'agrégation totalement compensatoire et une logique d'agrégation non compensatoire.

5.1. *Formules d'agrégation totalement compensatoires*

Les formules d'agrégation totalement compensatoires se distinguent principalement par les caractéristiques suivantes :

– Elles sont parfois complexes et difficiles à interpréter (voir le modèle BM25).

– Elles agrègent des éléments incommensurables, par exemple tf, idf et dl.

– Elles autorisent que de très faibles performances sur certains termes soient compensées par de bonnes performances sur d'autres. Des documents ayant des *configurations* très contrastées peuvent ainsi recevoir la même évaluation et être jugés équivalents. À titre d'illustration, selon le tableau suivant donnant les poids $w_{i,h}$ de deux

termes t_1 et t_2 et en supposant que la mesure de pertinence est donnée par la moyenne des poids des différents termes, les documents d_1 et d_2 sont jugés équivalents.

	t_1	t_2
d_1	0,5	0,5
d_2	0,1	0,9

– Elles sont sensibles à de faibles variations numériques. Nous pouvons ainsi avoir des inversions de préférence suite à des changements minimes des performances. Par exemple, en supposant les données du tableau précédent et que $w_{1,1}$ passe de 0,5 à 0,51, le document d_1 devient strictement meilleur que d_2 alors que lorsque $w_{1,1}$ devient 0,49, c'est d_2 qui devient strictement meilleur.

– Elles nécessitent parfois la détermination d'un jeu de valeurs de pondération qui est souvent entaché d'un *arbitraire* inévitable.

– Elles ne permettent pas de tenir compte de la part d'arbitraire inhérente à la construction des poids des termes et du fait que certaines différences de performances sont négligeables. Par exemple, le critère de fréquence peut avoir les expressions suivantes : $tf_{i,h}$, $\frac{tf_{i,h}}{max\_tf_i}$. Ainsi, dans l'exemple suivant, le document d_1 est jugé meilleur que d_2 selon le critère tf. La situation est inversée lorsque le critère $\frac{tf}{max\_tf}$ est retenu.

	$tf_{i,h}$	max_tf_i	$\frac{tf_{i,h}}{max\_tf_i}$
d_1	5	30	0,167
d_2	4	20	0,200

5.2. *Formules d'agrégation non compensatoires*

Utiliser des opérateurs d'agrégation non compensatoires implique souvent de juger un document sur la base de sa meilleure ou mauvaise performance, ce qui conduit à une perte d'information non négligeable. En effet, un seul terme est principalement considéré pour chaque document. Les autres termes servent éventuellement à départager les ex-aequos. Par ailleurs, on ne tient pas compte de l'insignifiance des faibles différences de performances. À titre d'illustration, considérons les données du tableau suivant donnant les poids des termes d'une requête et supposons que le score *rsv* est obtenu en utilisant l'opérateur $\min$. Dans ce cas, le document d_1 est jugé meilleur que d_2 bien qu'il soit plus judicieux de préférer d_2 puisque la différence de performances entre 0,5 et 0,49 est négligeable.

	t_1	t_2	t_3	t_4
d_1	0,5	0,5	0,5	0,5
d_2	0,49	0,9	0,9	0,9

Dans cette catégorie, nous trouvons principalement les opérateurs d'agrégation $\min$, $\max$ ainsi que l'opérateur d'agrégation lexicographique.

6. Une nouvelle famille de modèles de RI

Les approches classiques de RI agrègent les différentes sources de pertinence *d'abord au niveau des termes, puis au niveau de la requête*. Ainsi, en désignant par $g_j(d, t_h)$ la performance d'un document d selon un critère g_j mesurée par rapport à

un terme t_h de la requête, sur le tableau 1, la première phase consiste à agréger les données de chaque colonne, conduisant au calcul des poids w_h, et la deuxième phase consiste à agréger les éléments de la dernière ligne, conduisant au calcul du *rsv*.

d	t_1	t_2	$\dots$	t_{n_q}	
g_1	$g_1(d,t_1)$	$g_1(d,t_2)$	$\dots$	$g_1(d,t_{n_q})$	
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	
g_{n_f}	$g_{n_f}(d,t_1)$	$g_{n_f}(d,t_2)$	$\dots$	$g_{n_f}(d,t_{n_q})$	
w_h	w_1	w_2	$\dots$	w_{n_q}	rsv

Tableau 1. *Données de base pour le calcul du score de d*

Cette *agrégation en deux phases* est inévitable lorsque la requête se compose de plusieurs termes. Toutefois, l'ordre classique de ces deux phases d'agrégation consiste à agréger d'abord des données plutôt hétérogènes. Or en inversant l'ordre d'agrégation, c'est-à-dire en agrégeant d'abord au niveau de la requête, puis au niveau des termes, l'utilisation d'opérateurs analytiques en première phase est naturelle puisqu'on agrège des données homogènes. Il convient ensuite, en deuxième phase, d'agréger des données hétérogènes pour lesquelles des mécanismes d'agrégation plus adéquats peuvent être utilisés.

