Clustering en recherche d'information : Concentration vs. Distribution de l'information pertinente

Sylvain Lamprier, Tassadit Amghar, Bernard Levrat et Frédéric Saubion

LERIA - Université d'Angers
2, Bd Lavoisier 49000 Angers
{lamprier,amghar,levrat,saubion}@info.univ-angers.fr

RÉSUMÉ. *S'appuyant sur la Cluster Hypothesis, qui stipule que les documents pertinents à une requête tendent à être plus proches les uns des autres que des documents non pertinents, la plupart des systèmes de recherche d'information réalisant une catégorisation de leurs résultats visent à regrouper l'ensemble des documents pertinents dans un même groupe. Nous proposons ici, par la mise en place de nouvelles mesures d'évaluation, de reconsidérer les bénéfices résultant d'une telle concentration de l'information pertinente. Contrairement à ce qui est habituellement admis, nous montrons finalement que des systèmes réalisant une distribution de l'information pertinente peuvent s'avérer au moins aussi intéressants pour l'utilisateur que des systèmes regroupant l'ensemble des documents pertinents dans un cluster unique.*

ABSTRACT. *Relying on the Cluster Hypothesis, which states that relevant documents tend to be more similar one to each other than to non-relevant ones, most of information retrieval systems producing search results as a set of clusters seek to gather all relevant documents in the same cluster. We propose here, by the settlement of new evaluation measures, to reconsider the benefits of the entailed concentration of the relevant information. Contrary to what is commonly admitted, we finally show that systems realizing a distribution of the relevant information may be at least as useful for the user as systems gathering all relevant documents in a single group.*

MOTS-CLÉS : *Recherche d'information, Clustering, Évaluation*

KEYWORDS: *Information Retrieval, Clustering, Evaluation*

Sylvain Lamprier, Tassadit Amghar, Bernard Levrat et Frédéric Saubion

1. Introduction

Généralement, un système de recherche d'information retourne, en réponse à la requête d'un utilisateur, une liste de documents ordonnée selon une estimation de leur potentiel de pertinence (Manning *et al.*, 2008). Néanmoins, dans le but de réduire l'effort à fournir pour localiser les informations pertinentes, de nombreuses approches ont proposé des présentations alternatives des résultats (Hearst *et al.*, 1996, Tombros, 2002), la plupart utilisant les relations existant entre les documents retournés pour orienter l'utilisateur dans sa recherche (Croft, 1980). Dans ce contexte, les techniques de clustering ont été largement employées afin de faciliter l'accès à l'information en regroupant les résultats aux thématiques similaires[1]. L'utilisation de ce genre de processus en recherche d'information est principalement supportée par la *Cluster Hypothesis* (Jardine *et al.*, 1971), qui stipule que les documents pertinents à une requête tendent à être plus proches les uns des autres que des documents non pertinents et que donc, ces documents ont de grandes chances de figurer dans un même cluster (Manning *et al.*, 2008). Lorsque cette hypothèse se vérifie sur un corpus de textes donné, il suffit alors à l'utilisateur d'identifier le groupe le plus en rapport avec ses besoins pour localiser l'ensemble des documents pertinents.

Alors que la plupart des systèmes réalisant une catégorisation des résultats considèrent la *Cluster Hypothesis* comme un phénomène largement bénéfique, et s'y appuient pour tenter de créer le cluster le plus informatif possible, nous aurons plutôt tendance à la considérer comme un obstacle majeur à la production de groupes permettant à l'utilisateur de localiser rapidement les informations qu'il recherche. Tout d'abord, les documents pertinents tendant à se regrouper dans un seul et même cluster, la majorité des informations initialement présentées à l'écran risquent de ne pas correspondre aux besoins de l'utilisateur. Il suffit alors que le représentant du cluster contenant l'ensemble des documents pertinents n'ait pas été judicieusement choisi pour que l'utilisateur soit incapable de localiser les informations qu'il recherche. De plus, quand bien même le représentant s'avère intéressant, l'impression première que l'utilisateur peut avoir est que peu d'informations correspondent à son sujet (ou que le système de recherche est mauvais), ce qui peut le conduire à reformuler sa requête (ou changer de système) jugeant alors que la piste suivie ne lui permettra pas de satisfaire ses besoins informationnels.

Par ailleurs, bien que la validité de la *Cluster Hypothesis* ait été vérifiée à maintes reprises (Hearst *et al.*, 1996), elle ne fait qu'énoncer des tendances générales et il est probable que certains documents pertinents ne soient pas regroupés avec les autres. Or, le fait qu'une majorité de documents pertinents appartiennent à un même groupe conduit les documents pertinents malencontreusement contenus dans les autres clusters à se trouver isolés parmi des documents déconnectés du besoin de l'utilisateur. Ces documents ont alors peu de chances de trouver dans le représentant de leur cluster (document ou liste de termes) un "porte-parole" efficace. Lorsque la majorité des documents pertinents sont contenus dans un même cluster, l'utilisateur est amené à ne

1. Voir par exemple le meta-moteur de recherche *Vivisimo* (Koshman *et al.*, 2006).

s'intéresser qu'à ce cluster en particulier, pouvant alors passer à côté de documents qui auraient pu compléter sa recherche, en apportant des informations complémentaires, en faisant part d'un point de vue différent ou même, en traitant d'un aspect différent de la question. Le paradoxe est alors considérable : alors que l'on cherche à aider un utilisateur dans sa collecte d'informations, on risque de restreindre sa perception du sujet en l'incitant à ne visiter qu'un seul cluster susceptible de ne contenir que des documents abordant la question sous un même angle.

Enfin, le fait de rassembler l'ensemble des documents pertinents dans un même groupe ne permet pas de faire émerger la structure de l'information pertinente. Or, une même requête peut comprendre un certain nombre d'aspects bien distincts. L'utilisateur, face à un jeu de clusters dans lequel un unique groupe lui est présenté comme étant susceptible de lui être utile, n'a alors aucune idée de la multitude d'aspects que son sujet peut présenter. Lorsqu'il entre dans le cluster des documents pertinents, il est alors face à une liste ordonnée de documents, certes "filtrée" mais qu'il faut tout de même parcourir linéairement jusqu'à avoir le sentiment d'avoir collecté suffisamment d'informations. Un problème majeur se pose alors : l'utilisateur doit prendre la décision d'arrêter la collecte d'informations alors qu'il ne connait pas la diversité des textes en relation avec sa requête[2].

