Organiser les résultats d'une recherche d'information

Clustering, répartition de l'information et facilité d'accès

Sylvain Lamprier* — Tassadit Amghar** — Bernard Levrat**
Frédéric Saubion**

* LIP6, Université Pierre et Marie Curie
4, place Jussieu F-75005 Paris
Sylvain.Lamprier@lip6.fr

** LERIA, Université d'Angers
2, Bd Lavoisier F-49000 Angers
{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.*

DOI:10.3166/DN.13.1.9-39 © 2010 Lavoisier, Paris

1. Introduction

Traditionnellement, 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 (Voorhees *et al.*, 1997). Bien que ce type d'organisation des résultats soit encore adopté par la majorité des systèmes de recherche actuels, il n'apparaît pas nécessairement comme le plus à même d'aider l'utilisateur à appréhender les différentes informations retournées. Il impose en effet de parcourir linéairement la liste de résultats, en examinant un à un chacun des documents, jusqu'à avoir le sentiment que suffisamment d'informations ont été collectées. Outre le fait qu'un tel parcours puisse bien souvent s'avérer fastidieux, tout le problème est de savoir quand s'arrêter : à partir de quel moment est-on certain d'avoir récolté assez d'informations ? L'utilisateur doit prendre la décision d'arrêter la collecte d'informations alors qu'il ne connaît pas la diversité des textes en relation avec sa requête. Des études statistiques ont montré que généralement les utilisateurs de systèmes employant ce type de présentation des résultats ne s'intéressent qu'aux tout premiers documents de la liste (Silverstein *et al.*, 1998; Spink *et al.*, 2001)¹, alors que, pour diverses raisons, certains documents pertinents peuvent être précédés dans la liste par des documents déconnectés du besoin de l'utilisateur.

Dans le but de réduire l'effort à fournir pour localiser les informations pertinentes, et permettre ainsi une meilleure perception des informations qui ont trait à la requête exprimée, de nombreuses approches ont proposé des modes de présentation alternatifs (Hearst *et al.*, 1996; Tombros, 2002). La plupart de ces approches cherchent à mettre en évidence 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 similaires². 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 (en termes d'une mesure de similarité thématique) 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.

Néanmoins, bien que la validité de la *Cluster Hypothesis* ait été vérifiée à maintes reprises (Van Rijsbergen, 1979; Cutting *et al.*, 1992; Hearst *et al.*, 1996), un certain nombre de travaux émettent de sérieux doutes quant à la capacité à diviser l'espace des documents en deux zones bien distinctes, comprenant l'ensemble des documents per-

1. Ces études ont montré qu'une écrasante majorité des utilisateurs n'examinent pas plus d'une dizaine de descriptions de documents, considérant alors que si les informations qu'ils recherchent ne se trouvent pas dans ces premiers documents examinés, ils ont peu de chances de les trouver dans les documents suivants.

2. Voir par exemple le métamoteur de recherche *Vivisimo* (Koshman *et al.*, 2006).

tinents d'un côté et celui des non pertinents d'un autre. Pour Salton et Buckley (Salton *et al.*, 1994), l'hypothèse selon laquelle la totalité des documents pertinents à une requête puissent être regroupés dans une zone particulière de l'espace semble largement optimiste, affirmant même qu'une telle configuration ne s'observe que très rarement. Certes, les documents pertinents tendent à être plus proches les uns des autres que des non pertinents mais cela reste une tendance et il est difficile d'affirmer que des documents non pertinents ne puissent pas se glisser dans la zone des documents pertinents (et réciproquement). En raison d'un style d'écriture particulier, par exemple, le vocabulaire d'un document pertinent peut ne posséder que de faibles relations de similarité avec les autres documents pertinents du corpus. Selon (Bellot, 2000), « d'un côté l'existence d'homonymes fait que deux documents peuvent sembler proches tout en traitant de thématiques différentes, et de l'autre, l'existence de synonymes peut faire paraître éloignés deux documents pourtant proches thématiquement ». D'après des expérimentations décrites dans (de Loupy *et al.*, 1999) ou (Marteau *et al.*, 1999), les documents pertinents sont plus généralement répartis dans plusieurs zones de l'espace, il est alors difficile de trouver une unique séparation efficace entre les différents documents du corpus. Non seulement les différences de vocabulaire peuvent conduire à affaiblir le niveau de similarité entre certains documents pertinents, mais il est également possible que l'ensemble des documents pertinents abordent la question sous différents points de vue, dans différents contextes ou même, traitent de différents aspects, bien distincts, du sujet qui intéresse l'utilisateur. L'obtention d'un groupe qui puisse à la fois rassembler la totalité des documents pertinents tout en excluant l'ensemble des non pertinents est alors assez improbable.

En outre, alors que la plupart des systèmes réalisant une catégorisation des résultats considèrent la *Cluster Hypothesis* comme un phénomène largement bénéfique, et s'y appuient alors 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. La majorité des documents pertinents tendant à se regrouper dans un seul et même cluster, $k - 1$ des k clusters présentés sont supposés être largement déconnectés de la requête formulée, ce qui pose, selon nous, un certain nombre de problèmes.

Tout d'abord, la majorité des informations initialement présentées à l'écran risquent de ne pas correspondre aux besoins de l'utilisateur. Il suffit alors que la description du cluster contenant l'ensemble des documents pertinents n'ait pas été judicieusement choisie pour que l'utilisateur soit incapable de localiser les informations qu'il recherche. De plus, quand bien même cette description s'avère intéressante, 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, tel que nous venons de l'évoquer, la *Cluster Hypothesis* ne se vérifie pas toujours pleinement, elle ne fait qu'énoncer des tendances observées. Ainsi les plus fortes similarités entre documents pertinents ne concernent pas toujours la tota-

lité de ces documents. De plus, les techniques de clustering employées n'étant pas toujours infaillibles, certains documents, mêmes proches des autres documents pertinents, peuvent se retrouver dans un cluster différent. Le fait qu'une majorité de documents pertinents appartiennent à un même groupe conduit les documents pertinents malencontreusement contenus dans les autres clusters à être bien isolés parmi des documents non pertinents. Ces documents isolés 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 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. En tentant de regrouper l'ensemble des documents pertinents dans un même cluster, on prend le risque de restreindre les informations portées à la connaissance de l'utilisateur à un seul point de vue. 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 assez d'informations. Un problème majeur se pose alors : l'utilisateur doit prendre la décision d'arrêter la collecte d'informations alors qu'il ne connaît pas la diversité des textes en relation avec sa requête³.

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

3. 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.

