Contribution à la recherche d'information Une fonction de correspondance

Fonction de Correspondance

Fatou Kamara-Sangaré

Catégorie Jeunes Chercheurs
Laboratoire d'Analyse Numérique et d'Informatique (LANI)
UFR de Sciences Appliquées et de Technologie
Université Gaston Berger de Saint-Louis
Bp 234 Saint-Louis Sénégal

RÉSUMÉ. *Un Système de Recherche d'Information (SRI) dispose d'un modèle de recherche capable de déterminer le degré de similarité qui existe entre un document et une requête. Généralement, le mécanisme consiste à apparier les documents et la requête en utilisant une fonction de correspondance. Dans ce papier, nous proposons la définition d'une fonction de correspondance qui repose sur les termes contenus uniquement dans l'intersection de la requête et d'un document. Afin de montrer sa performance, nous utilisons les critères suivants : le taux de rappel, le taux de précision et la complexité temporelle.*

ABSTRACT. *An Information Retrieval System uses a retrieval model which is able to compute the degree of similarity between a document and an user query. A technique consists in match document and user query using a similarity function.*
In this paper we propose a new similarity function using only terms in intersection between documents and an user query. For an evaluation's performance, we compute precision ratio, recall ratio and run time which permits to retrieve documents matching user query.

MOTS-CLÉS : *Recherche d'Information, Fonction de Correspondance*

KEYWORDS: *Information Retrieval, Similarity Function.*

1. Introduction

Un Système de Recherche d'Information (SRI) permet de retrouver les documents pertinents par rapport aux besoins exprimés par l'utilisateur à travers une requête, à partir d'une base de documents. Avec la masse d'informations disponibles et les nouvelles technologies, beaucoup de contributions ont été proposées. Généralement, ces contributions [BAE 99, IHA 04a, IHA 04b] diffèrent sur la base de documents utilisée et sur les mécanismes de recherche appliqués à la base. Ainsi, un SRI est un système d'information qui permet l'accès à un ensemble de documents par leur contenu sémantique. Cet accès est le résultat d'une recherche consistant à définir un modèle capable de calculer ou de mesurer le degré de similarité qui existe entre les documents ou entre un document et une requête [BAE 99, SAL 71]. Les modèles les plus utilisés sont les modèles vectoriel, booléen et probabiliste.
La performance d'un modèle dépend d'une part de la base de documents utilisée et d'autre part de la fonction de correspondance appliquée à la base pour déterminer la similarité entre documents et requête. Dans ce présent travail, nous proposons une fonction de correspondance. Pour prouver sa performance, nous l'appliquons à une recherche sur une base de composants logiciels réutilisables. Un des problèmes liés à la réutilisation de composants [OUS 05] est la recherche. Comment retrouver rapidement et simplement un composant réutilisable pour un problème donné. Les termes **simplement** et **rapidement** ont un sens, puisqu'on ne doit pas passer plus de temps à retrouver un composant qu'on aurait mis à le développer. Un premier travail sur la recherche de composants logiciels a été fait dans le cadre d'une convention de recherche avec l'Institut Français du Pétrole (IFP, Pau) et PsRep (Pau Software Repository) [HOC 02]. L'environnement qui a été développé est un outil logiciel qui permet de cataloguer des composants logiciels et d'accéder à ces composants et leur description. Cet outil prend en compte les composants disponibles sous la forme de bibliothèques, de sous programmes, de procédures, de fonctions ou de classes. Notre objectif vise à améliorer certaines de ces fonctionnalités.
La suite de l'article s'organise comme suit. Dans la section 2, nous définissons la fonction de correspondance proposée et ses propriétés. Nous terminons dans la section 3 par une Application.

2. Définition et propriétés de la fonction proposée

Dans cette section nous décrivons quelques définitions et notations que nous utiliserons pour définir notre fonction et ses propriétés.

2.1. *Contexte*

Considérons les notations suivantes :

– D et T représentent respectivement le nombre de documents et le nombre de termes d'index ;

– B_{doc} et $base_{index}$ représentent respectivement la base des documents et l'ensemble des termes d'index ;

– B_{index} est constitué des transformées des documents en fonction des termes d'index ;

– d appartient à B_{doc} entraîne d est un document ;

– $\bar{d}$ appartient à B_{index} entraîne $\bar{d} = \{t_j\}_{\{0<j\leq T\}}$, $\bar{d}$ est un ensemble de termes d'index décrivant le contenu sémantique du document.