Le modèle multicritère de RI de (Farah *et al.,* 2006) s'inscrit dans cette nouvelle perspective. Dans ce modèle, nous distinguons deux phases essentielles qui sont la *modélisation de la pertinence* et l'*agrégation relationnelle des performances*.

La phase de modélisation de la pertinence consiste à développer les critères de pertinence en se basant sur les différents facteurs qui ont un impact sur la définition de la pertinence des documents. Sur le tableau 1, cela consiste à agréger d'abord les valeurs de chaque ligne. À titre d'exemple, considérant l'aspect relatif à la fréquence des termes de la requête dans le document, il s'agit d'attribuer au document une seule évaluation qui résume sa performance par rapport à cet aspect, en utilisant un opérateur d'agrégation analytique classique tel que la moyenne.

La phase d'agrégation consiste à utiliser des mécanismes d'agrégation relationnelle pour échapper aux limites des approches d'agrégation analytiques lorsqu'elles sont appliquées à des données hétérogènes et lorsque les valeurs à agréger sont entachées d'arbitraire. Dans ce modèle multicritère, le rangement est obtenu en deux étapes. Dans la première étape, on utilise des règles de décision simples dans des comparaisons par paires des documents à ranger. Ceci permet de juger si un document est globalement plus ou moins pertinent qu'un autre selon les performances des deux documents sur les critères retenus. Plusieurs relations binaires peuvent être construites selon le niveau d'exigence des règles de décision. Par exemple, une règle de décision simple consiste à accepter qu'un document d est globalement '*au moins aussi bon que*' d' lorsque $g_j(d) \geq g_j(d') - q_j, \forall j$, où q_j est un *seuil d'indifférence* permettant de modéliser l'arbitraire entachant la construction du critère g_j ainsi que l'imprécision de l'évaluation de la performance d'un document selon ce critère. La deuxième étape consiste à exploiter ces relations binaires pour construire un rangement final des

documents. Par exemple, une procédure élémentaire consiste à calculer pour chaque document d la différence entre les documents qui lui sont moins bons et ceux qui lui sont meilleurs. Les documents peuvent ainsi être rangés selon ce flux net.

Les premières expérimentations ont été menées dans le contexte TREC. Les résultats reportés dans (Farah *et al.,* 2006) concernent les requêtes TD ('*Topic Distillation*') de la tâche Web de TREC'2004 (Craswell *et al.,* 2004). Le tableau 2 compare les performances de l'approches multicritère (`mcm`) par rapport à d'autres stratégies d'agrégation classiques, sur la base des mesures de performances standards. On observe que les performances des stratégies d'agrégation classiques basées sur des logiques totalement compensatoires (`sum` et `prod`) ou non compensatoires (`max` et `min`) sont moins bonnes que celles des procédures d'agrégation basées sur des logiques partiellement compensatoires (`mcm`) utilisant des approches multicritères.

Run Id	AvP	R-p	r-r	S@1	S@5	S@10	Δ-AvP
mcm	17,08%	18,37%	58,04%	45,33%	74,67%	81,33%	–
max	8,02%	7,70%	21,40%	8,00%	33,33%	50,67%	-53,02%*
min	10,74%	12,91%	47,20%	32,00%	70,67%	77,33%	-37,13%*
prod	12,06%	14,02%	53,66%	37,33%	74,67%	80,00%	-29,41%*
sum	13,45%	14,37%	51,78%	36,00%	66,67%	82,67%	-20,73%*

Tableau 2. *Comparaison de différentes stratégies d'agrégation*

7. Conclusion

Dans ce papier, nous avons donné une nouvelle présentation des modèles théoriques de RI en distinguant la phase de construction des sources de pertinence de la phase de l'établissement du rangement final des documents. L'agrégation étant au coeur de ces modèles, nous avons présenté les principales caractéristiques des mécanismes d'agrégation utilisés dans ces modèles tout en mettant l'accent sur leurs limites. Nous avons aussi présenté une nouvelle famille de méthodes de RI qui agrègent les sources de pertinence de façon plus adéquate.