Figure 1. *Concentration vs. distribution de l'information pertinente*

meilleur choix⁴. La figure 1 illustre deux clusterings d'un même ensemble de documents, les documents pertinents étant représentés par des points noirs, les non pertinents par des points blancs et la requête par une étoile (les distances entre éléments correspondent à leurs distances thématiques) : vaut-il mieux, à l'instar de la plupart des systèmes, chercher à concentrer l'ensemble des documents pertinents dans un seul et même cluster (clustering de gauche), ou bien plutôt tenter de distribuer l'information pertinente sur l'ensemble des groupes afin d'en faire émerger la pluralité des aspects (clustering de droite) ? Nous pensons que, dans bien des cas, une distribution de l'information pertinente peut s'avérer préférable à sa concentration dans un unique cluster. La *Cluster Hypothesis* se pose alors 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. Selon le sujet qui intéresse l'utilisateur, la catégorisation la plus adaptée d'un même ensemble de documents n'est en effet pas nécessairement la même. Ainsi, (Tombros *et al.*, 2003) proposent par exemple de ne considérer, lors du calcul de la similarité de deux documents, que leurs phrases les plus proches de la requête. Cela permet de ne s'intéresser

4. Cette observation se limite bien entendu au cadre de la présente étude, à savoir la re-organisation des résultats d'une recherche d'information, et ne concerne pas les autres applications du clustering en recherche d'information (*e.g.*, *cluster-based retrieval* (Salton, 1971), modèles de langue (Liu *et al.*, 2004), exploration de corpus (Cutting *et al.*, 1992; McKeown *et al.*, 2002), etc.), pour lesquelles la *Cluster Hypothesis* peut s'avérer un phénomène très bénéfique.

qu'à des indices de proximité thématique correspondant au sujet qui intéresse l'utilisateur et ainsi de former des clusters susceptibles de mieux correspondre à ses attentes. Néanmoins, cette approche prend le risque d'ignorer de nombreuses parties de texte qui, bien que ne contenant pas les termes de la requête, auraient pu être utiles au rapprochement de documents aux thématiques fortement connectées. D'autres approches, telles que celles présentées dans (Chang *et al.*, 1997), (Iwayama, 2000) ou (Tombros *et al.*, 2004), proposent d'inclure une prise en compte de la requête dans les calculs de similarité entre documents. La *Query Sensitive Similarity Measure (QSSM)*, présentée dans (Tombros *et al.*, 2004), semble être l'approche la plus performante d'entre elles. Contrairement aux autres approches, qui modifient simplement les similarités en augmentant les poids des termes de la requête dans les représentations des documents, la mesure *QSSM* considère, pour le calcul de la similarité entre deux documents, un score de proximité commune à la requête (voir section 3). 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. Bien que ces approches, y compris cette mesure *QSSM* que nous étudions par la suite, aient été pour la plupart conçues dans le but d'augmenter les capacités des méthodes de clustering à regrouper l'ensemble des documents pertinents dans un même cluster, 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 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 $\mathcal{D} = \{D_1, \dots, D_n\}$, des n premiers

documents retournés par une recherche initiale, en k clusters $\mathcal{C} = \{C_1, \dots, C_k\}$ tels que⁵ :

$$\left\{ \begin{array}{l} \forall i \in \{1, \dots, k\}, C_i = \{D_1^i, \dots, D_{n_i}^i\} \neq \emptyset, \text{ avec } n_i = |C_i|, \\ \forall i, j \in \{1, \dots, k\}^2, i \neq j \Rightarrow C_i \cap C_j = \emptyset \\ \bigcup_{i=1}^k C_i = \mathcal{D} \end{array} \right. \quad [1]$$

Alors que l'évaluation des systèmes présentant une liste ordonnée de documents peut se faire en considérant les positions des documents pertinents dans cette liste, la caractérisation des performances des systèmes réalisant une catégorisation des documents peut s'avérer plus difficile. Bien sûr, il est possible de juger la qualité d'une méthode de clustering en considérant par exemple des critères de cohésion ou de séparation des groupes formés, mais ce qui nous intéresse surtout ici est l'accessibilité à l'information pertinente qu'offre le système évalué. Il s'agit alors d'évaluer la capacité qu'aura l'utilisateur à trouver les informations qui l'intéressent dans la partition présentée, ce qui semble plus difficile puisque de nombreux facteurs entrent en ligne de compte. Contrairement aux systèmes classiques où les documents sont présentés séquentiellement, les systèmes produisant des groupes de documents proposent un parcours des résultats plus complexe. On ne connaît alors pas *a priori* l'ordre dans lequel les documents seront présentés à l'utilisateur, puisque cela dépend de ses choix de clusters successifs.

2.1. Approches d'évaluation existantes

S'appuyant pour la plupart sur la *Cluster Hypothesis* (Manning *et al.*, 2008), les systèmes réalisant une catégorisation des résultats sont généralement évalués sur leur capacité à regrouper l'ensemble des documents pertinents dans un même cluster. L'approche d'évaluation la plus fréquemment utilisée est alors la mesure *MKI*, proposée dans (Jardine *et al.*, 1971), qui considère le nombre et 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 :

$$MKI(\mathcal{C}) = \min_{C_i \in \mathcal{C}} \left(1 - \frac{2 \times Precision(C_i) \times Rappel(C_i)}{Precision(C_i) + Rappel(C_i)} \right) \quad [2]$$

$$\text{Où } Precision(C_i) = \frac{|\mathcal{P} \cap C_i|}{|C_i|} \text{ et } Rappel(C_i) = \frac{|\mathcal{P} \cap C_i|}{|\mathcal{P}|}$$

avec $\mathcal{P} \subseteq \mathcal{D}$ correspondant à l'ensemble des documents pertinents. Selon (Jardine *et al.*, 1971), la mesure *MKI* 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.

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

Figure 2. *Parcours à travers les clusters*

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), de nombreuses approches ont proposé de reconstruire des listes ordonnées de documents à partir des clusters formés (Hearst *et al.*, 1996; Silverstein *et al.*, 1997; Bellot *et al.*, 1999). Pour ce faire, ces 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 $\mathcal{C} = \{C_1, \dots, C_k\}$ de k sous-ensembles ordonnés $C_i = \{D_1^i, \dots, D_{n_i}^i\}$ (D_j^i correspond alors au j -ième document du i -ième cluster C_i). Cette liste de listes (ou ensembles ordonnés) définit l'ensemble des chemins que l'utilisateur peut emprunter à travers les groupes présentés : les clusters, représentés par leur document le plus proche de la requête, sont présentés par ordre de potentiel de pertinence et l'utilisateur les consulte pour sélectionner le groupe qui lui semble le plus adapté à son besoin. Les documents de ce cluster sont présentés par ordre de potentiel de pertinence (ou de similarité avec le document représentant du groupe) et l'utilisateur peut les examiner les uns après les autres (tel que c'est le cas avec la liste de résultats présentée par les systèmes classiques) jusqu'à avoir trouvé ce qu'il recherche ou décidé que les informations correspondant à ses besoins sont susceptibles d'être localisées dans un autre cluster, auquel cas il choisit un autre groupe, examine les documents contenus dans ce groupe, etc. Les approches d'évaluation par reconstruction de listes ordonnées construisent leur liste finale $L = \{L_1, \dots, L_n\}$ selon les chemins que l'utilisateur peut emprunter dans les sous-ensembles ordonnés C_i correspondant aux k différents clusters. La seule contrainte sur la construction d'une telle liste concerne l'ordre dans lequel les documents d'un cluster sont examinés : le document D_{j+1}^i ne peut être d'indice inférieur à D_j^i dans la liste finale L (puisque logiquement l'utilisateur examine les documents dans l'ordre où ils sont listés). La figure 2 illustre deux parcours à travers les clusters, l'un valide, l'autre non valide. Les différents documents y sont représentés par des lettres (de A à L) et les trois clusters proposés y sont disposés horizontalement. Les groupes sont présentés à l'utilisateur par ordre d'estimation de pertinence (de haut en bas). Les documents sont ordonnés à l'intérieur de chaque cluster (de gauche à droite)

Figure 3. *Parcours en profondeur et parcours en largeur*

selon leur estimation de pertinence ou similarité avec le document représentant de leur groupe (A, E ou H). Le parcours de droite est invalide puisqu'il considère le quatrième document du premier cluster (D) avant le troisième (C), alors que l'utilisateur est supposé les examiner dans l'ordre.