Pour ces différentes raisons, nous pensons que le fait de chercher à regrouper l'ensemble des documents pertinents dans un même groupe n'est pas nécessairement le meilleur choix. Une distribution de l'information pertinente est alors certainement préférable à sa concentration dans un unique cluster. La *Cluster Hypothesis* se pose alors bien comme un obstacle à surmonter, la majorité des documents pertinents présentant une tendance naturelle à se regrouper. Un certain nombre d'approches, les approches de clustering orienté requête (Tombros, 2002), proposent de considérer la requête de l'utilisateur dans le processus de clustering, affirmant qu'une prise en compte du contexte dans lequel la catégorisation des documents est effectuée constitue un facteur déterminant pour l'obtention d'un partitionnement adapté à l'utilisation que l'on souhaite en faire. Bien que ces approches n'aient pas été nécessairement conçues dans un tel but, elles peuvent, selon nous, permettre de produire des clusters mieux organisés autour de la requête de l'utilisateur et ainsi, lorsque celle-ci comporte un certain nombre d'aspects bien distincts, de distribuer l'information pertinente dans des clusters différents. Cependant, les mesures d'évaluations existantes, qui présentent une forte tendance à favoriser les systèmes rassemblant la majorité des documents pertinents dans un même cluster, ne permettent pas de mettre en évidence les bénéfices retirés d'une telle distribution de l'information pertinente. Après avoir présenté ces

2. Notons néanmoins qu'un second processus de clustering peut être appliqué à ce cluster particulier, permettant ainsi de faire émerger une certaine structure de l'information pertinente. Dans ce cas cependant, le problème précédent risque d'être amplifié par le fait que l'on risque de donner à l'utilisateur l'impression qu'il se trouve face à un système lui permettant de percevoir l'ensemble des aspects de son sujet, alors que les documents abordant réellement la requête sous un angle différent sont susceptibles d'être contenus par les autres clusters et donc de n'être pas représentés par les groupes qui lui sont présentés.

mesures d'évaluation existantes, nous proposons, en section 2, la mise en place de nouvelles mesures permettant une comparaison plus équitable des différents types de systèmes. En section 3, nous réalisons un certain nombre d'expérimentations visant à vérifier la validité des mesures proposées. Différents types de systèmes sont finalement comparés en utilisant nos mesures d'évaluation.

2. Évaluer l'accès à l'information

L'objectif d'un système réalisant une catégorisation de ses résultats est de proposer un partitionnement d'un ensemble ordonné $\mathcal{D} = \{D_1, \ldots, D_n\}$, des n premiers documents retournés par une recherche initiale, en k clusters $\mathcal{C} = \{C_1, \ldots, C_k\}$ tels que[3] :

$$\begin{cases} \forall i \in \{1, \ldots, k\}, C_i = \{C_i^1, \ldots, C_i^{|C_i|}\} \neq \emptyset, \\ \forall i, j \in \{1, \ldots, k\}^2, i \neq j \Rightarrow C_i \cap C_j = \emptyset \\ \bigcup_{i=1}^{k} C_i = \mathcal{D} \end{cases} \quad [1]$$

S'appuyant pour la plupart sur la *Cluster Hypothesis* (Manning *et al.*, 2008), ces systèmes sont généralement évalués sur leur capacité à regrouper l'ensemble des documents pertinents dans un même cluster. La mesure la plus fréquemment utilisée est alors la mesure *MK1*, proposée dans (Jardine *et al.*, 1971), qui consiste à juger la qualité d'un ensemble de groupes de documents en considérant uniquement le meilleur cluster qu'il contient. Le score attribué à un système par cette mesure dépend du nombre et de la proportion de documents pertinents dans le meilleur cluster que l'on puisse trouver parmi l'ensemble $\mathcal{C}$ des clusters produits par le système :

$$MK1(\mathcal{C}) = \min_{C_i \in \mathcal{C}} \Big(1 - \frac{(1+\beta^2) \times Precision(C_i) \times Rappel(C_i)}{\beta^2 \times Precision(C_i) + Rappel(C_i)}\Big)$$
$$\text{Où} \quad Precision(C_i) = \frac{|\mathcal{P}ert \cap C_i|}{|C_i|} \quad \text{et} \quad Rappel(C_i) = \frac{|\mathcal{P}ert \cap C_i|}{|\mathcal{P}ert|} \quad [2]$$

avec $\mathcal{P}ert \subseteq \mathcal{D}$ correspondant à l'ensemble des documents pertinents. Selon (Jardine *et al.*, 1971), la mesure *MK1* présente l'avantage d'isoler la qualité du clustering des biais induits par l'utilisation d'une stratégie de recherche particulière. Néanmoins, ne permettant pas l'évaluation de systèmes proposant une distribution de l'information pertinente, cette mesure ne peut pas être utilisée dans notre étude.

Afin d'évaluer les systèmes réalisant un clustering des résultats de la même façon que les systèmes classiques (en utilisant par exemple la mesure de précision moyenne), certaines approches ont proposé de reconstruire des listes ordonnées de documents à partir des clusters formés (Hearst *et al.*, 1996, Bellot *et al.*, 1999). Pour ce faire, les

3. Dans ce qui suit, $|A|$ correspond au cardinal de l'ensemble A.

approches commencent généralement par définir un ordre entre les clusters (couramment selon la proximité à la requête de leur document qui en est le plus proche ou bien la similarité moyenne de leurs documents à la requête) et entre les documents de chaque groupe (selon leur similarité avec la requête ou avec le représentant du groupe) pour obtenir un ensemble ordonné $\mathcal{C} = \{C_1, \ldots, C_k\}$ de k sous-ensembles ordonnés $C_i = \{C_i^1, \ldots, C_i^{|C_i|}\}$ (C_i^j correspond alors au j-ième document du i-ième cluster C_i). Les approches d'évaluation par reconstruction de listes ordonnées construisent leur liste finale $L = \{L_1, \ldots, L_n\}$ selon les chemins que l'utilisateur peut emprunter dans cette liste de listes définissant l'ensemble des parcours possibles. La seule contrainte sur la construction d'une telle liste concerne l'ordre dans lequel les documents d'un cluster sont examinés : le document C_i^{j+1} ne peut être d'indice inférieur à C_i^j dans la liste finale L (puisque sauf exceptions, l'utilisateur examine les documents dans l'ordre où ils sont listés).