8. Bibliographie

Baeza-Yates R. A., Ribeiro-Neto B. A., *Modern Information Retrieval*, ACM Press , 1999.

Bordogna G., Pasi G., « Applications of OWA Operators to Information Retrieval », *in* , R. Yager, , J. Kacprzyk (eds), *The Ordered Weighted Averaging Operators – Theory and Applications*, Kluwer Academic Publ., Boston, p. 275-292, 1997.

Boughanem M., Loiseau Y., Prade H., « Rank-ordering documents according to their relevance in information retrieval using refinements of ordered-weighted aggregations », *Proceedings of AMR'05*, Springer-Verlag, p. 44-54, 2005.

Brin S., Page L., « The anatomy of a large-scale hypertextual Web search engine », *Proceedings of WWW'98*, Elsevier Science Publishers, p. 107-117, 1998.

Craswell N., Hawking D., « Overview of the TREC-2004 Web Track », *Proceedings of TREC'2004*, NIST Publication, 2004.

Dubois D., Fargier H., Prade H., « Beyond min aggregation in multicriteria decision : (Ordered) weighted min, discri-min, leximin », *in* , R. Yager, , J. Kacprzyk (eds), *The Ordered Weighted Averaging Operators – Theory and Applications*, Kluwer Academic Publ., Boston, p. 181-192, 1997.

Farah M., Vanderpooten D., « A Multiple Criteria Approach for Information Retrieval », *Proceedings of SPIRE'06*, LNCS 4209, Springer-Verlag, p. 242-254, 2006.

Fuhr N., Buckley C., « A probabilistic learning approach for document indexing », *ACM Trans. Inf. Syst.*, vol. 9, n° 3, p. 223-248, 1991.

Gao G., Nie J.-Y., Bai J., « Integrating word relationships into language models », *Proceedings of ACM-SIGIR'05*, ACM Press, p. 298-305, 2005.

Kleinberg J. M., « Authoritative sources in a hyperlinked environment », *Journal of ACM*, vol. 46, n° 5, p. 604-632, 1999.

Kwok K. L., « A network approach to probabilistic information retrieval », *ACM Trans. Inf. Syst.*, vol. 13, n° 3, p. 324-353, 1995.

Liu X., Croft W. B., « Cluster-based retrieval using language models », *Proceedings of ACM-SIGIR'2004*, ACM Press, p. 186-193, 2004.

Ogawa Y., Morita T., Kobayashi K., « A fuzzy document retrieval system using the keyword connexion matrix and a learning method », *Fuzzy sets and systems*, vol. 39, p. 163-179, 1991.

Ponte J. M., Croft W. B., « A language modeling approach to information retrieval », *Proceedings of ACM-SIGIR'98*, ACM Press, p. 275-281, 1998.

Robertson S. E., « The probability ranking principle in information retrieval », *Journal of Documentation*, vol. 33, n° 4, p. 294-304, 1977.

Robertson S. E., Spark Jones K., « Relevance weighting of search terms », *JASIS*, vol. 27, n° 3, p. 129-146, 1976.

Robertson S. E., Walker S., Jones S., Hancock-Beaulieu M., Gatford M., « Okapi at TREC-3 », *Proceedings of TREC'3*, NIST Publication, 1994.

Salton G., *Automatic Information organization and retrieval*, McGraw-Hill, 1968.

Salton G., Fox E. A., Wu H., « Extended Boolean information retrieval », *Commun. ACM*, vol. 26, n° 11, p. 1022-1036, 1983.

Salton G., McGill M. J., *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

Savoy J., Picard J., « Retrieval effectiveness on the web », *IPM*, vol. 37, n° 4, p. 543-569, 2001.

Sparck Jones K., « A statistical interpretation of term specificity and its application in retrieval », *Journal of Documentation*, vol. 28, p. 11-21, 1972.

Turtle H., Croft W. B., « Inference networks for document retrieval », *Proceedings of ACM-SIGIR'90*, ACM Press, p. 1-24, 1990.

Van Rijsbergen C. J., *Information Retrieval*, Second edn, Dept. of Computer Science, University of Glasgow, London : Butterworths, 1979.

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

Yager R. R., « On ordered weighted averaging aggregation operators in multicriteria decision making », *IEEE Trans. Syst. Man Cybern.*, vol. 18, n° 1, p. 183-190, 1988.

Zadeh L. A., « Fuzzy sets », *Information and Control*, vol. 8, n° 3, p. 338-353, June, 1965.