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 :

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

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 : le parcours en profondeur examine les clusters les uns après les autres en commençant par celui possédant le plus fort potentiel de pertinence, considérant alors la liste entière des documents de chaque cluster avant d'en changer, et le parcours en largeur considère séquentiellement les premiers documents non encore examinés de chaque liste, changeant de cluster après chaque examen de document. La figure 3 est une illustration de ces deux différents parcours de listes. Ces deux parcours représentent les deux extrêmes de l'ensemble des chemins qu'un utilisateur peut emprunter. Ce que l'on évalue dans les deux cas est la facilité d'accès aux documents pertinents, déterminée par leur distribution et leur position dans les listes des clusters. Alors que l'évaluation d'une liste résultant d'un parcours en profondeur revient à considérer dans quelle mesure le système a réussi à regrouper

l'ensemble de documents pertinents dans un même groupe, celle de la liste résultant d'un parcours en largeur traduit la répartition des documents pertinents dans les différents clusters et leur positionnement en tête de listes. Ces deux parcours 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. Il est alors difficile de déterminer la dominance d'un partitionnement sur un autre lorsque l'un des deux obtient un meilleur score par un parcours en profondeur et l'autre un meilleur score par un parcours en largeur ;

- la considération conjointe de ces deux critères pour comparer des clusterings produits par différentes méthodes tend à pénaliser ceux qui observent une distribution de l'information pertinente sur l'ensemble des clusters puisqu'il est plus difficile d'obtenir un score élevé selon un parcours en largeur que selon un parcours en profondeur. En effet, même si les différents groupes produits par une méthode donnée sont bien représentatifs des différents aspects de la requête formulée par l'utilisateur, il se peut que certains d'entre eux ne correspondent pas à ce qu'attendaient les annotateurs lors de l'établissement des jugements de pertinence des documents. Or, l'existence d'un seul groupe ne contenant aucun document pertinent suffit à réduire considérablement le score obtenu par un parcours en largeur. Par ailleurs, le recul d'un document pertinent d'une seule position dans la liste d'un cluster (ce qui correspond à une perte mineure) implique son recul de k positions dans la liste finale représentant un parcours en largeur (alors qu'il ne perd qu'une seule place dans un parcours en profondeur) ;

- les caractéristiques des partitionnements produits (tailles des clusters, degré de distribution de l'information pertinente) ont un fort impact sur les scores obtenus. Par exemple, les documents non pertinents situés en fin de cluster, qui ne devraient avoir aucun effet sur l'évaluation puisqu'ils n'ont que très peu de chance d'être examinés par l'utilisateur, ont un effet très négatif sur le score obtenu par un parcours en profondeur. Parallèlement, les écarts entre les nombres de documents pertinents contenus par chaque cluster ont un effet très négatif sur le score obtenu par un parcours en largeur : si un cluster contient plus de documents pertinents que les autres, un parcours en largeur risque de considérer ses derniers documents pertinents après certains documents non pertinents des autres groupes (même si ces documents pertinents ne sont pas classés après des documents non pertinents dans leur propre groupe) ;

- les scores obtenus dépendent fortement de l'ordre dans lequel sont présentés les clusters. Or, tandis que l'ordre dans lequel sont présentés les représentants des clusters peut avoir de l'importance (puisque l'utilisateur les examine logiquement dans l'ordre dans lequel ils lui sont présentés), cet ordre établi n'a pas d'impact direct sur

la route que l'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. Par exemple, le fait qu'un cluster soit présenté en cinquième position n'implique pas que le contenu de ce cluster sera examiné en cinquième position, tout dépend de l'intérêt que l'utilisateur lui porte (contrairement à l'ordre des documents dans les clusters que l'utilisateur est supposé suivre, puisqu'il n'y a pas de raison qu'il commence à examiner la fin de la liste présentée avant son début).

Dans (Leuski, 2001), un compromis est proposé 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 (en cas d'égalité le cluster choisi est celui dont le premier document non encore examiné possède la meilleure estimation de pertinence). Cette stratégie, qui peut être rapprochée des processus d'expansion de requêtes par réinjection de pertinence, 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 système⁶.

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 lors de parcours de listes correspondant à un partitionnement produit par une méthode cherchant à distribuer l'information pertinente sur l'ensemble des groupes. Le nombre de documents non pertinents considérés avant des documents pertinents est alors plus élevé, ce qui nuit au score d'évaluation attribué à la méthode. Par exemple, les partitionnements présentés par la figure 4, où les 1 représentent des documents pertinents et les 0 des documents non pertinents, obtiennent avec cette stratégie de parcours le même score

6. Notons qu'en utilisant cette stratégie de parcours, Leuski a montré que les systèmes réalisant une catégorisation des résultats pouvaient s'avérer au moins aussi performants que la plupart des systèmes utilisant des retours de pertinence (*Relevance Feedbacks*) pour enrichir la requête formulée par l'utilisateur (Leuski, 2001).

Figure 4. Deux partitionnements de 18 documents en 5 clusters

qu'avec un parcours en profondeur (le score de précision moyenne est de 1 pour le clustering 1 et de 0.662 pour le clustering 2), ce qui est alors largement en faveur de la méthode 1 qui regroupe tous les documents pertinents dans un même cluster. Enfin, ce parcours est lui aussi fortement dépendant de l'ordre dans lequel sont présentés les clusters, ce qui peut grandement biaiser les évaluations réalisées.

Au vu de ces différentes observations, nous proposons la mise en place d'un certain nombre de mesures qui visent à adopter un point de vue plus global lors de l'évaluation des systèmes proposant une catégorisation des résultats, cherchant à évaluer l'ensemble de groupes plutôt qu'un groupe en particulier, sans favoriser un type de système par rapport à un autre. L'objectif est alors d'estimer les capacités réelles qu'aurait un utilisateur à atteindre les informations pertinentes à partir de la liste de clusters qui lui est présentée.

2.2. Parcours optimal

Face à ces limites identifiées, nous avons dans un premier temps envisagé l'introduction d'une notion de « parcours optimal », qui correspond à la meilleure liste (*i.e.*, conduisant au meilleur score de précision moyenne) qu'il est possible de construire à partir de la liste de clusters produite par le système à évaluer. Cette liste représente le parcours effectué par un utilisateur très « chanceux », qui réalise les meilleurs changements de cluster au meilleur moment sans avoir aucune information sur les documents qu'il reste à examiner dans les différents groupes. Bien sûr, puisque très improbable, ce parcours n'est pas réaliste mais, dans un contexte d'évaluation, il peut permettre de comparer différents partitionnements de manière relativement équitable, en se basant sur le potentiel d'efficacité de chacun.

Contrairement aux approches de reconstruction de listes existantes qui requièrent la détermination d'un ordre entre les clusters, la seule contrainte imposée à ce parcours concerne l'ordre des documents à l'intérieur de chaque cluster. La liste optimale

correspond alors au meilleur enchevêtrement de listes C_i (ou inter-classement) qu'il est possible de produire en considérant un critère de précision moyenne (formule [3]).

Algorithme 1 : Algorithme d'évaluation par construction de la liste optimale

Données : Un ensemble ordonné $\mathcal{C}$ de k listes C_i contenant chacune n_i documents D_1^i à $D_{n_i}^i$ (avec un total de n documents).