Une fois la liste L produite, il est alors possible d'y appliquer une mesure d'évaluation classique, telle que la populaire mesure de précision moyenne (*Average Precision Ap*), qui correspond à la moyenne des précisions (proportion de documents pertinents dans un ensemble de documents) calculées après chaque document pertinent de la liste ordonnée. Si l'on dispose d'une fonction $Pert : \mathcal{D} \rightarrow \{0, 1\}$ retournant 1 si le document considéré est pertinent (0 sinon), la précision moyenne d'une liste $Ap : \mathcal{D} \rightarrow [0, 1]$ s'obtient par :

$$Ap(L) = \frac{1}{\sum_{i=1}^{n} Pert(L_i)} \times \sum_{i=1}^{n} \sum_{j=1}^{i} \frac{Pert(L_i) \times Pert(L_j)}{i} \qquad [3]$$

Cette mesure, qui considère le rang des éléments pertinents dans la liste produite, permet d'évaluer un parcours réalisé à travers les groupes proposés, et ainsi, de rendre compte de la capacité d'accès à l'information pertinente.

Les différences entre les approches d'évaluation par reconstruction de listes ordonnées résident alors dans les parcours réalisés à travers les clusters. Les deux approches de reconstruction de listes les plus répandues s'inspirent des parcours souvent réalisés dans les arbres de recherche : alors que le parcours en profondeur examine les clusters les uns après les autres en commençant par celui possédant le plus fort potentiel de pertinence, le parcours en largeur considère séquentiellement les premiers documents non encore examinés de chaque liste. Ces deux parcours, qui représentent les deux extrêmes de l'ensemble des chemins qu'un utilisateur peut emprunter, observent une corrélation inverse : alors que le parcours en profondeur favorise les systèmes regroupant l'ensemble des documents pertinents dans un même cluster, le parcours en largeur favorise les systèmes réalisant une distribution des documents pertinents sur l'ensemble des groupes. Bien que leur utilisation conjointe permet d'obtenir certaines indications sur les performances d'un système, ces deux parcours présentent, selon nous, un certain nombre de limites :

– L'amélioration du score d'une de ces deux évaluations tend à faire diminuer celui de l'autre. Cela rend difficile l'interprétation des résultats obtenus, d'autant plus que

l'équilibre de ces deux mesures n'est pas évident.

– La considération conjointe de ces deux critères pour comparer des clusterings produits par différentes méthodes tend à pénaliser ceux qui réalisent une distribution de l'information pertinente puisqu'il est plus difficile d'obtenir un score élevé selon un parcours en largeur que selon un parcours en profondeur.

– Les caractéristiques des partitionnements produits (tailles des clusters, degré de répartition de l'information pertinente) ont un fort impact sur les scores obtenus.

– Les scores obtenus dépendent fortement de l'ordre dans lequel sont présentés les clusters. Cela biaise les résultats puisque cet ordre n'a pas d'impact direct sur la route qu'un utilisateur emprunte à travers les clusters : à partir des descriptions présentées, l'utilisateur identifie les aspects qui semblent répondre au mieux à ses besoins et parcourt les clusters correspondants en priorité.

2.1. *Parcours moyen*

Au vu de ces différentes observations, nous cherchons à définir un compromis efficace entre ces deux parcours extrêmes, qui puisse permettre d'évaluer la capacité d'accès à l'information pertinente à partir d'un ensemble de groupes proposés, sans favoriser un type de système par rapport à un autre. Nous proposons alors de réaliser l'évaluation d'un partitionnement par considération du parcours "moyen" qu'il est possible d'effectuer à travers les groupes proposés. Cela revient à considérer l'espérance mathématique du score de précision moyenne pour un parcours effectué par un utilisateur "aveugle" (*i.e.,* qui n'oriente pas sa recherche selon les informations qu'il a déjà collectées dans les différents groupes). Étant donnée une fonction $exam : \{0, \ldots, n-1\} \times \{0, \ldots, |\mathcal{P}ert|\} \rightarrow \mathcal{D}$, définie telle que $exam(t, p)$ retourne le prochain document examiné après avoir déjà rencontré t documents dont p sont pertinents, cette espérance $ExpAP(\mathcal{C})$ peut être calculée par[4] :

$$\frac{\sum_{t=0}^{n-1} \sum_{i=1}^{k} \sum_{j=1}^{|C_i|} \sum_{p=0}^{|\mathcal{P}ert|} Pert(C_i^j) \times P(exam(t,p) = C_i^j) \times \frac{p+1}{t+1}}{|\mathcal{P}ert|} \quad [4]$$

Toute la difficulté réside alors dans le calcul de la probabilité $P(exam(t, p) = C_i^j)$. Il n'est en effet pas raisonnable de chercher à l'obtenir en testant, pour tous les documents et toutes les positions t et p envisageables, toutes les possibilités de parcours partiels permettant d'examiner C_i^j après t documents dont p pertinents. Afin de réduire le nombre de calculs à effectuer, nous choisissons alors de travailler avec des configurations $\vec{x}$, dont les composantes x_i (avec $i \in \{1, \ldots, k\}$) représentent les nombres de documents déjà examinés dans chaque cluster C_i (et respectent donc l'axiome suivant : $\forall i \in \{1, \ldots, k\}, x_i \leq |C_i|$). Avec $\mathcal{X}$ l'ensemble de toutes les configurations possibles, nous considérons alors une fonction $sel : \mathcal{X} \rightarrow \mathcal{C}$, de sélection du prochain

4. $P(A)$ dénote, de manière classique, la probabilité de l'évènement A.

cluster à parcourir à partir d'une configuration donnée, qui nous permet de définir la probabilité de sélectionner un cluster C_i (avec $i \in \{1, \ldots, k\}$) selon une configuration $\vec{x}$ de documents déjà examinés[5] :

$$P(sel(\vec{x}) = C_i) = \frac{1}{|\{j \in \{1, \ldots, k\}, x_j < |C_j|\}|} \quad [5]$$

