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

Mohamed Khazri* — Mohamed Tmar** — Mohamed Abid***

*,*** 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

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. 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. 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 : Classification, réseaux petits-mondes, recherche d'information.

KEYWORDS : Clustering, small-worlds networks, information retrieval.

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

1. Introduction

La recherche d'information est l'ensemble des processus permettant, à partir d'un besoin en information utilisateur exprimé par une requête sous forme d'un ensemble de mots clés, d'extraire et de présenter les seuls documents qui peuvent l'intéresser. L'émergence du besoin de tels systèmes ne cesse d'augmenter. La problématique de la recherche d'information a par la même été prise en main par plusieurs communautés de recherche et notamment les statiticiens et les linguistes. Par ailleurs, les caractéristiques caractérisant la nature de l'information sont de plus en plus variées complexes. Des milliards de documents sont disponibles, accessibles et potentiellement pertinents à des milliers d'utilisateurs du web.

Le référentiel auquel se projette la recherche d'information classique est le document, or dans le cas de très larges corpus, cette unité devient extrêmement difficile à localiser par un simple besoin en information étant noyée dans un contexte aussi étendu que le web.

Une des méthodes permettant de se livrer du concept du document à un concept plus concrêt est la classification [1]. Elle consiste à regrouper sous le même ensemble de critères plus ou moins similaires, un ensemble de documents tels que l'intéressement a postériori d'un utilisateur potentiel à l'un augmente la portabilité de son intéressement aux autres documents de la même classe, mais les techniques actuelles sont de nature purement statistiques et ne donnent satisfaction que lorsque les documents de la même classe sont mutuellement similaires : la construction de classes en se basant sur de telles méthodes pose alors des problèmes souvent liés aux contextes puisque la notion de centroïde de classe est au cœur de la plupart voire toutes ces méthodes. De conception initialement adaptée aux problèmes de routage, les réseaux petits monde [2] concrétisent la notion de classe autrement que celle issue de la classification statistique. Ils considèrent qu'un graphe d'objets peut être subdivisé, selon les liens entre objets, en sous-graphes ayant des propriétés intrinsèques aux réseaux réels ou réseaux de communauté.

Notre approche vise à mesurer l'apport des réseaux petits mondes en recherche d'information. Nous décrivons alors un corpus par un réseau de documents où les liens sont caractérisés par un poids qui reflète la similarité entre deux documents calculée selon les techniques de pondération traditionnelle en recherche d'information.

Afin de valider nos propositions nous avons effectué des expérimentations sur une collection de tests standard de recherche d'information. Cette collection provient de la campagne dévaluation TREC [3]. Les résultats obtenus montrent l'apport des réseaux petits mondes en recherche d'information.

Nous survolons dans la deuxième section la notion des réseaux petits mondes. Nous détaillons dans la troisième section la méthode que nous proposons, pour la quantification des liens qui relient chaque paire de documents, ainsi que la modélisation des réseaux petits mondes de documents. Dans la quatrième section nous décrivons les expérimentations effectuées et les résultats obtenus. La cinquième section conclut.

2. Les réseaux petits mondes : historique et propriétés

Les réseaux petits mondes sont utilisés pour des solutions de routage d'informations dans des réseaux de coummunication ou d'intéraction [4]. 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 peuvent être assimilés à un élément unique dont le comportement est unifié de 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.

[5] 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 [6]. 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 à 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 [5] :

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

Les propriètès des réseaux petits mondes paraissent intéressantes dans les problèmes de classification. D'autant plus que ces propriétés sont valuées. Comme application à la recherche d'information, nous présumons qu'un ensemble de documents peut constituer des réseaux petits pour moins qu'ils parlent du même sujet, et qu'une idée peut être transmise d'un document à un autre document si les auteurs partagent le même intérêt ou les mêmes idées. L'objectif de notre étude serait alors de construire des réseaux petits mondes de documents homogènes.

