Proposition pour l'intégration des réseaux petits mondes en recherche d'information

Mohamed Khazri, Mohamed Tmar, Mohamed Abid, Mohand Boughanem

Proposition pour l'intégration des réseaux petits mondes en recherche d'information

Mohamed Khazri* — Mohamed Tmar** — Mohamed Abid*** — Mohand Boughanem****

*,*** Ecole Nationale d'Ingénieurs de Sfax
4, route de Soukra
3038 Sfax

mohamed.khazri@yahoo.fr
Mohamed.Abid@enis.rnu.tn

** Institut Supérieur d'Informatique et du Multimédia de Sfax
4, route de tunis
3018 Sfax

mohamed.tmar@isimsf.rnu.tn

**** Institut de Recherche en Informatique de Toulouse
118, route de Narbonne
31000 Toulouse Cedex 9

bougha@irit.fr

RÉSUMÉ. Nous proposons dans ce papier une approche de classification d'un corpus de documents. Elle consiste en une représentation du corpus sous forme de graphe, où les liens sont définis par certains critères. Ces liens sont quantifiés par des mesures de simialrité. Nous visons à intégrer ce contexte dans l'approche de classification afin de constituer des réseaux petits mondes de documents homogènes. L'homogénéité des classes est valuée suivant les propriétés des réseaux petits mondes. Les classes, ainsi que leurs propriétés, nous servent au ré-ordonnancement de documents-résultats de recherche. Quelques expérimentations ont été menées sur un corpus issu de TREC [1] et les résultats obtenus montrent l'apport des réseaux petits mondes en recherche d'information.

ABSTRACT. We propose in this paper an approach for document clustering. It consists of representing the corpus as a document graph, where the links are defined by some criteria. These links are quantified by simialrity measures. We aim join this context into the approach of classification to constitute small-worlds networks of homogeneous documents. The homogeneity of the clusters is measured according to the properties of small worlds. The clusters, as well as their proprietes, allow to rerank search results. Some experiments were done on a corpus provided by TREC and the obtained results show the contribution of small-worlds networks in information retrieval.

MOTS-CLÉS : Recherche d'information, clusterisation, réseaux petits mondes, ré-ordonnancement

KEYWORDS : Information retrieval, clustering, small-worlds networks, re-ranking

1. Text REtrieval Conference, compagne d'evaluation internationale dont l'objectif est de promouvoir la recherche d'information.

1. Introduction

Les systèmes de recherche d'information préconisent une fonctionnalité très intéressante, voir indispensable, lors de tout processus de recherche : il s'agit de la reformulation automatique de la requête [1][2]. Cette fonctionnalité permet de rétablir les choix de l'utilisateur dans la perspective de retrouver plus de documents qui répondent à son besoin en information. Le besoin en information de l'utilisateur est cependant très vague : l'utilisateur ne sait en général pas ce qu'il cherche. Par ailleurs, il peut tolérer un résultat initial imprécis sous réserve de l'améliorer. Parmi les techniques les plus populaires, la reformulation par réinjection de pertinence (relevance feedback) est la plus utilisée.

Le feedback a toujours été vu sous l'angle de la reformulation de la requête, et plus précisément la repondération des termes qui apparaissent. Bien que problématique, la reformulation de la requête se base en général sur des techniques de reformulation très basiques ; l'algorithme de reformulation de Rocchio [3] a toujours été le plus utilisé, bien que simpliste.

La limite de ces méthodes est qu'elle se base uniquement sur les termes de la requête. Or, bien d'autres paramètres peuvent représenter un intérêt particulier pour un utilisateur. Par exemple, sur internet, la popularité d'une page [4] pourra être un élément fondamental de sélection et de pertinence.

En outre, la pertinence est trop subjective : elle dépend en particulier de l'utilisateur qui juge la pertinence selon son point de vue.

Faire recours à de nouvelles méthodes d'apprentissage est alors devenu une nécessité. Plusieurs modèles qui ont été auparavant délaissés, tels que la classification [5], sont repris en vue d'améliorer l'apprentissage en recherche d'information.

Zaragoza [6] propose de revoir l'apprentissage. Il considère que pour retrouver plus de documents pertinents et écarter plus de documents non pertinents, il suffit de retrouver la fonction qui sépare les deux ensembles. Cette fonction peut dépendre de tous les paramètres des documents et peut avoir des formes différentes selon les jugements des utilisateurs.
Retrouver cette fonction extrêmement complexe. En effet, aucune information à priori ne permet de prévoir sa forme, ni les paramètres qui devraient être pris en considération pour sa construction. De plus, l'optimisation de cette fonction est optimale.