De manière à évaluer la probabilité de rencontrer une configuration $\vec{x}$ donnée, nous définissons alors une notion de voisinage entre configurations par le biais d'une fonction $V : \mathcal{X} \times \mathcal{C} \rightarrow \mathcal{X}$ telle que $V(\vec{x}, C_i)$ représente la configuration permettant d'atteindre $\vec{x}$ en examinant le premier document non encore examiné dans le cluster C_i :

$$V(\vec{x}, C_i) = (y_1, \ldots, y_k) \mid y_i = x_i - 1 \ \wedge \ \forall j \in \{1, \ldots, k\}, j \neq i \Rightarrow y_j = x_j \quad [6]$$

Ainsi, la probabilité de rencontrer une configuration donnée $\vec{x}$ peut être évaluée par[6] :

$$P(\vec{x}) = \sum_{i \in \{1, \ldots, k\}, x_i > 0} P(V(\vec{x}, C_i)) \times P(sel(V(\vec{x}, C_i)) = C_i) \quad [7]$$

Il est alors possible de calculer la probabilité $P(exam(t, p) = C_i^j)$ qui nous intéresse :

$$P(exam(t, p) = C_i^j) = \sum_{\vec{x} \in config(C_i^j, t, p)} P(\vec{x}) \times P(sel(\vec{x}) = C_i) \quad [8]$$

où $config(C_i^j, t, p)$ correspond à l'ensemble des configurations permettant d'examiner C_i^j après avoir déjà rencontré t documents dont p pertinents, qui est définie par une fonction $config : \mathcal{D} \times \{0, \ldots, n-1\} \times \{0, \ldots, |\mathcal{P}ert|\} \rightarrow 2^{\mathcal{X}}$ telle que :

$$\begin{aligned} config(C_i^j, t, p) = \quad & \{(x_1, \ldots, x_k) \in \mathcal{X} \mid x_i = (j-1) \wedge \\ & \textstyle\sum_{l=1}^{k} x_l = t \ \wedge \ \sum_{a=1}^{k} \sum_{b=1}^{x_a} Pert(C_a^b) = p\} \end{aligned} \quad [9]$$

De tels calculs restent relativement complexes, mais leur emploi pour l'évaluation de partitionnements de 50 documents paraît tout à fait envisageable (autour de 10 secondes sur un *Pentium 4, 3GHz PC*). Par ailleurs, le score $ExpAP$ peut être estimé statistiquement, en effectuant des parcours aléatoires à travers les clusters (respectant les probabilités de sélection de cluster) et en calculant la moyenne des précisions moyennes obtenues sur les listes résultant de ces différents parcours. Des expérimentations[7] ont montré que cette estimation se révélait très proche du score réel que l'on aurait pu obtenir avec la formule 4, et très robuste pour des nombres de documents considérés supérieurs à 50, quels que soient le corpus utilisé et le nombre de clusters produits. Cette estimation peut alors être utilisée lorsque les calculs de l'espérance paraissent trop complexes.

5. Où $|\{j \in \{1, \ldots, k\}, x_j < |C_j|\}|$ correspond au nombre de groupes n'ayant pas encore été entièrement examinés dans la configuration $\vec{x}$.
6. Il est à noter que $P((0, \ldots, 0)) = 1$.
7. Les corpus utilisés dans ces expérimentations sont ceux décrits en section 3.

2.2. *Parcours orienté par la pertinence des documents*

Dans (Leuski, 2001), Leuski propose un compromis entre les deux extrêmes que représentent le parcours en profondeur et le parcours en largeur en considérant les jugements de pertinence établis lors de l'annotation du corpus. L'idée est de définir une stratégie de parcours qui tente de simuler le comportement qu'aurait pu avoir un utilisateur réel face aux différents clusters présentés par le système à évaluer : lorsque le nombre de documents non pertinents examinés dans un cluster dépasse le nombre de documents pertinents examinés dans ce même cluster, le processus stoppe le parcours de la liste correspondant au cluster courant pour s'intéresser à un autre groupe de documents. Le nouveau groupe choisi correspond alors au cluster dans lequel la meilleure proportion de documents pertinents a été observée sur les documents examinés. Cette stratégie simule le parcours qu'un utilisateur aurait pu emprunter en ce sens qu'elle se sert des éléments examinés pour orienter son parcours vers les clusters les plus susceptibles de contenir les informations pertinentes. La liste de documents finalement obtenue par le parcours réalisé est alors supposée mieux refléter la réalité que les parcours de clusters en profondeur ou en largeur communément employés, ce qui laisse augurer d'une meilleure évaluation du système. Néanmoins, si cette évaluation limite quelque peu les biais présentés par les parcours en profondeur et en largeur, elle favorise elle aussi les méthodes regroupant la plupart des documents pertinents dans un même groupe. En effet, l'intervalle existant entre le dernier examen d'un document pertinent dans un cluster et la prise de décision de changer de cluster implique des prises en compte de documents non pertinents. Or, le nombre de changements de clusters requis est plus important lorsque l'on considère un partitionnement où l'information pertinente est distribuée sur l'ensemble des groupes.

Les jugements de pertinence, utilisés pour orienter la stratégie de parcours proposée dans (Leuski, 2001), peuvent être inclus dans le calcul du score d'évaluation du parcours moyen afin de simuler un parcours réalisé par un utilisateur plus expérimenté. Ainsi, afin d'orienter plus probablement la recherche vers des clusters qui, selon leurs documents déjà examinés, présentent un fort ratio de documents pertinents, la probabilité $P(sel(\vec{x}) = C_i)$ peut être remplacée dans les formules 7 et 8 par :

$$\frac{0.5 + \sum_{j=1}^{x_i} Pert(C_i^j)}{x_i + 1} \Big/ \sum_{l \in \{1,\dots,k\}, x_l < |C_l|} \frac{0.5 + \sum_{j=1}^{x_l} Pert(C_l^j)}{x_l + 1} \qquad [10]$$

Au début du processus, tous les clusters possèdent la même probabilité de sélection. Selon le nombre et les positions des documents pertinents qu'ils contiennent, la probabilité de sélection de chaque cluster évolue au fur et à mesure que de nouveaux documents sont examinés. Contrairement à la stratégie de recherche proposée dans (Leuski, 2001), ce parcours ne requiert pas de déterminer un ordre entre les clusters. Les ratios de documents pertinents dans chaque cluster sont considérés en parallèle, ce qui permet d'écarter le biais, énoncé plus haut pour la stratégie de recherche proposée dans (Leuski, 2001), concernant l'intervalle existant entre le dernier examen d'un document pertinent dans un cluster donné et la prise de décision de changer de cluster.

