Un modèle de RI basé sur des critères d'obligation et de certitude

Leïla Kefi* — Catherine Berrut* — Eric Gaussier**

* *Laboratoire CLIPS-IMAG*
385, rue de la Bibliothèque - BP5, 38041Grenoble cedex 9, France
{leila.kefi, catherine.berrut}@imag.fr

** *Xerox Research Center Europe*
6 chemin de Maupertuis, 38240 Meylan, France
Eric.gaussier@xrce.xerox.com

RÉSUMÉ *Il existe un grand nombre de modèles de recherche d'information chacun ayant pour but de répondre au mieux aux attentes des utilisateurs. Le modèle que nous proposons se base sur une formulation précise de la requête reflétant le besoin de l'utilisateur : Chaque terme de la requête est augmenté par deux critères, l'un exprimant l'obligation ou non de l'apparition du terme dans les documents et l'autre exprimant la certitude de l'utilisateur quand au terme utilisé. Des expérimentations nous ont permis de vérifier qu'une telle formulation permet de gagner en précision.*

ABSTRACT. *There is a large number of IR models each one having for goal to answer users needs as well as possible. The model we propose bases on a precise query formulation reflecting the user's need: Each term of the query is increased by two criteria, one expressing the obligation or not that the term appears in the documents and the other expressing the user certainty about the term he used. Experiments enabled us to check that such a formulation increase the system precision.*

MOTS-CLÉS : *modèle de RI, formulation de la requête, obligation et certitude.*

KEYWORDS: *IR Model, query formulation, obligation and certainty.*

1. Introduction

Il existe un grand le nombre de modèles de recherche d'information ; Leur principale différence réside dans la façon dont les documents disponibles et le besoin en information de l'utilisateur sont représentés et mis en correspondance [VanR79]. Dans la plupart des modèles existants, la requête est représentée sous forme d'un ensemble de termes pondérés (selon le modèle considéré) et de ce fait, ils ne permettent pas à l'utilisateur de préciser clairement son besoin et, retournent une masse importante de documents pas tous pertinents. Pourtant dans certains domaines, notamment le domaine professionnel, les utilisateurs ont besoin d'une information précise, réponse à un besoin d'information précis, d'où le besoin d'avoir un modèle orienté précision, autant au niveau de l'expression de la requête qu'au niveau des réponses retournées par le système.

Notre objectif est de donner à l'utilisateur la possibilité de façonner sa requête, le plus facilement possible et avec le plus de précision possible. Nous nous intéressons donc à l'ajout de certains critères sur les termes de la requête afin d'augmenter l'expressivité du système de recherche. En nous inspirant de certains modèles augmentant les termes de la requête par un critère d'obligation/option, nous proposons un modèle qui permet aussi l'ajout d'un critère de certitude/incertitude et qui permet l'utilisation multiple des termes au niveau de l'index et de la requête. L'utilisateur peut ainsi formuler sa requête en fonction de ce qu'il veut exactement retrouver dans les documents. Des expérimentations sur un corpus technique nous permettent de nous positionner par rapport aux modèles booléen et vectoriel.

2. Motivations

Dans un but de formulation précise mais néanmoins aisée de la requête, nous nous intéressons à certains points que nous pensons être utiles pour son expressivité.

Notion d'obligation

Un terme marqué comme obligatoire dans une requête doit absolument apparaître dans les documents retrouvés, alors qu'un terme optionnel peut y apparaître ou non. Cette notion d'obligation n'est pas nouvelle : Le moteur de recherche Altavista, par exemple, a utilisé un tel critère (préfixe "+") afin de fournir une syntaxe plus simple et plus intuitive, permettant ainsi de résoudre la difficulté rencontrée par les utilisateurs pour exprimer des requêtes booléennes. [Denos97-97a] a aussi proposé d'associer aux termes de la requête un critère d'obligation/option dans le but de regrouper les documents pertinents dans des classes par rapport aux termes optionnels qu'ils contiennent. L'utilisateur peut ainsi mieux comprendre la relation entre sa requête et les documents retrouvés ce qui aide à sa reformulation.