Ce challenge reste très vague et très problématique en recherche d'information. Une technique très prometteuse émerge aujourd'hui qui vise à mesurer le lien entre les réseaux petits mondes (RPM, small-world en anglais [7]) et la recherche d'information.

Les propriétés des réseaux petits mondes permettent de revoir la structure des systèmes de recherche d'information selon un point de vue différent de l'habituel. Notre perspective fondamentale repose sur la proposition de nouvelles méthodes d'apprentissages en recherche d'information en faisant appel aux réseaux petits mondes.

Nous proposons dans ce papier une méthode basée sur la classification des documents. Nous considérons qu'une classe homogène peut être utilisée comme moyen efficace pour estimer les scores d'autres documents. Nous admettons qu'une classe est homogène et compacte si elle admet certaines propriétés : celles des réseaux petits mondes. Un ensemble de documents est alors transformé en un graphe ou les noeuds sont des documents et les liens matérialisent des ressemblances vis-à-vis de certains paramètres.

Le papier est organisé comme suit : dans la deuxième section, nous présentons la notion des réseaux petits mondes et leurs utilisations. La troisième section présente notre approche de constitution de réseaux petits mondes de documents. Dans la quatrième section, nous détaillons les expérimentations que nous avons menées et nous exposons les résultats obtenus. La cinquième section conclut.

2. Réseaux petits mondes : définitions et propriétés

Les réseaux petits mondes sont utilisés pour des solutions de routage d'informations dans des réseaux de coummunication ou d'interaction [8]. Le principe fondamental de la théorie des réseaux petits mondes repose sur le fait qu'un ensemble d'individus d'une population de nature particulière peut être assimilé à un élément unique dont le comportement est unifié à l'ensemble des individus qui le constituent.

Ce principe est très proche de celui de la classification, mais les réseaux petits mondes procurent des propriétés intrinsèques à la logique de la communauté ou des réseaux réels.

Une relation entre individus est caractérisée par son contenu (nature de la ressource qui est échangée), sa direction (directe ou indirecte) et sa force (intensité de l'échange). A partir de la notion de relation, on peut définir celle du lien : un lien assure la connexion entre deux individus à travers une ou plusieurs relations. Deux individus peuvent ainsi être reliés par un lien constitué d'une seule relation (appartenir à la même organisation, par exemple), ou par un lien reposant sur des relations multiples (partager de l'information, assister ensemble à des conférences, s'entraider financièrement, etc.). Les liens varient aussi en termes de contenu, de direction et de force [9].

Young [10] formalise via un graphe ce qu'est un réseau relationnel : un réseau social se représente par un graphe G composé d'un ensemble fini V de sommets et d'un ensemble E d'arcs non orientés. Chaque sommet i représente un individu du système. Un arc i, j relie deux individus i et j si et seulement si i et j sont voisins, par exemple ils sont réciproquement influencés par leurs actions. Chaque lien i, j est supposé être doté d'une intensité $\beta_{ij} = \beta_{ji} > 0$ avec $\beta_{ij} > \beta_{ik}$ signifie que i attache plus d'importance aux actions de j qu'à celles de k.

Milgram [11] décrit une expérimentation dont le protocole consiste en la transmission du courrier d'un destinataire à son émetteur par le canal des connaissances (passage du courrier de main en main). Il constate que la plupart des couples émetteur-destinataire, bien qu'apparemment distants, sont en fait connectés par une chaîne très courte de connaissances intermédiaires. Cette chaîne est typiquement de longueur six. Le paradigme des six degrés de séparation a été popularisé par [12].

La seconde propriété exhibée est le clustering. [13] montrent que dans la plupart des réseaux réels, la probabilité d'existence d'un lien entre deux individus est plus importante si les deux individus en question ont une connaissance en commun (ou plusieurs). Pour le dire autrement, la probabilité que deux amis en connaissent un troisième est supérieure à la probabilité que deux personnes choisies aléatoirement dans la population en connaissent une troisième. Watts et Strogatz ont défini un coefficient de clustering, habituellement noté C, qui est la probabilité que deux connaissances d'une personne aléatoirement choisie se connaissent entre elles. Ils montrent que pour un grand nombre de réseaux, ce coefficient varie de quelques pour cent à 50 pour cent.