2.3. *Parcours orienté par la proximité des documents pertinents*

Tel que c'est le cas pour la prise en compte des retours de pertinence, il est possible d'inclure une considération du contenu des documents examinés dans le calcul du score d'évaluation du parcours moyen. Il s'agit alors d'orienter plus probablement la recherche vers des clusters dont le contenu des documents examinés semble correspondre aux informations portées par les documents pertinents. La probabilité $P(sel(\vec{x}) = C_i)$ peut alors être remplacée, dans les formules 7 et 8, par :

$$\frac{\dfrac{0.5 + \sum_{j=1}^{x_i} \sum_{D \in \mathcal{P}ne(\vec{x})} \dfrac{Sim(D, C_i^j)}{|\mathcal{P}ne(\vec{x})|}}{x_i + 1}}{\sum_{l \in \{1,\ldots,k\}, x_l < |C_l|} \dfrac{0.5 + \sum_{j=1}^{x_l} \sum_{D \in \mathcal{P}ne(\vec{x})} \dfrac{Sim(D, C_l^j)}{|\mathcal{P}ne(\vec{x})|}}{x_l + 1}} \qquad [11]$$

où $\mathcal{P}ne(\vec{x})$ correspond à l'ensemble des documents pertinents n'ayant pas encore été examinés dans la configuration $\vec{x}$. Les probabilités de sélection des clusters dépendent alors des similarités moyennes des documents examinés dans chacun d'entre eux avec les documents pertinents n'ayant pas encore été rencontrés. Afin de travailler avec des valeurs de similarité suffisamment dispersées pour que les différences puissent influer sur la recherche, le processus d'évaluation commence par identifier les similarités minimales et maximales de chaque document pertinent et utilise ces valeurs pour répartir ses similarités sur $[0, 1]$. De la même façon qu'avec le parcours moyen utilisant des jugements de pertinence, tous les clusters possèdent initialement la même probabilité d'être sélectionnés. Ces probabilités évoluent ensuite selon le contenu des documents rencontrés dans chaque cluster. Le fait de ne considérer que les documents pertinents non rencontrés permet d'orienter la recherche vers d'autres clusters lorsque tous les documents pertinents connectés à la thématique d'un cluster donné ont déjà été examinés. Ce critère d'évaluation, fondé sur la proximité des pertinents plutôt que sur des retours binaires de pertinence, prend alors en compte les informations contenues par l'ensemble des documents, qu'ils soient pertinents ou non, pour orienter la recherche vers les clusters les plus pertinents, ce qui nous semble mieux correspondre aux comportements réels des utilisateurs.

Certains systèmes présentent les clusters de documents par des labels décrivant leur contenu. L'utilisateur peut alors s'appuyer sur ces descriptions pour choisir les clusters qui lui semblent le mieux correspondre à ses besoins. Pour effectuer une évaluation réaliste, il est envisageable, dans le cas de tels systèmes, d'inclure une prise en compte de la similarité de ces labels avec les documents pertinents dans les calculs des probabilités de sélection donnés par la formule 11. Par ailleurs, avec des systèmes décrivant le contenu des clusters en sélectionnant un document représentatif dans chacun d'entre eux, il est probable que les utilisateurs examinent l'ensemble des représentants de cluster avant de se lancer dans le parcours des groupes qui leur sont proposés. Les k représentants de cluster sont alors susceptibles d'être les k premiers documents à

être rencontrés (probablement dans l'ordre où ils sont présentés). Pour être plus réaliste, notre mesure d'évaluation peut alors inclure cette observation dans ses calculs, en fixant $P((1,\ldots,1)) = 1$ et en considérant $P(exam(i-1, \sum_{j=1}^{i-1} Pert(C_j^1)) = C_i^1) = 1$ pour chacun des représentants de cluster[8] (les autres probabilités d'examen sont fixées à 0 pour ces documents). Néanmoins, étant donné le faible impact qu'elle a sur les résultats finaux[9], une telle orientation de la recherche peut être omise.

3. Expérimentations

Quatre collections de documents tirées du corpus de la conférence TREC-1 ont été utilisées dans nos expérimentations. Le corpus ZIFF contient des articles informatiques, FR des extraits du registre fédéral des États Unis et AP et WSJ des articles de presse édités par l'Associated Press et le Wall Street Journal respectivement. Le même jeu de requêtes, les topics 1-50 de TREC, a été utilisé pour chaque corpus.

Les expérimentations présentées concernent le partitionnement des premiers documents retournés, en réponse à chaque requête, par le très populaire système Smart (Salton *et al.*, 1988). Deux algorithmes de clustering sont utilisés : les classiques méthodes K-Means (Tou *et al.*, 1974) et Group-Average (Tombros, 2002). Alors que la méthode K-Means produit une partition de l'ensemble de documents en optimisant la similarité de chaque élément avec le centre du cluster auquel il appartient, la méthode hiérarchique Group-Average construit un clustering en fusionnant, à chaque itération, les deux plus proches clusters jusqu'à ce que le nombre de clusters désiré soit atteint. Deux mesures de similarité inter-documents, qui s'appuient sur le modèle vectoriel (Baeza-Yates *et al.*, 1999) où les textes sont représentés par des vecteurs des poids de leurs termes significatifs[10], sont utilisées dans nos expérimentations : la mesure *Cosine* et la *Query-Sensitive Similarity Measure*. Selon la mesure *Cosine*, la similarité $Sim(D_i, D_j)$ de deux documents D_i et D_j est donnée par :

$$\frac{\sum_{l=1}^{t} w_{D_i,l} \times w_{D_j,l}}{\sqrt{\sum_{l=1}^{t} w_{D_i,l}^2 \times \sum_{l=1}^{t} w_{D_j,l}^2}} \text{ où } w_{D_i,l} = (1 + \ln(tf_{D_i,l})) \times \ln \frac{N}{n_l} \quad [12]$$

où t représente le nombre de termes uniques du corpus, $tf_{D_i,l}$ le nombre d'occurrences du terme l dans le document D_i et n_l le nombre de documents, parmi les N documents du corpus, dans lesquels le terme l apparaît.