– Q et $\bar{Q}$ représentent respectivement la requête de l'utilisateur en langage naturel et sa transformée en termes d'index.

– Le poids du j^{e} terme dans le i^{e} document détermine l'importance de ce terme dans le document. Il est quantifié par un scalaire w_{ji} positif tel que :
$w_{ji} = tf_{ij} \times idf_j$ si le j^{e} terme est dans le i^{e} document 0 sinon.
Où tf_{ij} est la fréquence d'apparitions du j^{e} terme dans le i^{e} document et idf_j est l'inverse de la fréquence du document du terme indexé t_j.

– Partant de l'idée que la requête de l'utilisateur exprime son besoin informationnel, alors tout terme d'index appartenant à la transformée de la requête $\bar{Q}$ est important. Cela implique que tous les termes de $\bar{Q}$ ont le même poids qui est une constante q défini par : $q = min(idf_j)_{t_j \in base_{index}}$ si le j^{e} terme est dans $\bar{Q}$.

Définition 2.1 *La fonction cosinus [BAE 99] est la fonction la plus utilisée pour mesurer la similarité entre un document et une requête, elle est défi nie par :*

$$\begin{array}{rccl} cosinus: & B_{index} \times \{\bar{Q}\} & \longrightarrow & R_+ \\ & (\bar{d}_i\ ,\ \bar{Q}) & \longmapsto & \dfrac{\sum\limits_{t_j \in \bar{d}_i \cap \bar{Q}} w_{ji}\, q_j}{\left(\sum\limits_{t_j \in \bar{d}_i} w_{ji}^2\right)^{1/2} \left(\sum\limits_{t_j \in \bar{Q}} q_j^2\right)^{1/2}}. \end{array} \qquad [1]$$

2.2. *Définition de la fonction de correspondance et ses propriétés*

Définition 2.2 *La fonction de correspondance proposée est une variante de la fonction cosinus (2.1). Elle utilise uniquement les poids des termes dans les documents. Elle est défi nie par :*

$$\begin{array}{rccl} simis: & B_{index} \times \{\bar{Q}\} & \longrightarrow & R_+ \\ & (\bar{d}_i\ ,\ \bar{Q}) & \longmapsto & \dfrac{\sum\limits_{t_j \in \bar{d}_i \cap \bar{Q}} w_{ji}}{1 + \sum\limits_{t_j \in \bar{d}_i \cap \bar{Q}} w_{ji}}. \end{array} \qquad [2]$$

Théorème 2.1 *La fonction cosinus peut s'écrire sous la forme :*

$$cosinus\left(\bar{d}_i\,,\,\bar{Q}\right) \;=\; simis\left(\bar{d}_i\,,\,\bar{Q}\right) \,+\, f\left(\bar{d}_i\,,\,\bar{Q}\right), \qquad [3]$$

où $f\left(\bar{d}_i\,,\,\bar{Q}\right)$ est une fonction négligeable (Preuve en annexe (page 6)). Ce qui entraîne que les fonctions simis et cosinus sont équivalentes.

Propriétés 2.1 *Soient $\bar{d}_i, \bar{d}_k$ et $\bar{Q}$ des éléments de $B_{index} \cup \{\bar{Q}\}$, la fonction $simis$ vérifi e les propriétés suivantes :*

1.) $simis$ *est symétrique ;*

2.) $0 \leq simis(\bar{d}_i\,,\,\bar{Q}) \leq 1$ *;*

3.) $\bar{Q} \cap \bar{d}_i = \emptyset \implies simis(\bar{d}_i\,,\,\bar{Q}) = 0$ *;*

4.) $simis\left(\bar{d}_i, \bar{Q}\right) \geq simis\left(\bar{d}_k, \bar{Q}\right) \Longleftrightarrow Q$ *est plus proche du document d_i que du document d_k ;*

5.) elle est applicable à tout modèle utilisant une fonction de correspondance ;

6.) elle ne dépend que des termes appartenant à l'intersection de la requête et d'un document versus la fonction cosinus (2.1) utilisant tous les termes de la requête et d'un document.

3. Application

L'évaluation de $simis$ repose sur une étude comparative entre le modèle vectoriel utilisant la fonction $cosinus$ (MVC) et le modèle vectoriel utilisant la fonction $simis$ (MVS).