La troisième propriété des réseaux est celle relative à l'inclinaison de la distribution des degrés au sein du réseau ; le degré d'un sommet i étant égal au nombre de sommets adjacents à ce sommet. Dans une population de N individus, on s'intéresse donc à la forme de la distribution du nombre d'individus ayant n voisins (n variant de 0 à $N-1$). Le réseau, basé sur un graphe aléatoire, est réputé doté de caractéristiques de distribution des degrés éloignées de celles des réseaux réels.

Les propriétés des réseaux réels sont ainsi étudiées en sociométrie et servent de base à la réflexion de modélisation. L'enjeu est ici de construire des algorithmes de génération de réseaux dont la topologie est dotée des propriétés des réseaux réels.

Watts et al. [13] définissent les réseaux petits mondes par deux propriétés : la longueur du chemin caractéristique L, et l'indice de clustéring C. Ces deux indices permettent de faire la différence entre les réseaux petits mondes, les réseaux générés aléatoirement et les réseaux réguliers [14]. L est la moyenne des longueurs des plus courts chemins entre deux nœuds du graphe. Soit $d_{min}(i,j)$ la longueur du plus court chemin entre deux nœuds i et j et N le nombre total de nœuds alors :

$$L = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{min}(i,j)$$

L'indice de clustering C est un indice de la richesse de la cohésion locale. Il est défini de la manière suivante : si un nœud s a k voisins alors il peut exister au maximum $n = \frac{k \times (k-1)}{2}$ arcs entre ces k nœuds. Soit m le nombre d'arcs qu'il y a effectivement entre ces k nœuds alors le coefficient de clustering C_s associé au nœud est $\frac{m}{n}$. Le coefficient global C est égal à la moyenne des C_s.

$$C = \frac{1}{N} \sum_{i=1}^{N} C_s$$

La valeur prise par l'indice de clustering variera entre zéro pour un graphe totalement déconnecté et un pour un graphe complet. Pour savoir si on a affaire à un graphe de type petit monde, on compare les coefficients C et L à ceux d'un graphe aléatoire ayant le même nombre de nœuds (N) et le même nombre moyen d'arcs par nœud (k). Pour un graphe petit monde on a alors [13] :

$$L \geq L_{rand} \approx \frac{log(N)}{log(k)} \quad \text{et} \quad C >> C_{rand} \approx \frac{k}{N}$$

Les propriétés des réseaux petits mondes paraissent intéressantes dans les problèmes de classification. D'autant plus que ces propriétés sont valuées. Comme application à la

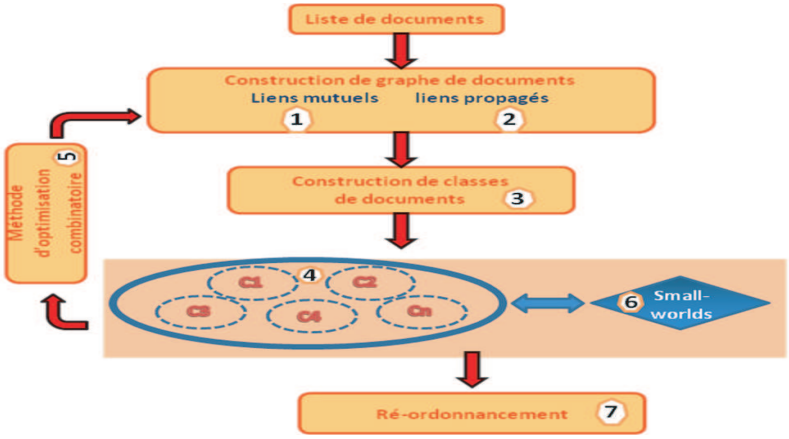

Figure 1. *Processus de Notre approche.*

recherche d'information, nous présumons qu'un ensemble de documents peut constituer des réseaux petits mondes pour autant qu'ils parlent du même sujet, et qu'une idée peut être transmise d'un document à un autre si les auteurs partagent le même intérêt ou les mêmes idées. L'objectif général de ce travail serait alors de construire des réseaux petits mondes de documents homogènes. L'homogénéité des classes est valuée suivant les propriétés des réseaux petits mondes. Les classes, ainsi que leurs propriétés, nous servent au ré-ordonnancement de documents-résultats de recherche du système OKAPI [15]. La figure 1 illustre le processus général de notre approche.