Notion de certitude

Un terme marqué comme certain dans une requête doit absolument apparaître dans les documents retrouvés tel qu'il a été mentionné dans la requête, alors qu'un terme incertain peut y apparaître sous une forme "proche" (ex. synonyme). L'utilisateur qui sait exactement ce qu'il recherche (ce qu'il veut trouver dans les documents) marquera le(s) terme(s) de la requête dont il est sûr comme certain(s) et l'utilisateur qui hésite quand aux termes à utiliser pour exprimer sa requête les marquera comme incertains. Cette distinction paraît encore plus utile dans les milieux professionnels utilisant des termes spécialisés et précis concernant des données connues des utilisateurs.

Utilisation multiple d'un terme dans l'index et la requête

Généralement, dans les approches classiques, un élément contenu dans un document plusieurs fois n'est représenté qu'une seule fois dans le document indexé. Ainsi, une première image représentant un bateau et une seconde en représentant deux, seront toutes deux indexées par un terme "bateau". Pourtant, dans certains contextes, il est plus pertinent d'utiliser le terme autant de fois que nécessaire pour une meilleure précision du système. (Et pareillement pour la requête).

3. Description formelle du modèle proposé

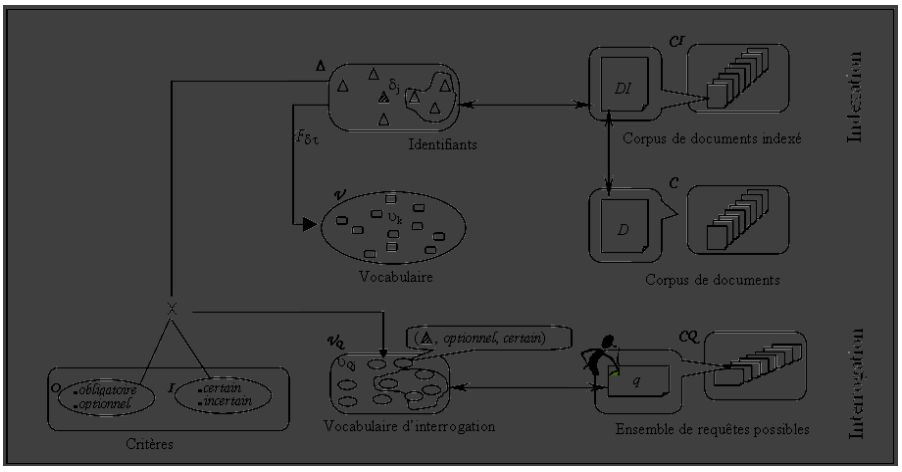

Figure 1. *Vue globale du modèle proposé*

Soient un corpus C de documents *Di* et un vocabulaire $\mathcal{V}$ défini par un ensemble de termes τ_k (nous désignons par terme tout élément d'indexation : un mot clé, un concept, un objet graphique, etc.) Une relation de proximité $\mathcal{P}roche$ (pour l'incertitude) relie certains termes de $\mathcal{V}$. Elle peut se présenter selon différentes formes (synonymie, spécialisation/généralisation, etc.)

Afin de permettre l'utilisation multiple de termes, nous définissons un ensemble d'identifiants rattaché au vocabulaire et qui servira de base pour l'indexation et la formulation de la requête : $\Delta = \{ \delta_1,\ldots, \delta_k,\ldots, \delta_{\mathcal{N}\Delta} \}$. A chaque identifiant correspond un terme de $\mathcal{V}$ selon la fonction $\mathcal{F}_{\delta\tau}$: $\mathcal{F}_{\delta\tau}(\delta_k) = \tau_j \in \mathcal{V}$

3.1. Le document indexé et la requête

Un document indexé DI rattaché au document D est un ensemble d'identifiants : $DI \in \mathcal{P}(\Delta)$ et $DI = \{\delta_{D.1} ..., \delta_{D.j}, \ldots, \delta_{D.\mathcal{N}D}\}$

Afin de permettre l'expression des critères d'obligation et de certitude, nous introduisons deux ensembles $\mathcal{O}=\{obligatoire,optionnel\}$ et $\mathcal{I}=\{certain,incertain\}$ et définissons le vocabulaire d'interrogation: $\mathcal{V}_Q = \Delta \times \mathcal{O} \times \mathcal{I}$. Trois fonctions permettent d'obtenir pour chaque triplet de $\mathcal{V}_Q$, l'identifiant ($\mathcal{F}_\delta$), le critère d'obligation ($\mathcal{F}_{Ob}$) et le critère de certitude ($\mathcal{F}_{Cer}$) :