3. Une Méthode de construction 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. 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) [7] [8], si un document $\vec{d_i}$ et 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 autre que le lien direct entre eux. Naturellement, ce lien doit être affaibli en fonction de la longueur du chemin : plus celle-ci est élevée, plus le lien est faible. Soit k

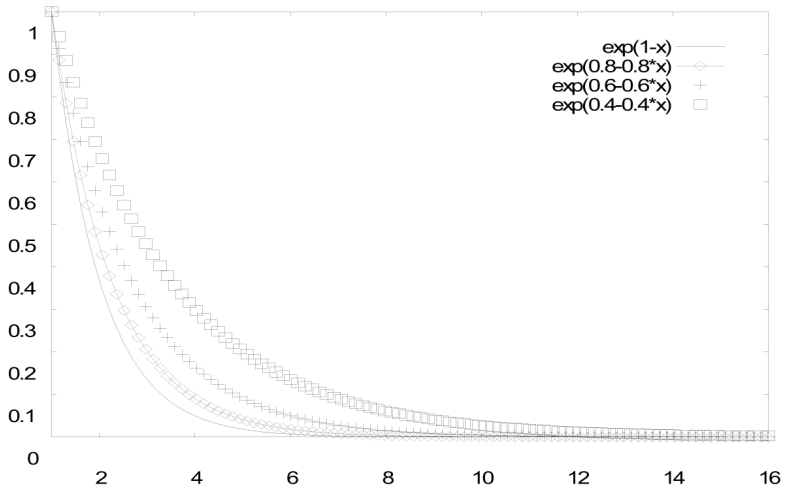

Figure 1. *Allure du coefficient d'affaiblissement pour différentes valeurs de* α

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 $\vec{d}_{l_1}, \vec{d}_{l_2}, ..., \vec{d}_{l_k}$ où les $\vec{d}_{l_m}$ sont mutuellement différents est alors :

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

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. La figure 1 montre l'effet de α sur l'application de la fermeture transitive.

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: $(i, j) \leftarrow argmax_{(l,k), X_l \in E, X_k \in E, k \neq l} Sim(v_i, v_j)$
5: $D_i \leftarrow D_i \cup D_j$
6: $E \leftarrow E - \{D_j\}$
7: $v_i^* = \frac{\sum_{d \in D_i} d}{|d_i|}$
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 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. Nous démontrons que les classes ainsi construites représentent des réseaux petits mondes.

4. Expérimentations et résultats

Dans cette section, nous présentons les résultats d'expérimentation effectuée pour évaluer notre approche. Nous nous basons sur un corpus issu de la compagne d'évaluation

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
$R353$	0.85	8.77	2.66	0.9
$R354$	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 1. *Classes construites, les propriétés L et C vérifient que les classes construites vérifient généralement les propriétés des réseaux petits mondes.*

TREC. Le contenu de chaque document est pondéré selon la formule de pondération suivante :

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

où t est un terme, d est un document, $tf(t,d)$ est le nombre d'occurences du terme t dans le document d, N est le nombre total de documents et $N(t)$ est le nombre de documents contenant le terme t. Nous constatons que les classes construites vérifient globalement les propriétés des réseaux petits mondes (voir tableau 1).

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 proporiétés 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 clustérisation moyen :

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

où C_1 et C_2 sont les coefficients moyens de clustérisation 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 clustérisation :

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

où c_{d_1} et c_{d_2} sont les coefficients de clustérisation 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 [3] qui enregistre les meilleurs résultats parmi tous les participants officiels de TREC. Les résultats obtenus sont illustrés par le tableau 2. Nous observons que notre approche donne des résultats globalement comparables à ceux d'OKAPI ce qui montrent l'apport des réseaux petits

Requête	Nombre de documents pertinents	OKAPI	NOTRE APPROCHE
R351	48	0.107	0.036
R352	246	0.026	0.010
R353	122	0.049	0.044
R354	361	0.016	0.010
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	0.041

Tableau 2. *Comparatif des résultats (précision exacte) avec OKAPI.*

mondes en recherche d'information.

5. Conclusion

Nous avons présentés 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 reseaux petits mondes et la recherche d'information, à savoir l'existance de liens mutuels et de liens propoagés 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, ainsi que la prise en compte des propriétés des réseaux petits mondes pour le réordonnancement des documents.

6. Bibliographie

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

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

[3] 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.

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

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

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

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

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