3. Constructionn de réseaux petits mondes

Nous considérons qu'une classe est un réseau petit monde de documents. Il est alors nécessaire de construire d'abord un graphe de documents.

3.1. Construction du graphe de documents

Le graphe de documents est constitué de nœuds (chaque nœud représente un document) et de liens valués. Le poids d'un lien entre deux documents est fonction de la similarité entre ces documents vis-à-vis de l'ensemble des critères. Un critère peut être toute valeur qui caractérise un document, par exemple la longueur du document, le nombre d'apparitions d'un terme, la popularité du document, etc. Chaque critère possède un poids, ce poids traduit l'intérêt porté par l'utilisateur à celui-ci et peut être identifié par son jugement. Le poids du lien entre deux vecteurs documents $\vec{d_i} = (d_{i_1}, d_{i_2}, ..., d_{i_n})^T$ et $\vec{d_j} = (d_{j_1}, d_{j_2}, ..., d_{j_n})^T$, où d_{x_y} est la valeur du critère y dans le document x, est donné par :

$$L\left(\vec{d_i}, \vec{d_j}\right) = \sum_{k=1}^{n} w_k S_k \left(\vec{d_i}, \vec{d_j}\right)$$

où w_k est le coefficient (poids) du critère k et $S_k\left(\vec{d_i}, \vec{d_j}\right)$ est la similarité entre $\vec{d_i}$ et $\vec{d_j}$ vis-à-vis du critère k. Le choix des critères ainsi que leurs coefficients pourra être réalisé par apprentissage. Cette phase d'analyse est similaire à la repondération des termes pendant la phase d'expansion et de reformulation automatique de la requête en recherche d'information. Nous avons opté pour une méthode d'apprentissage mais elle ne sera pas expérimentée dans ce papier. Pour notre part, nous présumons que les critères discriminants sont les termes d'indexation et que leurs coefficients sont égaux ($w_k = 1 \; \forall \; k \in \{1, 2, \ldots, n\}$). En effet, l'objectif de ce papier est de montrer l'apport des propriétés des réseaux petits mondes dans la recherche d'information. La similarité (S_k) entre deux documents $\vec{d_i}$ et $\vec{d_j}$ vis-à-vis du critère k est donnée par :

$$S_k\left(\vec{d_i}, \vec{d_j}\right) = d_{i_k} \times d_{j_k}$$

Si le coefficient de chaque critère est égal à 1 alors la similarité entre deux documents $\vec{d_i}$ et $\vec{d_j}$ sera le produit scalaire des vecteurs associés :

$$S_k\left(\vec{d_i}, \vec{d_j}\right) = \sum_{k=1}^{n} w_k \times S_k\left(\vec{d_i}, \vec{d_j}\right) = \sum_{k=1}^{n} d_{i_k} \times d_{j_k}$$

3.2. Propagation des liens

Cette première construction du graphe induit les relations directes entre deux documents. Or, les relations peuvent être propagées par transitivité. Cette hypothèse est une conséquence directe des propriétés des réseaux petits mondes. En effet, partant de la loi de communauté (le monde est petit) [11] [12], si un document $\vec{d_i}$ est en relation avec $\vec{d_j}$ et celui-ci est en relation avec $\vec{d_k}$, alors $\vec{d_i}$ est en relation avec $\vec{d_k}$. Il convient alors à ce stade d'analyse de compléter la construction du réseau par fermeture transitive. C'est-à-dire renforcer le lien entre chaque paire de documents s'il existe un chemin additionnel entre eux. Naturellement, ce lien doit être affaibli en fonction de la longueur du chemin : plus la longueur est élevée, plus le lien est faible. Soit k la longueur d'un chemin entre deux documents, nous utilisons le coefficient d'affaiblissement suivant :

$$e^{\alpha - \alpha \times k}$$

Ce facteur est associé aux poids des liens figurant sur ce chemin. Plus les poids augmentent, plus le poids du lien est important. Nous choisissons le poids le moins élevé parmi tous les poids qui figurent sur ce chemin. Le poids additionné du chemin $\vec{d}_{l_1}, \vec{d}_{l_2}, \ldots, \vec{d}_{l_k}$ où les $\vec{d}_{l_m}$ sont mutuellement différents est alors :

$$e^{\alpha - \alpha \times k} \times \min_{1 \leq i < k} S\left(\vec{d_{l_m}}, \vec{d_{l_{m+1}}}\right)$$