Résultat : Le score O de précision moyenne du parcours optimal.

```

1  début
2  |  $O = 0$ ;  $L = ()$ ;  $L' = ()$ ;  $L'' = ()$ ;  $stable = 0$ ;  $i = 1$ ;  $j = 1$ ;  $pos = 0$ ;
3  | Concaténation des  $k$  listes  $C_i$  :  $L = C_1.C_2 \dots C_{k-1}.C_k$ ;
4  |  $O = Ap(L)$ ; /* Formule [3] */
5  | tant que  $stable = 0$  faire
6  | |  $stable = 1$ ;
7  | | pour  $i$  de 1 à  $k$  faire
8  | | | pour  $j$  de 1 à  $n_i$  faire
9  | | | |  $L'' = L$ ;
10 | | | |  $pos =$  Position de  $D_j^i$  dans  $L$ ;
11 | | | | pour  $step$  de 1 à  $(n - pos)$  faire
12 | | | | | pour  $cur$  de 1 à  $n$  faire
13 | | | | | |  $L'_{Move(L,cur,pos,step)} = L_{cur}$ ; /* Formule [4] */
14 | | | | | fin
15 | | | | | si  $Ap(L') > O$  alors
16 | | | | | |  $O = Ap(L')$ ;  $L'' = L'$ ;  $stable = 0$ ;
17 | | | | | fin
18 | | | | fin
19 | | | |  $L = L''$ ;
20 | | | fin
21 | | fin
22 | fin
23 fin
  
```

Cette liste optimale ne peut pas être obtenue en énumérant tous les enchevêtrements de listes possibles (il en existe $n! / \prod_{i=1}^k n_i!$). Néanmoins, elle peut être construite relativement efficacement en concaténant les k listes C_i et en y réalisant des permutations de fragments de listes permettant une augmentation du score de précision moyenne tout en respectant la contrainte d'ordre entre les documents d'un même groupe (la liste finale L ne peut contenir un document D_{j+1}^i d'indice inférieur à un document D_j^i). L'algorithme 1 décrit un processus d'évaluation par construction de la liste optimale. Ce processus utilise une fonction $Move(L, cur, pos, step)$ qui retourne

la nouvelle position d'un élément L_{cur} dans la liste L après avoir essayé de permuter l'élément L_{pos} avec les $step$ éléments qui le suivent dans la liste L :

$$\begin{aligned} Move(L, cur, pos, step) = \\ \text{Si } cur \notin [pos, pos + step] \text{ alors } cur; \\ \text{Sinon } cur + Cl(L_{pos}, L_{cur}) \times (\sum_{x=cur}^{pos+step} 1 - Cl(L_{pos}, L_x)) \\ - (1 - Cl(L_{pos}, L_{cur})) \times (\sum_{x=pos}^{cur} Cl(L_{pos}, L_x)). \end{aligned} \quad [4]$$

avec $Cl : \mathcal{D} \times \mathcal{D} \rightarrow \{0, 1\}$ une fonction qui retourne 1 si les deux documents concernés appartiennent au même cluster. Si L_{cur} n'est pas situé entre les positions pos et $pos + step$, sa position reste inchangée. Sinon, si il appartient au même groupe que L_{pos} , il est déplacé pour être positionné après tous les éléments n'appartenant pas à leur cluster et étant situés entre sa position cur et la position $pos + step$. Enfin, si L_{cur} n'appartient pas au même groupe que L_{pos} , il est avancé dans la liste pour se positionner devant tous les éléments appartenant au groupe de L_{pos} et étant situés entre les positions pos et cur . Un tel processus converge vers la liste optimale en un temps raisonnable : dans l'ensemble de nos expérimentations (voir section 3.2), l'algorithme n'a jamais requis plus de 3 itérations pour atteindre une liste stable.

Cette évaluation par construction de la liste optimale, qui ne requiert pas la détermination d'un ordre entre les différents clusters mais considère le meilleur parcours qu'il est possible d'effectuer à travers les clusters d'un partitionnement proposé, isole les performances du système considéré des biais pouvant résulter du choix d'une stratégie de parcours particulière. Elle permet alors une meilleure interprétation des résultats. Néanmoins, cette évaluation est dépendante du nombre de clusters produits par le système considéré. Par exemple, une méthode de clustering produisant autant de clusters que de documents à catégoriser ($k = n$), obtiendrait un score égal à 1. Plus le nombre de clusters produits est important, plus le score d'évaluation risque d'être élevé. Par conséquent, cette mesure ne peut être utilisée que pour comparer des systèmes présentant un même nombre de clusters à l'utilisateur. Par ailleurs, les caractéristiques du partitionnement évalué ont un impact assez important sur l'espérance du score attribué par la mesure : une méthode de clustering produisant des clusters de tailles homogènes a plus de chances d'obtenir un bon score d'évaluation qu'une méthode ayant tendance à regrouper la majorité des documents dans un même groupe. Enfin, un système ayant obtenu un score inférieur à un autre selon cette évaluation par construction de la liste optimale, peut très bien permettre un plus grand nombre de parcours efficaces et être alors au moins aussi intéressant pour l'utilisateur. Cette évaluation n'est donc pas exempte de biais.

2.3. Parcours moyens

Alternativement, nous proposons 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 utilisateur⁷.

Étant donnée une fonction $exam : \{0, \dots, n-1\} \times \{0, \dots, |Pert|\} \rightarrow \mathcal{D}$, définie telle que $exam(t, p)$ retourne le prochain document examiné après avoir déjà rencontré t documents dont p sont pertinents⁸, cette espérance $ExpAP(\mathcal{C})$ peut être calculée de la manière suivante⁹ :

$$\frac{\sum_{t=0}^{n-1} \sum_{i=1}^k \sum_{j=1}^{|C_i|} \sum_{p=0}^{|Pert|} Pert(D_j^i) \times P(exam(t, p) = D_j^i) \times \frac{p+1}{t+1}}{|Pert|} \quad [5]$$

Toute la difficulté réside alors dans le calcul de la probabilité $P(exam(t, p) = D_j^i)$. 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 D_j^i 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 $i \in \{1, \dots, k\}$) représentent les nombres de documents déjà examinés dans chaque cluster C_i (et respectent donc l'axiome suivant : $\forall i \in \{1, \dots, k\}, x_i \leq |C_i|$). 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 cluster à parcourir à partir d'une configuration donnée, qui nous permet de définir la probabilité de sélectionner un cluster C_i (avec $i \in \{1, \dots, k\}$) selon une configuration $\vec{x}$ de documents déjà examinés¹⁰ :

$$P(sel(\vec{x}) = C_i) = \frac{1}{|\{j \in \{1, \dots, k\}, x_j < n_j\}|} \quad [6]$$

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 :

$$V(\vec{x}, C_i) = (y_1, \dots, y_k) \mid y_i = x_i - 1 \wedge \forall j \in \{1, \dots, k\}, j \neq i \Rightarrow y_j = x_j \quad [7]$$

Ainsi, la probabilité de rencontrer une configuration donnée $\vec{x}$ peut être évaluée par¹¹ :

$$P(\vec{x}) = \sum_{i \in \{1, \dots, k\}, x_i > 0} P(V(\vec{x}, C_i)) \times P(sel(V(\vec{x}, C_i)) = C_i) \quad [8]$$

7. Dans un premier temps, nous considérons un utilisateur « aveugle » (i.e., qui n'oriente pas sa recherche selon les informations qu'il a déjà collectées dans les différents groupes).

8. $Pert \subseteq \mathcal{D}$ représente le sous-ensemble des documents pertinents contenus dans l'ensemble $\mathcal{D}$ des n documents partitionnés. La fonction $Pert : \mathcal{D} \rightarrow \{0, 1\}$, qui lui est associée, retourne 1 si le document considéré appartient à ce sous-ensemble (et 0 sinon).

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

10. Où $|\{j \in \{1, \dots, k\}, x_j < n_j\}|$ correspond au nombre de groupes n'ayant pas encore été entièrement examinés dans la configuration $\vec{x}$.