Illustration par un exemple
Supposons que nos bases sont constituées comme suit :

$$\begin{array}{lcl}
B_{doc} & = & \{d_1,\ d_2,\ d_3,\ d_4,\ d_5,\ d_6\}\ ; \\
Base_{index} & = & \{t_1,\ t_2,\ t_3,\ t_4,\ t_5,\ t_6,\ t_7,\ t_8\}\ ; \\
B_{index} & = & \left\{\bar{d}_1 = \{t_1\ \ t_1\ \ t_3\ \ t_4\}\,,\ \ \bar{d}_2 = \{t_2,\ \ t_5,\ \ t_5,\ \ t_5,\ \ t_6,\ \ t_3\}\,,\right. \\
 & & \bar{d}_3 = \{t_4,\ \ t_6,\ \ t_3,\}\,,\ \ \bar{d}_4 = \{t_8,\ \ t_3,\ \ t_6,\ \ t_3,\ \ t_7,\ \ t_3\}\,, \\
 & & \left.\bar{d}_5 = \{t_7,\ \ t_8,\ \ t_7,\ \ t_4\}\,,\ \ \bar{d}_6 = \{t_1,\ \ t_5,\ \ t_7,\ \ t_3\}\right\}.
\end{array}$$

Soit une requête Q de l'utilisateur dont sa représentation en fonction des termes d'index est définie par $\bar{Q} = \{t_1, t_2, t_5\}$. En utilisant MVC et MVS, les documents d_1, d_2 et d_6 sont restitués. En effet, MVS et MVC restituent les documents dont leur représentation a une intersection non vide avec la représentation de la requête. Le degré de similarité qui existe, par exemple, entre le document d_2 et la requête Q est calculé ci-dessous en fonction de $simis$ et $cosinus$:

$$\left\{\begin{array}{lcl}
simis(\bar{d}_2,\ \ \bar{Q}) & = & \dfrac{w_{22} + w_{52}}{1 + w_{22} + w_{52}}\ ; \\
cosinus(\bar{d}_2,\ \ \bar{Q}) & = & \dfrac{w_{22} * q_2 + w_{52} * q_5}{(w_{22}^2 + w_{52}^2 + w_{62}^2 + w_{32}^2)^{1/2}(q_2^2 + q_3^2 + q_5^2 + q_6^2)^{1/2}}.
\end{array}\right.$$

Rappelons que w_{ji} est le poids du j^{e} terme dans le i^{e} document. La pertinence d'un document par rapport à une requête est calculée en fonction de la pertinence des poids des termes contenus dans leur intersection. Etant donné pour une requête, tous ses termes sont importants, nous prenons en compte pour la fonction $simis$ les poids des termes dans les documents.

Expérimentation
Pour évaluer la performance de la fonction $simis$ dans un système de recherche d'information, nous avons les critères suivants : **taux de rappel**, **taux de précision**, **complexité temporelle**. Notre application repose sur le modèle vectoriel pour rechercher des composants logiciels. Dans ce papier, un composant est assimilé à une classe objet et la base est ainsi constituée des composants issus des bibliothèques scientifiques décrites dans le tableau (**Tableau 1**). Généralement, les résultats obtenus pour la base

langage	sources	composants
JAVA	Jama Jampack MatrixToolkit LinAI	106
C++	Mv++ Sparselib++ Lapack++ ML++ HCL	60

Tableau 1. *Description de la base de composants logiciels*

de composants considérée nous indiquent un taux de rappel et un taux de précision qui sont à 100% pour certaines requêtes, ce qui veut dire que tous les documents pertinents sont restitués. Cependant pour d'autres requêtes, le taux de précision reste faible. Les deux fonctions sont équivalentes en termes de taux de rappel et de taux de précision. Le temps d'exécution est plus court pour MVS que pour MVC.

Discussion et Perspectives
Mathématiquement, nous venons de démontrer que les fonctions $simis$ et $cosinus$ sont équivalentes d'après le théorème(2.1), $simis$ a toutes les propriétés de $cosinus$. $simis$ et $cosinus$ ont le même taux de rappel et le même taux de précision du fait qu'elles restituent les mêmes documents d'une base pour une requête donnée, en plus $simis$ est moins complexe que $cosinus$.
Afin de mieux évaluer les performances de la fonction proposée, nous établirons une étude comparative en maintenant les mêmes critères d'évaluation entre MVC et MVS dans des collections de test existantes, par exemple TREC (Text REtrieval Conference).

4. Bibliographie

[BAE 99] BAEZA-YATES R., RIBEIRO-NETO B., *Modern Information Retrieval*, Addison Wesley,ACM Press New York, ISBN 0-201-39829-X, 1999.