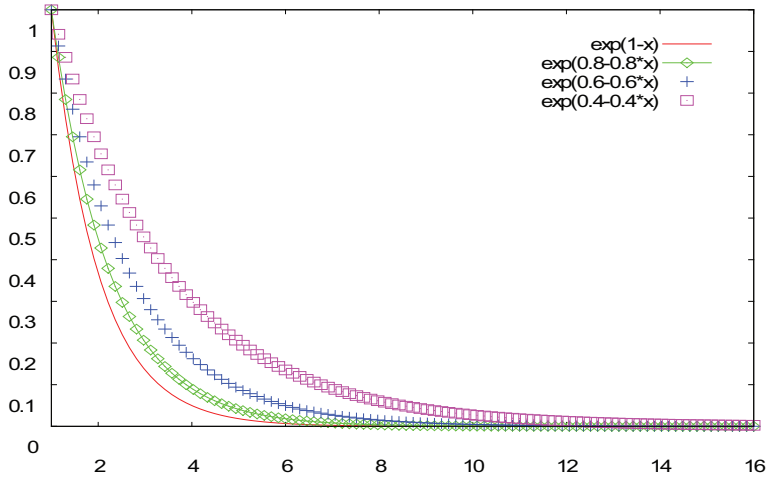

Figure 2. *Allure du coefficient d'affaiblissement pour différentes valeurs de* α

Les poids des liens sont additionnés pour tous les chemins de longueur k et pour toutes les valeurs de k possibles (de 1 à $N-2$). Le poids du lien entre deux documents est alors actualisé comme suit :

$$S\left(\vec{d_i},\vec{d_j}\right) \leftarrow S\left(\vec{d_i},\vec{d_j}\right) + \sum_{k=1}^{N-2} e^{\alpha-\alpha\times k} \sum_{\substack{\left(d_{l_1}\dots d_{l_k}\right)\in(C-\{d_i,d_j\})^k \\ d_{l_s}\neq d_{l_t}\forall s,t\in\{1\dots k\},s\neq t}} \min_{1\leq m<k} S\left(\vec{d_{l_m}},\vec{d_{l_{m+1}}}\right)$$

La valeur de α permet d'atténuer l'effet de transitivité et doit être fixée par expérimentation(dans nos expérimentations $\alpha = 1$). Par ailleurs, d'après les études effectuées par Milgram [11], qui se basent sur l'hypothèse de la loi des six degrés de séparation (qu'il serait possible d'atteindre n'importe quel individu par une chaîne de six poignées de mains entre des individus qui se connaissent deux-à-deux), l'application de la fermeture transitive se limite à $k = 6$, ceci permet également de réduire la complexité de l'algorithme. La forme serait alors :

$$S\left(\vec{d_i},\vec{d_j}\right) \leftarrow S\left(\vec{d_i},\vec{d_j}\right) + \sum_{k=1}^{6} e^{\alpha-\alpha\times k} \sum_{\substack{\left(d_{l_1}\dots d_{l_k}\right)\in(C-\{d_i,d_j\})^k \\ d_{l_s}\neq d_{l_t}\forall s,t\in\{1\dots k\},s\neq t}} \min_{1\leq m<k} S\left(\vec{d_{l_m}},\vec{d_{l_{m+1}}}\right)$$

Ceci permet également de réduire la complexité de l'algorithme. Par ailleurs, la figure 2 montre que l'effet de transitivité s'annule ou presque pour les valeurs de k supérieures à 6, ce qui confirme notre choix de la fonction $e^{\alpha-\alpha\times k}$. Même pour les valeurs de k inférieurs à 6, la fermeture transitive est de complexité élevée. En effet, pour un nombre N de documents, on peut identifier un nombre θ de chemins de longueurs différentes :

$$\begin{aligned}
\theta &= C_N^2 \times \sum_{k=1}^{6} A_{N-2}^k \\
&= \frac{N!}{2! \times (N-2)} \times \sum_{k=1}^{6} \frac{(N-2)!}{k!} \\
&= \frac{N!}{2!} \times \sum_{k=1}^{6} \frac{1}{k!} = \frac{N!}{2} \times 1.72 \\
&\approx N!
\end{aligned}$$