11. Il est à noter que $P((0, \dots, 0)) = 1$.

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

$$P(exam(t, p) = D_j^i) = \sum_{\vec{x} \in config(D_j^i, t, p)} P(\vec{x}) \times P(sel(\vec{x}) = C_i) \quad [9]$$

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

$$config(D_j^i, t, p) = \{(x_1, \dots, x_k) \in \mathcal{X} \mid x_i = (j-1) \wedge \sum_{l=1}^k x_l = t \wedge \sum_{a=1}^k \sum_{b=1}^{x_a} Pert(D_b^a) = p\} \quad [10]$$

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, 3 GHz PC*). Par ailleurs, le score *ExpAP* peut être estimé statistiquement, en effectuant des parcours aléatoires à travers les clusters¹² et en calculant la moyenne des précisions moyennes obtenues sur les listes résultant de ces différents parcours. Des expérimentations¹³ ont montré que cette approximation se révélait très proche du score réel que l'on aurait pu obtenir avec la formule [5] et très robuste pour des nombres de documents considérés supérieurs à 50, quels que soient le corpus utilisé et le nombre de clusters produits : le plus grand écart enregistré entre deux estimations d'une même instance (100 estimations, comprenant chacune 1000 parcours aléatoires, ont été réalisées pour chaque instance) est égal à 0.0012 pour des clusterings aléatoires de 100 documents et de 0.0008 pour des clusterings aléatoires de 500. Étant donnés les écarts observés entre scores obtenus sur des instances différentes, ces écarts paraissent tout à fait négligeables. Cette estimation statistique peut alors être utilisée lorsque les calculs de l'espérance de précision moyenne d'un parcours de clusters paraissent trop complexes.

Cette évaluation par calcul de l'espérance du score de précision moyenne pour un partitionnement donné est indépendante du nombre de clusters produits par le système considéré (vérifié expérimentalement en section 3.1.3). En effet, si le nombre de routes efficaces augmente avec le nombre de clusters produits, le nombre de mauvaises routes augmente aussi, ce qui permet de trouver un équilibre qui autorise la comparaison de scores obtenus sur des partitionnements aux nombres de clusters différents. Avec une telle évaluation, seules comptent les positions des documents pertinents dans les listes correspondant aux clusters considérés. Le parcours moyen, qui résume l'ensemble des parcours possibles, représente un bon compromis entre parcours en profondeur et parcours en largeur. Par ailleurs, une telle mesure ne souffre pas des biais induits par la

12. Afin de respecter les différentes probabilités, plutôt que de choisir une route donnée parmi l'ensemble des parcours possibles, le processus construit le parcours itérativement, en sélectionnant, à chaque itération et selon la probabilité de sélection donnée par la formule [6], un cluster dans lequel examiner le premier document non encore rencontré.

13. Les corpus utilisés dans ces expérimentations sont ceux décrits en section 3.

détermination d'un ordre entre les clusters et permet alors certainement de réaliser une évaluation plus fiable que les deux approches de reconstruction en largeur et en profondeur. D'un point de vue plus pragmatique, cette mesure rend compte de la capacité qu'aura un utilisateur novice (qui n'a pas l'habitude d'utiliser un système présentant des catégories de résultats et donc ne s'oriente pas selon les informations collectées dans les différents groupes) à atteindre les documents pertinents à partir de la liste de clusters présentés.

2.4. Parcours orienté par la pertinence des documents

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. Dans ce sens, 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 $P(sel(\vec{x}) = C_i)$, de choisir le cluster C_i après avoir examiné x_l documents dans chaque cluster C_l , peut être remplacée dans les formules 8 et 9 par :

$$\frac{0.5 + \sum_{j=1}^{x_i} Pert(D_j^i)}{x_i + 1} \bigg/ \sum_{l \in \{1, \dots, k\}, x_l < n_l} \frac{0.5 + \sum_{j=1}^{x_l} Pert(D_j^l)}{x_l + 1} \quad [11]$$

Une telle probabilité de sélection de cluster permet de déterminer un score d'évaluation correspondant à l'espérance de précision moyenne obtenue par un utilisateur se servant de ses observations pour s'orienter vers les clusters semblant contenir le plus fort ratio de documents pertinents. 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 moyen orienté par des retours de pertinence ne requiert pas, encore une fois, 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.

2.5. 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 8 et 9, par :

$$\frac{0.5 + \sum_{j=1}^{x_i} \sum_{D \in \mathcal{P}ne(\vec{x})} \frac{Sim(D, D_j^i)}{|\mathcal{P}ne(\vec{x})|}}{x_i + 1} \quad [12]$$

$$\sum_{l \in \{1, \dots, k\}, x_l < n_l} \frac{0.5 + \sum_{j=1}^{x_l} \sum_{D \in \mathcal{P}ne(\vec{x})} \frac{Sim(D, D_j^l)}{|\mathcal{P}ne(\vec{x})|}}{x_l + 1}$$

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 influencer 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 $[0, 1]$. Avec une telle normalisation, le document D_x le plus proche du document pertinent D_p se voit attribuer une similarité $Sim(D_x, D_p) = 1$ et le document D_y qui en est le plus éloigné obtient une similarité $Sim(D_y, D_p) = 0$, ce qui permet de donner un poids plus important aux clusters contenant des documents très proches des documents pertinents. 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 ces labels dans les calculs des probabilités de sélection donnés par la formule [12]. Une initialisation du numérateur et du dénominateur de cette formule selon la similarité des documents pertinents avec les descriptions des clusters (plutôt que selon la valeur arbitraire de 0.5) peut alors permettre de favoriser les clusters

paraissant les plus intéressants. La probabilité de sélection $P(sel(\vec{x}) = C_i)$ devient alors :

$$\sum_{l \in \{1, \dots, k\}, x_l < n_l} \frac{\sum_{D \in \mathcal{P}_{ne}(\vec{x})} \frac{Sim(D, Label(C_l))}{|\mathcal{P}_{ne}(\vec{x})|} + \sum_{j=1}^{x_l} \sum_{D \in \mathcal{P}_{ne}(\vec{x})} \frac{Sim(D, D_j^l)}{|\mathcal{P}_{ne}(\vec{x})|}}{x_l + 1} \quad [13]$$

Avec des systèmes décrivant le contenu des clusters en sélectionnant un document représentatif dans chacun d'entre eux (tel que réalisé pour nos expérimentations), 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 $P((1, \dots, 1)) = 1$ et en considérant $P(exam(i-1, \sum_{j=1}^{i-1} Pert(D_j^i)) = D_i^i) = 1$ pour chacun des représentants de cluster¹⁴ (les autres probabilités d'examen sont fixées à 0 pour ces documents). Néanmoins, étant donné le faible impact qu'elle a sur les résultats finaux¹⁵, une telle orientation de la recherche peut être omise.

3. Expérimentations

Nous proposons ici de réaliser un certain nombre d'expérimentations visant à rendre compte des tendances observées par les mesures proposées face à des partitionnements de types différents.

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.

14. En supposant d'avoir placé les représentants en début de cluster : $\forall_{i \in \{1, \dots, k\}, Rep(C_i) = D_i^i}$

15. Les premiers documents des clusters ont, dans tous les cas, une forte probabilité d'être examinés en début de recherche.