$\mathcal{F}_\delta$: $\mathcal{V}_Q \rightarrow \mathcal{V}$ $(\delta, o, c) \rightarrow \delta$ $\quad$ $\mathcal{F}_{Ob}$: $\mathcal{V}_Q \rightarrow \mathcal{O}$ $(\delta, o, c) \rightarrow o$ $\quad$ $\mathcal{F}_{Cer}$: $\mathcal{V}_Q \rightarrow \mathcal{I}$ $(\delta, o, c) \rightarrow c$

Une requête est un sous-ensemble du vocabulaire d'interrogation ($q \in \mathcal{P}(\mathcal{V}_Q)$):

$$q = \{\delta_{q.1},\ldots, \delta_{q.j},\ldots, \delta_{q.\mathcal{N}q}\} \text{ et } \delta_{q.j} = (\mathcal{F}_\delta(\delta_{q.j}), \mathcal{F}_{Ob}(\delta_{q.j}), \mathcal{F}_{Cer}(\delta_{q.j}))$$

On peut aussi écrire: $q = q_{ObCer} \cup q_{ObInc} \cup q_{OpCer} \cup q_{OpInc}$, où :

$q_{ObCer} = \{\delta_q = \mathcal{F}_\delta(\delta_{q.j}) / \mathcal{F}_{ob}(\delta_{q.j}) = obligatoire$ et $\mathcal{F}_{cer}(\delta_{q.j}) = certain\}$, et ainsi de suite…

3.2. La correspondance entre la requête et les documents

Dans une requête q, certains identifiants référencent des termes obligatoires et d'autres référencent des termes optionnels. Un document D est pertinent si:

- A chaque identifiant obligatoire de q correspond un identifiant de DI: les deux référencent le même terme ou des termes proches (dépend du critère de certitude).

- A <u>certains</u> identifiants optionnels de q correspondent des identifiants de DI. Nous considérons deux sous-ensembles: q'_{OpCer} (contenant les identifiants optionnels et certains ayant un correspondant dans DI) et q'_{OpInc} (contenant les identifiants optionnels et incertains ayant un correspondant dans DI).

De manière similaire, DI va contenir des identifiants concernés par q (correspondant à des identifiants de q) et des identifiants non concernés. Nous considérons donc le sous-ensemble de DI qui contient les identifiants correspondant à des identifiants de q. Nous notons cet ensemble DI'.

La correspondance est caractérisée par une fonction $\mathcal{F}_{Corresp}$ qui, à chaque identifiant δ dans $q_{ObCer} \cup q_{OpInc} \cup q'_{OpCer} \cup q'_{OpInc}$ associe un identifiant unique δ' dans *DI'* vérifiant le critère de certitude.

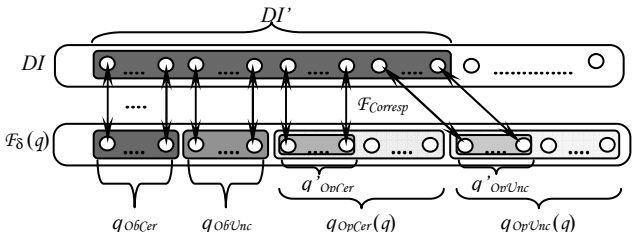

Figure 2. *Correspondance un à un entre identifiants de DI et de q*

Ainsi, un document *D* répond à une requête *q* si :

$\exists\ \Delta'_{OpCer} \subseteq \Delta_{OpCer}(q)$, $\Delta'_{OpInc} \subseteq \Delta_{OpInc}(q)$ et $DI' \subseteq DI$ tels que : $\exists$ une fonction bijective $\mathcal{F}_{Corresp}$ définie :

$$\mathcal{F}_{Corresp} : \Delta_{ObCer} \cup \Delta_{OpInc} \cup \Delta'_{OpCer} \cup \Delta'_{OpInc} \rightarrow DI'$$

$$\delta \rightarrow \delta' \text{ telle que :}$$

$$\begin{cases} \text{Si } \delta \in \Delta_{ObCer} \cup \Delta'_{OpCer},\ \mathcal{F}_{\delta\tau}(\delta') = \mathcal{F}_{\delta\tau}(\delta) \\ \text{Si } \delta \in \Delta_{ObInc} \cup \Delta'_{OpInc},\ \mathcal{F}_{\delta\tau}(\delta') = \mathcal{F}_{\delta\tau}(\delta) \text{ ou } \mathcal{F}_{\delta\tau}(\delta') \in \mathcal{P}roche_{\mathcal{T}}(\mathcal{F}_{\delta\tau}(\delta)) \end{cases}$$

4. Expérimentations

Les expérimentations ont été menées selon la méthodologie standard de RI dans laquelle une collection de test constituée de : (i) 1034 graphiques commentés extraits de 12 manuels techniques représentés par une liste de termes extraits par un processus d'indexation automatique [kefi05], (ii) 25 requêtes et (iii)des jugements de pertinence pour chaque requête, est utilisée afin de calculer la précision et le rappel.