Or, dans notre processus, la fermeture transitive se fait par calcul de puissances matricielles ($M\,[N \times N]$) où M est la matrice d'adjacence. Afin de réduire la complexité, nous envisageons de procéder à la diagonalisation de cette matrice ($M = A \times U \times A^{-1}$, où U est une matrice diagonale), M^k devient alors :

$$\begin{aligned}
M^k &= \left(A \times U \times A^{-1}\right)^k \\
&= \underbrace{A \times U \times A^{-1} \times A \times U \times A^{-1} .... \times A \times U \times A^{-1}}_{kfois} \\
&= A \times \underbrace{U \times A^{-1} \times A \times U \times A^{-1} .... \times A \times U \times}_{kfois} A^{-1} \\
&= A \times U^k \times A^{-1}
\end{aligned}$$

U est une matrice diagonale, soit :

$$U = \begin{pmatrix} u_1 & 0 & \dots & 0 \\ 0 & u_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & u_n \end{pmatrix} \; alros \; U^k = \begin{pmatrix} u_1^k & 0 & \dots & 0 \\ 0 & u_2^k & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & u_n^k \end{pmatrix}$$

le calcul de la puissance de M ne nécessite alors que le calcul des puissances de la diagonale de U et le produit de 3 matrices A, U^k et A^{-1}.

3.3. Construction des classes de documents

Nous utilisons une méthode de classification hiérarchique. Ayant un nombre N de documents, la classification hiérarchique démarre avec N classes où chaque document représente à lui seul une classe, puis progressivement on unifie les classes les plus homogènes à chaque itération. Le nombre de classes à la deuxième itération sera alors $N-1$, puis $N-2$, etc.
L'algorithme 1 illustre la classification hiérarchique. Le processus s'arrête lorsque le nombre de classes vaut $c < N$.

Algorithm 1 : Classification hiérarchique

1: $\forall i \in \{1, 2 \dots N\}, D_i = \{d_i\}, v_i \leftarrow d_i$
2: $E \leftarrow \{D_1, D_2 \dots D_N\}$
3: **répéter**
4: $(i, j) \leftarrow argmax_{(l,k),X_l \in E, X_k \in E, k \neq l} Sim(v_l, v_k)$
5: $D_i \leftarrow D_i \cup D_j$
6: $E \leftarrow E - \{D_j\}$
7: $v_i^* = \frac{\sum\limits_{d \in D_i} d}{|d_i|}$
8: **jusqu'à** $|E| = c$

L'identification du nombre de classes c est décisif de la qualité de classification et de la nature des classes construites. A chaque itération, nous calculons une valeur d'inertie intra-classe. Ayant c classes, l'inertie intra-classe est calculée comme suit :

$$I_w = \frac{\sum_{i=1}^{c} \frac{\sum_{(d,d') \in D_i^2, d \neq d'} S(\vec{d}, \vec{d}')}{|D_i|}}{c}$$

où D_i est une classe.

L'inertie intra-classe permet de quantifier l'homogénéité des classes. Nous estimons qu'un nombre raisonnable de classes correspond au point où l'évolution de l'inertie (la pente de la tangente à la courbe d'inertie) est proche d'une valeur qu'on détermine par expérimentation.

La similarité entre deux classes de documents est fonction des poids des liens entre les documents de chaque classe et est donnée par :

$$S^*(D_i, D_j) = \frac{\sum_{(d_k, d_l) \in D_i \times D_j} S\left(\vec{d_i}, \vec{d_j}\right)}{|D_k| \times |D_l|}$$

La distance entre deux classes de documents est donnée par la distance moyenne entre deux documents appartenant chacun à une classe. Le calcul de la similarité entre classes de documents est nécessaire pour la classification.

4. Expérimentations et résultats

Dans cette section, nous présentons les résultats d'expérimentation effectuée pour évaluer notre approche. Nous nous basons sur les 1000 premiers résultats du système OKAPI pour dix requêtes issues de la compagne d'évaluation TREC.

4.1. OKAPI

OKAPI est un système basé sur un modèle de recherche probabiliste. La fonction de pondération des termes utilisée est $BM25$. Etant donné un terme t, une requête q est un document d, le poids w de d selon q et le terme est calculé par la fonction suivante :

$$w = \frac{(k_1 + 1) \times t_f}{k + t_f} \times log \frac{N - n + 0.5}{n + 0.5} \times \frac{(k_3 + 1) \times qt_f}{k_3 + qt_f} \oplus k_2 \times nq \times \frac{advl - dl}{advl + dl}$$

Le tableau 1 illustre les variables de la fonction $BM25$.