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 significatifs¹⁶, sont utilisées dans nos expérimentations : la mesure *Cosine* et la *Query-Sensitive Similarity Measure*. Selon la mesure *Cosine*, la similarité de deux documents D_i et D_j est donnée par :

$$\text{Cosine}(D_i, D_j) = \frac{\sum_{l=1}^t w_{D_i,l} \times w_{D_j,l}}{\sqrt{\sum_{l=1}^t w_{D_i,l}^2 \times \sum_{l=1}^t w_{D_j,l}^2}} \quad [14]$$

$$\text{où } w_{D_i,l} = (1 + \ln(tf_{D_i,l})) \times \ln \frac{N}{N_l}$$

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) (Tombros *et al.*, 2004) réalise, pour le calcul de la similarité entre deux documents D_i et D_j , un produit entre leur score de similarité thématique classique (en termes de la mesure *Cosine*) et un score de proximité à la requête du vecteur $D_i \cap D_j$ correspondant à l'intersection de leurs deux représentations vectorielles (seuls les termes communs sont conservés dans ce vecteur, le poids de chaque terme correspond à la moyenne de ses poids dans les vecteurs individuels des deux documents) :

$$QSSM(D_i, D_j) = \text{Cosine}(D_i, D_j) \times \text{Cosine}(R, D_i \cap D_j) \quad [15]$$

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 termes 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 termes de *Cosine*). Dans l'ensemble de nos expérimentations, nous utiliserons les notations données par le tableau 1.

Quatre collections de documents tirées du corpus de la conférence TREC-1 ont été utilisées : le corpus ZIFF contient des articles informatiques, FR des extraits du

16. 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).

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

Tableau 1. *Notations des mesures d'évaluation*

Corpus	ZIFF	AP	WSJ	FR
Nb. documents	75180	84510	98732	25960
Nb. termes par document	297	217	204	927
Nb. termes des 1000 documents les + courts	19	14	11	49
Nb. termes des 1000 documents les + longs	7915	512	1235	11304
Nb. requêtes	31	33	39	13
Nb. documents pertinents par requête	82.87	43.24	61.69	29
Nb. documents pertinents retournés par requête	12.19	9.75	23.04	4.23

Tableau 2. *Statistiques des collections de documents*

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 tableau 2 présente les statistiques de ces corpus. Alors que le corpus FR contient un grand nombre de très longs documents, AP regroupe quant à lui des documents relativement courts. ZIFF et FR présentent la plus grande variété en termes de tailles de documents. Le même jeu de requêtes, les topics 1-50 de TREC, a été utilisé pour chaque corpus. Dans les expérimentations présentées dans la suite de ce rapport, nous utilisons deux formes des mêmes requêtes : la partie *Title* des requêtes (Tit.), qui contient 2.95 termes en moyenne (soit 2.95 uniques), et la partie *Narrative* (Nar.), qui contient une moyenne de 88.08 termes (soit 52.6 uniques). Le tableau 2 donne le nombre de requêtes pour lesquelles au moins un document dans les 50 premiers retournés par le système de recherche documentaire initial est pertinent. Comme pour les résultats expérimentaux présentés dans la suite, seules ces requêtes sont prises en compte pour le calcul des deux dernières statistiques présentées dans le tableau (le nombre moyen de documents pertinents dans chaque corpus et le nombre moyen de documents pertinents retournés parmi les 50 premiers par le système initial).

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. Dans le but 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 attentivement chacun des documents qui leur seraient présentés pour orienter leur recherche vers les groupes qui leur semblent le plus probablement contenir les informations pertinentes. Pour chaque requête considérée¹⁷, chaque système expérimenté¹⁸ et chaque individu, le processus expérimental est le suivant :

- 1) la description narrative de la requête est présentée au sujet qui la lit pour saisir précisément les informations qu'il doit rechercher ;
- 2) les k (dans nos expérimentations, $k = 3$) représentants de cluster sont présentés au sujet qui les lit pour avoir une idée des différences entre chaque groupe ;
- 3) le sujet choisit un indice de cluster (parmi ceux des clusters non encore entièrement examinés) en fonction de son sentiment concernant le cluster étant le plus susceptible de contenir les informations qu'il recherche ;
- 4) le premier document non encore examiné du cluster choisi est présenté et le sujet le lit pour saisir de quoi il retourne, ce qui est susceptible de changer ses impressions sur le contenu des différents groupes ;
- 5) si il reste des documents pertinents à localiser, le processus se poursuit à l'étape 3. Sinon, la précision moyenne de la liste résultant du parcours réalisé est calculée et le processus se termine.

Même si les différents sujets ont parfois suivi des routes différentes, les expérimentations réalisées permettent de capturer les variations du degré de capacité à atteindre les documents pertinents à partir des listes de représentants présentées. L'objectif est alors de calculer la corrélation entre les variations observées et celles générées par les différentes mesures de reconstruction de liste (en termes de précision moyenne).

17. Les topics 1-20 de *TREC-I* sont utilisés dans cette étude.

18. Chaque individu est confronté successivement aux quatre systèmes G,C , K,C , G,Q et K,Q pour chacune des requêtes. Néanmoins, l'ordre dans lequel ces systèmes sont utilisés est différent pour chaque sujet. Le corpus utilisé est le corpus AP , qui contient les documents les plus courts.

La corrélation entre valeurs observées o_i et valeurs mesurées m_i est évaluée par un coefficient $R_{o,m}^2$:

$$R_{o,m}^2 = 1 - \frac{\sum_{i=1}^N (o_i - m_i)^2}{\sum_{i=1}^N (o_i - \bar{o})^2} \quad [16]$$

où o_i correspond à la moyenne des scores obtenus par les 10 sujets sur la i -ième requête en utilisant le système considéré (m_i le score obtenu par la mesure concernée sur cette même requête) et $\bar{o}$ représente la moyenne des scores o_i sur l'ensemble des 20 requêtes. Ce que l'on cherche à évaluer ici est la capacité des mesures proposées à suivre les variations des scores obtenus par les sujets. Puisque le but final est de déterminer si les mesures sont capables d'ordonner correctement les différents systèmes selon leurs performances, le fait qu'une mesure obtienne des scores plus élevés qu'une autre ne nous importe pas, seule compte la corrélation avec les variations de score observées. Par conséquent, plutôt que de considérer directement les valeurs obtenues, nous normalisons chaque distribution de scores en remplaçant chaque valeur s_i (observée ou obtenue par une mesure) par :

$$s_i = \frac{s_i - \bar{s}}{\sigma_s} \quad [17]$$

où $\bar{s}$ correspond à la moyenne des scores sur l'ensemble des requêtes et σ_s à leur écart type. En utilisant une telle normalisation, la corrélation évaluée ne dépend que des variations observées (puisque chaque distribution présente alors une moyenne de 0 et un écart type de 1).

$R_{o,m}^2$	PRO	LAR	LEU	OPT	PM1	PM2	PM3
G, C	0.85	0.80	0.89	0.87	0.83	0.92	0.92
K, C	0.51	0.76	0.62	0.73	0.81	0.86	0.90
G, Q	0.72	0.72	0.83	0.82	0.78	0.89	0.91
K, Q	0.47	0.71	0.55	0.70	0.74	0.81	0.88

Tableau 3. *Corrélation entre variations observées*

Le tableau 3 présente, pour les quatre différents systèmes, les coefficients de corrélation obtenus entre les distributions de valeurs observées et celles des valeurs calculées par les différentes approches de reconstruction de liste. 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 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*¹⁹. 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*, paraît 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

Afin d'évaluer l'influence qu'ont les différentes caractéristiques des clusterings considérés (*i.e.*, les variations de tailles de leurs clusters et le degré de répartition des documents pertinents dans leurs clusters) sur les espérances des scores obtenus par les différentes mesures, nous proposons de réaliser des expérimentations impliquant des clusterings aléatoires de documents. Différents types de clustering sont produits pour étudier l'influence du degré de répartition des documents dans les clusters : selon le type de clustering concerné, chaque document est affecté à un cluster spécifique (le premier par exemple) selon une probabilité donnée, prise parmi les valeurs {0.2, 0.33, 0.5, 0.66, 1}. Alors qu'une probabilité de 1 conduit au regroupement de l'ensemble des documents dans un même groupe, une probabilité de 0.2 tend à produire des groupes de tailles relativement homogènes. Pour chaque requête (topics 1-50 de *TREC*), 100 clusterings de chaque type ont alors été produits pour étudier l'influence des tailles de clusters sur les scores attribués par les mesures d'évaluation. Ces clusterings concernent la répartition en cinq clusters des 20 premiers documents retournés par une recherche initiale sur le corpus *AP*. Les représentants de clusters sont ici sélectionnés aléatoirement.