- Une requête qui représentée dans notre modèle ne contient que des termes optionnelles n'est pas réaliste. Nous posons donc la contrainte qu'au moins un terme de la requête soit obligatoire.

- Pour chacun des modèles vectoriel et booléen, nous considérons deux cas : « enrichi » ou « simple » selon qu'il utilise ou non la fonction $\mathcal{P}roche$ définie dans notre modèle pour enrichir la requête.

- Afin d'ordonner les documents retournés par notre modèle, nous utilisons une variante de la fonction de similarité définie dans [Salton83]:

$$S(q, DI) = 1 - \sqrt{\sum_{\delta_{D.j} = \mathcal{F}^{-1}_{correp}(\delta_{q.k})} (1 - w_{D.j})^2 \Big/ \sum_{\delta_{D.j} = \mathcal{F}^{-1}_{correp}(\delta_{q.k})} 1}$$

($w_{D.j}$: poids du j^{ième} terme de D)

Nous présentons les valeurs moyennes de précision pour chacun des modèles ainsi que les valeurs de la précision à 5 documents (voir table 1). La conclusion que nous pouvons tirer à partir de ces résultats, est que tenir compte de critères d'obligation et de certitude est bénéfique pour les performances de recherche. Les résultats obtenus par notre système sont largement supérieurs à ceux des modèles

vectoriel et booléen (enrichis ou non). Une extension de ces modèles permettant de tenir compte de ces critères pourrait être envisagée. Il s'agirait, dans ce cas, de considérer deux sous-requêtes l'une portant sur les termes obligatoires et la l'autre sur les optionnels, combinées afin de donner un score global tel que : $Sim(q,D)= \alpha Sim(q_{ob},D)+(1-\alpha)Sim(q_{op},D)$. α étant un paramètre additionnel qu'il faut fixer pour chaque collection. Ceci nous permet de remarquer que notre modèle restera toujours une solution est plus directe qu'une extension des modèles existants.

Modèles / Précision	Modèle proposé	Vectoriel simple	Vectoriel enrichi	Booléen simple	Booléen enrichi
Moyenne	0,608	0,427	0,386	0,320	0,345
à 5 docs	0,808	0,688	0,496	0,458	0,416

Table 1. *Valeurs de la précision moyenne* et de la précision à 5 documents

5. Conclusion

Le modèle que nous proposons offre à l'utilisateur la possibilité d'exprimer ses besoins en fonction de ce qu'il souhaite voir absolument apparaître ou non dans les documents et en fonction de sa certitude quand à ce qu'il désire voir dans ces documents. Nous avons pu constater que ce modèle permet d'atteindre les meilleures valeurs de précision, ce qui est un atout certain pour certaines applications telles que celles dédiées à des professionnels.

Nous pensons que ce modèle est particulièrement bien adapté à la documentation technique, dans laquelle la présence de graphiques rend l'utilisation des critères d'obligation et de certitude encore plus utile en raison de la nature visuelle de cette information.

6. Bibliographie

[Denos97] N. Denos, Modélisation de la pertinence en recherche d'information - modèle conceptuel, formalisation et application, Ph.D. thesis, Université Joseph Fourier, 1997.

[Denos97a]N. Denos, Modelling system relevance through user criteria - A conceptual and a formal model, Rapport de recherche, Équipe Mrim, CLIPS-IMAG, TR-97-001, 1997.

[kefi05]L.Kefi, C. Berrut, E. Gaussier, Indexation Complexe de documents: vers une vérification qualitative, in Inforsid, Grenoble, pp521-538, 24-27 mai, 2005

[VanR79] C.J. van Rijsbergen. Information Retrieval, 2nd edition. Dept of Computer science, University of Glasgow, 1979

[Salton83] G. Salton, E. A. Fox, and H.Wu. *Extended Boolean information retrieval*. Communications of the ACM, 26(12):1022-1036, December 1983.