Variables	Déscription
N	le nombre total de documents dans la collection
n	le nombre de documents contenant le terme t,(n<N)
t_f	fréquence du terme t dans le document d
qt_f	fréquence du terme t dans la requête q
nq	nombre de termes dans la requête q
dl	longueur du document d
$avdl$	longueur moyenne des documents dans la collection
k	$k_1 \times (1_b + b \times \frac{dl}{avdl})$
$k1, k2, k3, b$	Paramètres constants

Tableau 1. *Paramètres de la fonction $BM25$.*

4.2. Collection

les expérimentations que nous avons effectuées ont été réalisées sur une collection de documents résultats issue du système OKAPI pour dix requêtes. Le tableau 2 montre les caractéristiques de la base.

Requêtes	Nombre total de documents	Nombre de documents pertinents
R351	958	48
R352	966	246
R353	900	122
R354	888	361
R355	711	45
R356	852	17
R357	924	270
R358	583	51
R359	815	28
R360	784	151

Tableau 2. *Caractéristiques de la base.*

4.3. Processus de ré-ordonnancement

le processus de ré-ordonnancement et d'évaluation (voir figure 1) se déroule comme suit :

• *Construction de classes de documents et classification :* Le contenu de chaque document est pondéré selon la formule de pondération suivante :

$$w_t\left(t,d\right) = tf\left(t,d\right) \times log\left(\frac{N}{N\left(t\right)}\right)$$

où t est un terme, d un document, $tf\left(t,d\right)$ le nombre d'occurences du terme t dans le document d, N le nombre total de documents et $N(t)$ le nombre de documents

contenant le terme t. Nous constatons que les classes construites vérifient globalement les propriétés des réseaux petits mondes (voir tableau 3).

Requête	C_{moyen}	L_{moyen}	L_{rand}	C_{rand}
$R351$	0.78	15.18	2.47	0.56
$R352$	0.74	13.44	1.16	0.52
$R353$	0.9	8.77	2.66	0.85
$R354$	0.38	11.58	1.26	0.21
$R355$	0.47	3.5	4.4	0.08
$R356$	0.57	2.1	2.9	0.053
$R357$	0.43	9	3.7	0.08
$R358$	0.49	11.20	1.180	0.069
$R359$	0.85	2.38	2.25	0.092
$R360$	0.72	2.52	2.48	0.13

Tableau 3. *Classes construites, les propriétés L et C vérifient que les classes construites vérifient généralement les propriétés des réseaux petits mondes.*

• *Ré-ordonnancement :* Le but de ces expérimentations est de mettre en valeur l'adaptation des propriétés des réseaux petits mondes en recherche d'information. Pour cela, nous avons adapté ces propriétés sur la construction d'une liste triée par ordre décroissant de pertinence-système des résultats pour chacune de ces requêtes. Nous établissons la liste triée sur la base des classes construites selon deux critères de préférence :

- R_c : une relation d'ordre inter-classe telle que si deux classes D_1 et D_2 vérifient $D_1 R_c D_2$ alors D_1 est potentiellement meilleure que D_2. Le critère de préférence que nous utilisons est le coefficient de clusterisation moyen :

$$D_1 R_c D_2 \Leftrightarrow C_1 > C_2$$

où C_1 et C_2 sont les coefficients moyens de clusterisation de D_1 et D_2.

- R_d : une relation d'ordre intra-classe (donc inter-documents) telle que si deux documents d_1 et d_2 vérifient $d_1 R_d d_2$ alors d_1 est potentiellement plus pertinent que d_2. Le critère de préférence que nous utilisons est le coefficient de clusterisation :

$$d_1 R_d d_2 \Leftrightarrow c_{d_1} > c_{d_2}$$

où c_{d_1} et c_{d_2} sont les coefficients de clusterisation de d_1 et d_2.

Pour mesurer la performance de notre méthode, nous avons comparé les valeurs de précision exacte obtenues avec celles obtenues par le système OKAPI, qui enregistre les meilleurs résultats parmi tous les participants officiels de TREC. Les résultats obtenus sont illustrés par le tableau 4. Nous observons que notre approche donne des résultats globalement comparables à ceux d'OKAPI, ce qui montre l'apport des réseaux petits mondes en recherche d'information.