La figure 5 trace les distributions des scores obtenus par les différentes mesures pour les différents types de clustering. Selon les courbes présentées, à l'exception du parcours optimal *OPT* dont l'espérance diminue pour les clusterings regroupant la plupart des documents dans un même groupe, l'ensemble des mesures paraissent présenter une espérance constante quelle que soit la diversité de tailles de clusters.

Plus intéressante est l'étude de l'influence du degré de répartition des documents pertinents dans les clusters : après avoir aléatoirement distribué dans 5 clusters l'ensemble des documents non pertinents considérés, chaque document pertinent est alloué au premier cluster avec une probabilité de 0.2, 0.33, 0.5, 0.66 ou 1 selon le type

19. 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.

Figure 5. Influence de la répartition des documents dans les clusters

Figure 6. Influence de la répartition des documents pertinents dans les clusters

de partitionnement concerné. Ces différentes probabilités d'affectation des documents pertinents au premier groupe nous permettent d'obtenir des clusterings présentant différents degrés de concentration de l'information pertinente : 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 6 présente alors les distributions des scores obtenus 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 6, 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 le degré de concentration de l'information pertinente 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 des similarités des représentants avec les documents pertinents. Nous pensons que cette diminution d'espérance n'existerait pas avec des partitionnements 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.

3.1.3. Influence du nombre de clusters produits

Afin d'évaluer l'influence du nombre de clusters produits sur l'espérance des différentes mesures, nous produisons, pour chaque requête (topics 1-50 de *TREC*) et chaque nombre de clusters entre 3 et 7, 100 clusterings aléatoires des 50 premiers documents retournés par une recherche initiale. Ces clusterings sont produits en choisissant aléatoirement k documents (avec k le nombre de clusters à produire), qui représentent alors les centres des clusters, et en affectant chaque document non sélectionné au cluster dont le centre lui est le plus proche. Un tel processus nous permet de travailler avec des clusters n'étant pas totalement dénués de signification, tout en écartant les biais qui auraient pu survenir en choisissant de travailler avec une méthode de clustering particulière (qui peut favoriser un nombre de clusters donné selon les distributions de similarités considérées). La figure 7 trace les distributions de scores obtenus par les différentes mesures selon le nombre de clusters produits. Selon les courbes obtenues, on peut noter que la plupart des mesures ne sont pas influencées par le nombre de clusters produits (à l'exception de *LAR*, *OPT* et *PM1*). La diminution d'espérance du parcours moyen lorsque le nombre de clusters augmente peut s'expliquer par le fait que plus le nombre de clusters est élevé, plus le nombre de clusters ne contenant pas de document pertinent est susceptible d'être élevé, d'autant plus lorsque l'on considère la *Cluster Hypothesis* qui fait tendre la majorité des documents pertinents à se regrouper dans un même groupe. L'utilisation des retours de pertinence *PM2* ou de la proximité des documents examinés avec les documents pertinents *PM3* pour orienter le parcours semble résoudre ce problème. Les informations collectées par ces mesures au cours de la recherche permettent d'orienter leur parcours vers les clusters contenant les documents pertinents. Observant une espérance relativement constante quel que soit le nombre de clusters produits, ces mesures peuvent être utilisées pour comparer des systèmes réalisant des clusterings de résultats aux nombres de groupes différents.

Figure 7. Influence du nombre de clusters produits

3.2. Comparaison des systèmes

L'objectif de cette section est de comparer nos 4 différents systèmes, en utilisant les nouvelles mesures proposées en section 2, afin de mesurer l'impact que peut avoir la prise en compte de la requête sur les clusters présentés à l'utilisateur. Les résultats expérimentaux donnés dans cette section concernent tous des partitionnements, en cinq clusters, des 50 premiers résultats d'une recherche initiale. [0.5cm]

		ZIFF			AP			WSJ			FR		
		Co	Sep	EcD	Co	Sep	EcD	Co	Sep	EcD	Co	Sep	EcD
Tit	G, C	0.56	0.47	15.94	0.57	0.47	15.32	0.57	0.47	16.10	0.65	0.44	13.04
	K, C	0.57	0.45	7.06	0.58	0.45	6.50	0.58	0.45	6.00	0.66	0.41	6.28
	G, Q	0.56	0.46	13.62	0.57	0.46	14.87	0.57	0.46	14.51	0.65	0.43	13.17
	K, Q	0.56	0.45	7.03	0.57	0.45	7.17	0.58	0.45	6.37	0.64	0.42	6.93
Nar	G, C	0.58	0.46	14.01	0.59	0.46	12.84	0.58	0.46	13.20	0.61	0.46	14.28
	K, C	0.58	0.45	5.82	0.59	0.45	5.86	0.59	0.45	5.87	0.63	0.44	6.02
	G, Q	0.58	0.45	13.10	0.59	0.45	12.50	0.58	0.45	12.43	0.61	0.45	12.63
	K, Q	0.58	0.45	5.86	0.59	0.45	5.96	0.58	0.45	5.53	0.63	0.44	6.10

Tableau 4. Statistiques des partitionnements produits

Le tableau 4 présente les statistiques des partitionnements produits par les différents systèmes : *Co* correspond à la cohésion moyenne des groupes formés, *Sep* à leur séparation (c'est-à-dire la dissimilarité moyenne entre groupes) et *EcD* à l'écart type entre tailles de clusters. Les critères de cohésion et de séparation des clusters sont calculés selon la mesure *Cosine*. Selon les résultats, la méthode *Group-Average* tend à produire un cluster contenant une grande part des documents et à isoler les documents marginaux dans les autres groupes, ce qui conduit à l'obtention de groupes très dissimilaires mais présentant un faible degré de cohésion. La méthode *K-means*, quant à elle, paraît produire des clusters de tailles plus homogènes, dont les éléments sont