La *Query Sensitive Similarity Measure* (*QSSM*), présentée dans (Tombros *et al.*, 2004), semble être l'approche de clustering orienté requête la plus performante. Cette mesure réalise, pour le calcul de la similarité entre deux documents, un produit entre

8. En supposant d'avoir placé les représentants en début de cluster : $\forall_{i\in\{1,\ldots,k\}}, Rep(C_i) = C_i^1$
9. Les premiers documents des clusters ont, dans tous les cas, une forte probabilité d'être examinés en début de recherche.
10. Ici, les termes sont les stemmes des mots porteurs de sens, obtenus après avoir supprimé les mots trop fréquents (stop-list) et avoir appliqué un processus de stemmatisation (Porter, 1980).

leur score de similarité thématique classique (en terme de *Cosine*) et un score de proximité à la requête du vecteur correspondant à l'intersection de leurs deux représentations vectorielles (où le poids de chaque terme correspond à la moyenne de ses poids dans les vecteurs individuels des deux documents). De cette manière, les documents contenant des termes de la requête différents, ce qui peut être considéré comme un indicateur de thématiques différentes, obtiennent un score de similarité minoré. À l'inverse, deux documents partageant un grand nombre de termes de la requête, et qui donc ont des chances de correspondre à un même aspect du sujet, voient leur similarité renforcée par cette prise en compte de leur proximité commune à la requête.

Quatre systèmes sont étudiés dans nos expérimentations :

– G, C : la méthode hiérarchique *Group-Average* utilisant la mesure *Cosine* ;

– K, C : la méthode *K-Means* utilisant la mesure *Cosine* ;

– G, Q : la méthode hiérarchique *Group-Average* utilisant la mesure *QSSM* ;

– K, Q : la méthode *K-Means* utilisant la mesure *QSSM*.

Dans les expérimentations réalisées, nous avons choisi d'ordonner les documents de chaque cluster selon leur similarité avec leur représentant (en terme de *Cosine*). Chaque représentant correspond au document le plus proche de la requête parmi les documents contenus dans le cluster concerné. Lorsque nécessaire, un ordre entre les groupes de documents est déterminé selon la similarité du représentant de chaque groupe avec la requête (en terme de *Cosine*). Dans l'ensemble de nos expérimentations, nous utiliserons les notations données par la table 1.

NbP	Nombre de représentants pertinents
$MK1$	Qualité du groupe optimal
PRO	Parcours en profondeur
LAR	Parcours en largeur
LEU	Parcours proposé dans (Leuski, 2001)
$PM1$	Parcours moyen
$PM2$	Parcours orienté par la pertinence des documents
$PM3$	Parcours orienté par la proximité des documents pertinents

Tableau 1. *Notations des mesures d'évaluation*

3.1. *Étude des mesures proposées*

3.1.1. *Degré de corrélation avec le comportement des utilisateurs*

L'objectif des mesures est de rendre compte de la capacité qu'aura un utilisateur à atteindre les documents qu'il recherche à partir de la liste de clusters qui lui est présentée. Pour être de bons indicateurs de cette capacité, les parcours de clusters doivent se rapprocher au maximum des routes qu'un utilisateur réel aurait pu suivre. Nous proposons alors ici d'évaluer la corrélation entre les scores obtenus par nos mesures

et ceux résultant de parcours d'utilisateurs réels. Pour ce faire, nous avons demandé à dix personnes volontaires de chercher à localiser le plus rapidement possible, pour 20 différentes requêtes, les documents pertinents en utilisant les partitionnements produits par nos quatre systèmes. Afin d'obtenir des résultats représentatifs de recherches réelles, nous leur avons expliqué que les différents groupes présentés étaient supposés représenter différents aspects de l'ensemble des documents considérés et qu'ils devaient alors lire chacun des documents qui leur seraient présentés pour orienter leur recherche vers les groupes qui leur semblent le plus probablement contenir les informations pertinentes. Ainsi, avant chaque examen de document, l'utilisateur doit choisir un index de groupe à explorer. Le premier document non encore examiné du cluster sélectionné lui est alors présenté et, selon le contenu de celui-ci, doit prendre la décision de continuer dans l'examen de ce cluster (le documents sont ordonnés dans les clusters selon leur similarité avec leur représentant), ou bien de passer à un autre groupe, qui lui semble mieux correspondre à la description de la requête qui lui a été fournie. En fin de processus, la moyenne des scores de précision moyenne obtenus sur les parcours suivis par les différents sujets est calculée pour chacune des requêtes. La corrélation entre les variations des scores obtenus par des utilisateurs réels et celles des scores calculés par les mesures d'évaluation[11] est estimée au moyen d'un coefficient $R^2_{o,m}$ dont les valeurs sont données, pour les différents systèmes, par la table 2.

$R^2_{o,m}$	PRO	LAR	LEU	PM1	PM2	PM3
G, C	0.85	0.80	0.89	0.83	0.92	0.92
K, C	0.51	0.76	0.62	0.81	0.86	0.90
G, Q	0.72	0.72	0.83	0.78	0.89	0.91
K, Q	0.47	0.71	0.55	0.74	0.81	0.88

Tableau 2. *Corrélation entre variations observées*

La méthode *Group-Average* conduit bien souvent à l'obtention d'un cluster contenant un très grand nombre de documents. Les choix proposés aux utilisateurs sont alors bien moins nombreux puisque les autres clusters sont souvent très vite épuisés et qu'il ne reste alors plus qu'un seul cluster contenant des documents non examinés. Il est alors bien plus aisé de suivre les variations observées. Dans ce cas, le fait que le parcours en profondeur obtienne un meilleur coefficient de corrélation que le parcours en largeur s'explique par le fait que les documents contenus par les plus petits clusters sont souvent bien marginaux et n'ont alors pas grand chose à voir avec le sujet de la recherche. Avec l'algorithme *K-Means* qui produit des clusters de tailles

11. Puisque l'objectif est d'évaluer la capacité des mesures à suivre les variations de scores des utilisateurs réels, et pas nécessairement d'obtenir les mêmes valeurs, chaque distribution de scores s (observés ou obtenus par une mesure) est normalisée selon la moyenne $\bar{s}$ et l'écart-type σ_s qu'elle observe sur l'ensemble des requêtes i : $\forall i \in \{1, \ldots, 20\}, s_i = (s_i - \bar{s})/\sigma_s$.

plus homogènes, il est plus difficile de suivre une route optimale puisque les documents pertinents sont mieux répartis dans les clusters. Cette difficulté est par ailleurs accrue lors de l'utilisation de la mesure *QSSM*[12]. Les coefficients de corrélation reportés montrent que notre parcours orienté par la pertinence des documents $PM2$, et *a fortiori* orienté par la proximité des documents pertinents $PM3$, parait bien mieux suivre les variations de score observées avec les individus impliqués dans l'étude que les autres approches d'évaluation quel que soit le système utilisé. Ils sont donc susceptibles d'être de meilleurs indicateurs de la capacité à atteindre les documents pertinents à partir d'une liste de clusters présentée à l'utilisateur.