Notons que dans le processus d'expérimentation, nous ne réalisons pas l'apprentissage des poids des critères (w_k). Les poids considérés sont identiques pour tous les

Requête	Nombre de documents pertinents	OKAPI	NOTRE APPROCHE
R351	48	0.107	0.09
R352	246	0.026	0.024
R353	122	0.049	0.063
R354	361	0.016	0.018
R355	45	0.049	0.032
R356	17	0.011	0.009
R357	270	0.025	0.024
R358	51	0.102	0.033
R359	28	0.046	0.016
R360	151	0.037	0.077

Tableau 4. *Comparatif des résultats (précision exacte) avec OKAPI.*

termes ($w_k = 1 \; \forall \; k \in \{1, 2, \ldots, n\}$). Pour les requêtes $R360$, $R354$ et $R353$, les résultats sont nettement meilleurs que OKAPI, ce qui signifie que si les critères w_k sont tous identiques, alors cette situation est plus adaptée à ces requêtes que les autres. Nous envisageons d'intégrer la phase d'apprentissage.

5. Conclusion

Nous avons présenté dans ce papier une approche de classification de documents. La classification vise à construire des classes qui vérifient les propriétés des réseaux petits mondes. L'originalité de ce travail réside dans la mesure du lien entre les réseaux petits mondes et la recherche d'information, à savoir l'existence de liens mutuels et de liens propagés ainsi que la prise en compte de la loi de six degré de séparation dans le contexte de la recherche d'information. Les expérimentations sur la collection TREC avaient pour objectif de tester l'apport des propriétés des réseaux petits mondes en recherche d'information. Les résultats obtenus montrent cet apport et s'avèrent très encourageants. Les perspectives à court terme concernent l'apprentissage des critères de sélection afin de mieux cibler la recherche et d'ajouter correctement ces poids pour que la construction des réseaux petits mondes en soit orientée.

6. Bibliographie

[1] S. ROBERTSON, S. JONES , « Relevance weithing of search terms », *JASIS, vol.27*, 129–146, 1976.

[2] M. BOUGHANEM, S. SOULE-DUPAY , « Query modification based on relevance feedback propagation in adhoc environnement information », *In processing and management, vol.35*, 121–139, 1999.

[3] J. ROCCHIO, « Relevance feedback in information retrieval », *MART retrieval system : experiments in automatic document processing*, 313–323, 1971.

[4] O. KURLAND, « PageRank without hyper-links : structural reranking using links induced by language model », *In processing of 28th Annual Internationel ACM SIGIR Conference on Research and developpement in information retrieval*, 2005.

[5] G. SAPORTA, « Probabilités, Analyse des données et statistiques », *Edition technip*, 254-255, 1990.

[6] H. ZARAGOZA, « Relevance weighting for query independent evidence », *Microsoft research, U.K*, 2005.

[7] D.J, WATTS, « Small Worlds », *Princeton university press,Princeton*, 1999.

[8] J. M, KLEINBERG, « Navigation in a small world », *Nature*, 406-845, 2000.

[9] P. COHENDET AND A. KIRMAN AND J.B ZIMMERMANN, « Emergence, formation et dynamique des réseaux, modèles de la morphogène », , 1990.

[10] H. YOUNG, « Diffusion in Social Networks », *ED Economic Studies, The Brookings Institution*, 2, 1999.

[11] S. MILGRAM, « The small-world problem », *Psychology Today*,1,60-67, 1967.

[12] J. GUARE, « six degrees of separation : A play », *New York : Vintage*, 1990.

[13] D. J, WATTS AND H. STROGATZ, « Collective dynamics of small-world networks », *Sature*, 393, 440-442, 1998.

[14] B. BOLLOBÀS, « Random Graphs », *Academic Press, NewYork*, 2001.

[15] M. BEAULIEU, M. GATFORD, X. HUANG, S.E, ROBERTSON, S. WALTER, P. WILLIAMS, « Okapi at TREEC-5 », *Proceeding of the 5th Text REtrieval Conference*, 143-166, 1996.

[16] S. ROBERTSON AND S. WALKERS, « On relevance weights with little relevance information », *In proceeding of the ACM SIGIR*, 16–24, 1997.