[HOC 02] HOCINE A., RAFFINAT P., « Etude et prototypage d'un système de catalogage de composants », rapport, 2002, Rapport fi nal NÂ° 25992, Convention IFP-UPPA.

[IHA 04a] IHADJADENE M., *Les systèmes de Recherche d'Information : modèle conceptuels*, Lavoisier, 2004.

[IHA 04b] IHADJADENE M., *Méthodes et Modèles avancés pour la Recherche d'Information*, Lavoisier, 2004.

[OUS 05] OUSSALAH M., *INGENIERIE DES COMPOSANTS :Concepts, techniques et outils*, Vuibert, 2005.

[SAL 71] SALTON G., « The SMART retrieval system », *Prentice Hall, Englewood Cliffs, ,* 1971, p. 88-89.

Annexe

Preuve du théorème 2.1 (3)

Nous avons

$$cosinus\left(\bar{d}_i\,,\,\bar{Q}\right) \;=\; simis\left(\bar{d}_i\,,\,\bar{Q}\right) \;-\; simis\left(\bar{d}_i\,,\,\bar{Q}\right) \;+\; cosinus\left(\bar{d}_i\,,\,\bar{Q}\right). \quad [4]$$

Posons $f\left(\bar{d}_i\,,\,\bar{Q}\right) \;=\; cosinus\left(\bar{d}_i\,,\,\bar{Q}\right) \;-\; simis\left(\bar{d}_i\,,\,\bar{Q}\right)$.

Considérons les inégalités suivantes :

$$\begin{array}{rcl} -\left(\sum_{t_j\in\bar{d}_i} w_{ji}^2\right)^{1/2}\left(\sum_{t_j\in\bar{Q}} q_j^2\right)^{1/2} & \leq & -\left(\sum_{t_j\in\bar{d}_i\cap\bar{Q}} w_{ji}^2\right)^{1/2}\left(\sum_{t_j\in\bar{d}_i\cap\bar{Q}} q_j^2\right)^{1/2}; \\ -\left(\sum_{t_j\in\bar{d}_i\cap\bar{Q}} w_{ji}^2\right)^{1/2}\left(\sum_{t_j\in\bar{d}_i\cap\bar{Q}} q_j^2\right)^{1/2} & \leq & -\sum_{t_j\in\bar{d}_i\cap\bar{Q}} w_{ji}q_j. \end{array} \quad [5]$$

Il s'agit de montrer que f défi nie ci-dessus est négligeable. Par défi nition, la fonction f s'écrit comme suit :

$$f\left(\bar{d}_i\,,\,\bar{Q}\right) \;=\; \frac{\sum_{t_j\in\bar{d}_i\cap\bar{Q}} w_{ji}\,q_j}{\left(\sum_{t_j\in\bar{d}_i} w_{ji}^2\right)^{1/2}\left(\sum_{t_j\in\bar{Q}} q_j^2\right)^{1/2}} \;-\; \frac{\sum_{tj\,\in\,\bar{d}_i\cap\bar{Q}} w_{ji}}{\sum_{tj\,\in\,\bar{d}_i\cap\bar{Q}} w_{ji}\;+1}. \quad [6]$$

D'après les inégalités (5), nous en déduisons

$$f\left(\bar{d}_i\,,\,\bar{Q}\right) \;\leq\; \frac{\sum_{t_j\in\bar{d}_i\cap\bar{Q}} w_{ji}\,q_j}{\left(\sum_{t_j\in\bar{d}_i} w_{ji}^2\right)^{1/2}\left(\sum_{t_j\in\bar{Q}} q_j^2\right)^{1/2}\cdot\sum_{tj\,\in\,\bar{d}_i\cap\bar{Q}} w_{ji}}. \quad [7]$$

Compte tenu de l'implication suivante :

$$\forall t_j\in\bar{Q} \Longrightarrow q_j = q,\ \text{alors}\ f\left(\bar{d}_i\,,\,\bar{Q}\right) \;\leq\; \frac{1}{\sqrt{n_1\times n_2}\times q}.$$

Où n_1 et n_2 représentent respectivement le cardinal de $\bar{Q}$ et le cardinal de $\bar{d}_i$, et $w_{ij} >= q$. Nous en concluons que $\frac{1}{\sqrt{n_1\times n_2}\times q}$ est négligeable d'où la fonction f l'est aussi.