3.1.2. *Influence du degré de répartition des documents pertinents*

Afin d'être capable de comparer les différents types de systèmes de manière équitable, le degré de concentration / distribution de l'information pertinente ne doit pas avoir un impact significatif sur les scores obtenus par les mesures d'évaluation utilisées. Dans le but de vérifier l'impartialité de nos mesures, nous proposons de réaliser des expérimentations sur des clusterings aléatoires présentant différents degrés de distribution des documents pertinents dans les clusters : après avoir aléatoirement distribué dans 5 clusters l'ensemble des documents non pertinents considérés, nous affectons chacun des documents pertinents au premier des cinq clusters avec une probabilité de 0.2, 0.33, 0.5, 0.66 ou 1 selon le type de clustering souhaité. Alors qu'une probabilité de 1 conduit à la production d'un cluster contenant la totalité des documents pertinents, une probabilité de 0.2 favorise la distribution de l'information pertinente. La figure 1 présente les différentes distributions de scores obtenus[13], par les mesures sur ces différents types de clustering.

Tel qu'attendu, l'espérance du parcours en profondeur augmente avec la probabilité d'affecter tous les documents pertinents dans un même cluster. L'espérance du parcours en largeur diminue dans le même temps. Conformément à notre intuition, le parcours proposé par (Leuski, 2001) semble largement favoriser les clusterings rassemblant tous les documents pertinents dans le même groupe. Selon les courbes présentées par la figure 1, nos parcours $PM2$ et $PM3$ paraissent constituer les mesures les plus équitables puisque leur espérance ne semble que très légèrement affectée par les variations du degré de répartition des documents pertinents dans les clusters. L'espérance du parcours $PM3$ semble néanmoins diminuer légèrement lorsque la concentration des documents pertinents augmente. Cela peut s'expliquer par le fait que les clusters considérés ne représentent pas de réelles thématiques puisqu'ils ont été produits de manière aléatoire. Ce parcours a alors des difficultés à déterminer les clusters les plus potentiellement intéressants à partir du contenu des premiers documents des cluster. Cette diminution d'espérance ne s'observe pas avec des partitionnements

12. Il est à noter que cette difficulté plus élevée ne traduit en rien une moins bonne efficacité du système, il ne s'agit ici que de la difficulté à trouver la meilleure route possible, et non de la difficulté accrue à atteindre les documents pertinents.
13. Les valeurs représentent des moyennes obtenues, pour chaque requête et chaque type de partitionnement, sur 100 clusterings des 50 premiers documents (du corpus AP) retournés. Les représentants de clusters sont ici sélectionnés aléatoirement.

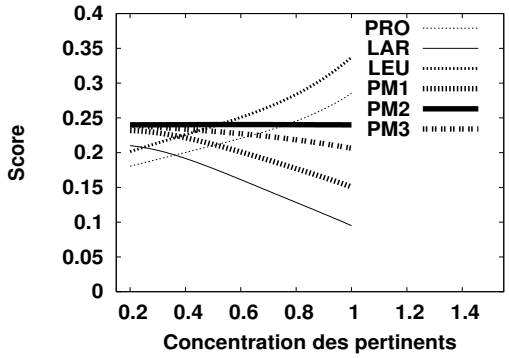

Figure 1. *Influence de la répartition des documents pertinents dans les clusters*

produits par des méthodes de clustering considérant les similarités existant entre les documents. Selon les résultats obtenus, on peut légitimement considérer les parcours $PM2$ et $PM3$ comme les plus à même de comparer des systèmes produisant des partitionnements de types différents[14].

3.2. *Comparaison des systèmes*

	ZIFF				AP				WSJ				FR			
	G,C	K,C	G,Q	K,Q	G,C	K,C	G,Q	K,Q	G,C	K,C	G,Q	K,Q	G,C	K,C	G,Q	K,Q
NbP	2.25	2.61	2.32	**2.69**	1.84	**2.38**	1.93	2.36	1.85	2.58	1.97	**2.59**	2.49	3.02	2.71	**3.21**
$MK1$	**0.62**	0.68	**0.62**	0.68	0.59	0.64	**0.58**	0.64	**0.66**	0.72	0.67	0.71	**0.55**	0.62	0.56	0.62
PRO	**0.54**	0.51	0.53	0.50	**0.55**	0.53	**0.55**	0.52	0.62	0.61	**0.63**	0.60	**0.54**	0.52	0.52	0.52
LAR	0.50	**0.54**	0.51	**0.54**	0.46	0.51	0.46	**0.52**	0.49	0.54	0.49	**0.56**	0.48	0.56	0.49	**0.58**
LEU	**0.55**	0.53	**0.55**	0.52	0.56	0.55	**0.57**	0.55	0.65	0.64	**0.66**	0.63	0.56	0.56	0.56	**0.57**
$PM1$	0.52	**0.55**	0.52	**0.55**	0.49	0.52	0.49	**0.53**	0.55	0.58	0.55	**0.61**	0.52	0.63	0.54	**0.65**
$PM2$	0.55	**0.57**	**0.57**	**0.57**	0.55	0.58	0.56	**0.59**	0.65	0.69	0.67	**0.70**	0.57	0.65	0.57	**0.67**
$PM3$	0.55	0.59	0.56	**0.60**	0.54	0.58	0.56	**0.61**	0.66	0.68	0.67	**0.70**	0.56	0.65	0.58	**0.71**

Tableau 3. *Résultats des systèmes*

La table 3 présente les résultats moyens obtenus par nos quatre systèmes réalisant une catégorisation, en cinq clusters, des 50 premiers documents retournés par la recherche initiale en réponse à chacune des requêtes. Alors que les résultats montrent que le système utilisant la méthode *Group-Average* obtient généralement le meilleur score selon la mesure *MK1*, ce système semble produire un clustering des documents à