Tit.	ZIFF				AP				WSJ				FR			
	<i>G, C</i>	<i>K, C</i>	<i>G, Q</i>	<i>K, Q</i>	<i>G, C</i>	<i>K, C</i>	<i>G, Q</i>	<i>K, Q</i>	<i>G, C</i>	<i>K, C</i>	<i>G, Q</i>	<i>K, Q</i>	<i>G, C</i>	<i>K, C</i>	<i>G, Q</i>	<i>K, Q</i>
<i>NbP</i>	1.10	1.69	1.19	1.65	0.82	1.73	1.01	1.37	0.96	2.17	1.08	1.98	1.75	2.50	1.75	2.30
<i>MK1</i>	0.68	0.72	0.70	0.73	0.62	0.66	0.63	0.67	0.68	0.75	0.70	0.77	0.57	0.67	0.57	0.68
<i>PRO</i>	0.41	0.42	0.42	0.41	0.46	0.41	0.43	0.42	0.52	0.57	0.52	0.55	0.50	0.49	0.48	0.47
<i>LAR</i>	0.35	0.40	0.34	0.39	0.32	0.38	0.33	0.34	0.42	0.49	0.40	0.48	0.47	0.53	0.47	0.52
<i>LEU</i>	0.43	0.44	0.43	0.42	0.48	0.44	0.44	0.43	0.55	0.60	0.53	0.57	0.51	0.51	0.50	0.49
<i>OPT</i>	0.52	0.65	0.52	0.62	0.54	0.66	0.53	0.62	0.63	0.80	0.62	0.78	0.70	0.73	0.69	0.71
<i>PM1</i>	0.38	0.42	0.36	0.40	0.36	0.40	0.36	0.36	0.47	0.54	0.45	0.53	0.49	0.56	0.48	0.53
<i>PM2</i>	0.42	0.46	0.42	0.43	0.46	0.46	0.45	0.44	0.54	0.63	0.51	0.60	0.50	0.58	0.50	0.56
<i>PM3</i>	0.43	0.46	0.40	0.42	0.45	0.48	0.46	0.46	0.53	0.63	0.51	0.61	0.51	0.60	0.51	0.59
Nar.	ZIFF				AP				WSJ				FR			
	<i>G, C</i>	<i>K, C</i>	<i>G, Q</i>	<i>K, Q</i>	<i>G, C</i>	<i>K, C</i>	<i>G, Q</i>	<i>K, Q</i>	<i>G, C</i>	<i>K, C</i>	<i>G, Q</i>	<i>K, Q</i>	<i>G, C</i>	<i>K, C</i>	<i>G, Q</i>	<i>K, Q</i>
<i>NbP</i>	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
<i>MK1</i>	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
<i>PRO</i>	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
<i>LAR</i>	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
<i>LEU</i>	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
<i>OPT</i>	0.68	0.76	0.69	0.75	0.72	0.77	0.72	0.79	0.74	0.82	0.76	0.83	0.74	0.78	0.74	0.82
<i>PM1</i>	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
<i>PM2</i>	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
<i>PM3</i>	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 5. Résultats des systèmes

plus proches les uns des autres. Enfin, la mesure *QSSM* semble conduire à une légère baisse des deux critères de cohésion et de séparation des clusters.

Le tableau 5 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 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*.

Selon nous, la *Query Sensitive Similarity Measure* souffre du fait qu'elle tend à minorer la similarité entre documents partageant un grand nombre de termes n'apparaissant pas dans la requête, que ces documents partagent ou non un grand nombre

de termes de la requête. De plus, un document pertinent contenant peu de termes de la requête ne peut pas, par l'utilisation d'une telle mesure, bénéficier de sa proximité avec d'autres documents pertinents, tel que cela pourrait être le cas avec des mesures de similarité classiques. Néanmoins, lorsque la requête est suffisamment longue, ces biais sont équilibrés par les bénéfices tirés de l'augmentation de la distance entre documents ne possédant pas de termes de la requête en commun (ce qui peut ne pas survenir lorsque la requête est courte puisque l'ensemble des documents considérés sont susceptibles de contenir la totalité des termes de la requête). Par la considération des mesures d'évaluation existantes, aucune amélioration significative n'est cependant observée. 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 contre-balançée 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 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., Méthodes de classification et de segmentation locales non supervisées pour la recherche documentaire, Phd thesis, Université d'Avignon, Janvier, 2000.
- 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.
- Chang C., Hsu C., « Customizable multi-engine search tool with clustering », *WWW'97*, ACM, 1997.
- Croft W. B., « A model of cluster searching bases on classification. », *Information Systems*, vol. 5, n° 3, p. 189-195, 1980.
- Cutting D. R., Pedersen J. O., Karger D., Tukey J. W., « Scatter/Gather : A Cluster-based Approach to Browsing Large Document Collections », *SIGIR'92*, p. 318-329, 1992.
- Hearst M. A., Pedersen J. O., « Reexamining the cluster hypothesis : Scatter/gather on retrieval results », *SIGIR'96*, Zürich, CH, p. 76-84, 1996.
- Iwayama M., « Relevance feedback with a small number of relevance judgements : incremental relevance feedback vs. document clustering », *SIGIR'00*, ACM, p. 10-16, 2000.
- 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*, DEC, 2006.
- Leuski A., « Evaluating document clustering for interactive information retrieval », *CIKM '01*, ACM, New York, NY, USA, p. 33-40, 2001.
- Liu X., Croft W. B., « Cluster-based retrieval using language models », *SIGIR '04*, ACM, New York, NY, USA, p. 186-193, 2004.
- Loupy (de) C., Bellot P., El-Bèze M., Marteau P.-F., « Query Expansion and Automatic Classification », *TREC-7, NIST special publication 500-242*, p. 443-450, 1999.
- Manning C. D., Raghavan P., Schütze H., *Introduction to Information Retrieval*, Cambridge University Press, July, 2008.
- Marteau P.-F., Loupy (de) C., Bellot P., El-Bèze M., « Systèmes & Sécurité », *Le traitement automatique du langage naturel, outil d'assistance à la fonction d'intelligence économique : vers une architecture Push-Pull d'accès à l'information*, vol. 5, n° 4, p. 8-41, 1999.

- McKeown K. R., Barzilay R., Evans D., Hatzivassiloglou V., Klavans J. L., Nenkova A., Sable C., Schiffman B., Sigelman S., « Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster », *Proc. Human Language Technology Conference*, 2002.
- Porter M., « An algorithm for suffix stripping », *Program*, vol. 14, n° 3, p. 130-137, 1980.
- Salton G., « Cluster search strategies and the optimization of retrieval effectiveness », *The SMART Retrieval System : Experiments in Automatic Document Processing*, Prentice-Hall, p. 223-242, 1971.
- Salton G., Buckley C., « Term-weighting approaches in automatic text retrieval », *Information Processing & Management*, vol. 24, n° 5, p. 513-523, 1988.
- Salton G., Buckley C., « Automatic Structuring and Retrieval of Large Text Files », *Communications of the ACM*, vol. 37, n° 2, p. 97-108, 1994.
- Silverstein C., Henzinger M., Marais H., Moricz M., Analysis of a Very Large AltaVista Query Log, Technical Report n° 1998-014, Digital SRC, 1998.
- Silverstein C., Pedersen J. O., « Almost-Constant-Time Clustering of Arbitrary Corpus Subsets. », *SIGIR*, ACM, p. 60-66, 1997.
- Spink A., Wolfram D., Jansen M. B. J., Saracevic T., « Searching the Web : the public and their queries », *J. Am. Soc. Inf. Sci. Technol.*, vol. 52, n° 3, p. 226-234, 2001.
- Tombros A., The Effectiveness of Query-Based Hierarchic Clustering of Documents for Information Retrieval, PhD thesis, University of Glasgow, UK, 2002.
- Tombros A., Jose J. M., Ruthven I., « Clustering Top-Ranking Sentences for Information Access », in , T. Koch, , I. Sølberg (eds), *ECDL*, vol. 2769 of *Lecture Notes in Computer Science*, Springer, p. 523-528, 2003.
- 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.
- Van Rijsbergen C., *Information Retrieval, 2nd edition*, Departement of Computer Science, Information Retrieval Group, University of Glasgow, 1979.
- Voorhees E. M., Harman D., « Overview of the Fifth Text REtrieval Conference (TREC-5). », *TREC'97*, NIST Special Publication 500-238, p. 1-28, 1997.