14. Des expériences additionnelles ont par ailleurs montré que le nombre et la taille des clusters considérés n'ont pas non plus d'impact significatif sur les scores obtenus par nos mesures.

partir duquel l'accès aux documents pertinents est plus difficile qu'avec celui proposé par le système utilisant la méthode *K-Means*. En effet, alors que ce système obtient des résultats légèrement supérieurs selon un parcours en profondeur, nous observons une forte dominance de celui utilisant la méthode *K-Means* lorsque le parcours en largeur est considéré. Or, la stratégie de parcours LEU donne, bien souvent, de meilleurs scores au système utilisant la méthode Group Average : cette mesure, qui favorise les approches regroupant la plupart des documents pertinents, ne peut pas être utilisée pour comparer des systèmes produisant des clusterings de types différents. Nos mesures, qui comparent les systèmes de manière plus équitable, permettent de mettre en évidence, dans tous les cas, la supériorité des systèmes utilisant la méthode *K-Means*.

Par la considération des mesures d'évaluation existantes, l'utilisation de la mesure QSSM n'apporte pas d'amélioration significative des résultats. Les meilleurs scores obtenus par le parcours en largeur sont généralement minorés par une diminution du score obtenu par la recherche en profondeur. Il semble alors que l'augmentation de la capacité à regrouper l'ensemble des documents pertinents dans un même cluster, lorsque ceux-ci abordent le sujet de l'utilisateur sous un même angle, est contrebalancée par les pénalités que les évaluations attribuent à la mesure *QSSM*, pour avoir conduit à la distribution de l'information pertinente dans plusieurs groupes lorsque la requête comporte plusieurs aspects bien distincts. Les approches d'évaluation que nous proposons, et tout particulièrement le parcours orienté par la proximité des pertinents $PM3$, permettent de de mettre en valeur les bénéfices tirés de l'utilisation de mesures telle que la mesure *QSSM* : une prise en compte de la requête dans le processus de clustering peut effectivement permettre d'améliorer l'accès à l'information en mettant en évidence, lorsque le sujet de la requête est suffisamment large, différents aspects de l'information recherchée par l'utilisateur.

4. Conclusion

L'application de techniques de clustering sur les résultats d'une recherche d'information a pour but d'en faire émerger les thématiques principales. Cependant, le niveau de diversité des textes considérés implique bien souvent un faible degré de finesse du clustering réalisé et certaines thématiques émergentes peuvent se trouver en forte déconnection avec la requête formulée. La plupart des systèmes réalisant une catégorisation de leurs résultats y voient un effet bénéfique puisque cela permet de regrouper la plupart des textes pertinents dans un même cluster, donnant ainsi la possibilité à un utilisateur de filtrer les résultats retournés en ne parcourant que le cluster contenant les informations qui l'intéressent. Adoptant un point de vue différent, nous considérons ce phénomène comme largement négatif pour de multiples raisons, un tel regroupement de l'information pertinente ne permettant notamment pas de présenter à un utilisateur les différents aspects de sa requête et risquant alors de lui en restreindre la perception à un unique point de vue. Selon nous, une distribution de l'information pertinente sur l'ensemble des groupes formés peut s'avérer bien plus intéressante que sa concentration dans un unique cluster. Tout dépend du niveau d'accessibilité des textes pertinents

à partir de la liste de descriptions de clusters présentée. Or, nous nous sommes aperçus que la plupart des mesures d'évaluation présentaient une forte tendance à favoriser les systèmes regroupant l'ensemble des textes pertinents dans un même cluster. Nous avons alors cherché à établir des mesures réalisant une estimation plus équitable de la capacité d'accès à l'information pertinente. À la lumière des approches proposées, qui semblent refléter efficacement le comportement d'un utilisateur réel face à une liste de clusters, nous avons mis en évidence les bénéfices potentiels résultant d'une prise en compte de la requête dans le processus de clustering pour distribuer l'information pertinente dans des clusters distincts. Pour la première fois, la concentration de l'information pertinente n'est alors pas perçue comme le meilleur moyen d'aider l'utilisateur dans sa recherche. À ce titre, nous pensons que les mesures proposées ici sont susceptibles de remettre en cause nombre d'hypothèses concernant l'utilisation de techniques de clustering en recherche d'information.

5. Bibliographie

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

Bellot P., El-Bèze M., « Query Length, Number of Classes and Routes through Clusters : Experiments with a Clustering Method for Information Retrieval », *ICSC '99*, Springer-Verlag, London, UK, p. 196-205, 1999.

Croft W. B., « A model of cluster searching bases on classification. », *Information Systems*, vol. 5, n° 3, p. 189-195, 1980.

Hearst M. A., Pedersen J. O., « Reexamining the cluster hypothesis : Scatter/gather on retrieval results », *SIGIR'96*, Zürich, CH, p. 76-84, 1996.

Jardine N., Van Rijsbergen C. J., « The use of hierarchic clustering in information retrieval. », *Information Storage and Retrieval*, vol. 7, n° 5, p. 217-240, 1971.

Koshman S., Spink A., Jansen B. J., « Web searching on the vivisimo search engine », *Journal of the American Society for Information Science and Technology*, vol. 57, n° 14, p. 1875-1887, 2006.

Leuski A., « Evaluating document clustering for interactive information retrieval », *CIKM '01*, ACM, New York, NY, USA, p. 33-40, 2001.

Manning C. D., Raghavan P., Schütze H., *Introduction to Information Retrieval*, Cambridge University Press, 2008.

Porter M.F., « An algorithm for suffix stripping », *Program*, vol. 14, n° 3, p. 130-137, 1980.

Salton G., Buckley C., « Term-weighting approaches in automatic text retrieval », *Information Processing & Management*, vol. 24, n° 5, p. 513-523, 1988.

Tombros A., The Effectiveness of Query-Based Hierarchic Clustering of Documents for Information Retrieval, PhD thesis, University of Glasgow, UK, 2002.

Tombros A., Van Rijsbergen C. J., « Query-Sensitive Similarity Measures for Information Retrieval », *Knowledge Information Systems*, vol. 6, n° 5, p. 617-642, 2004.

Tou J. T., Gonzalez R. C., *Pattern recognition principles*, Applied Mathematics and Computation, Reading, Mass. : Addison-Wesley, 1974, 1974.