Mesures de la qualité des systèmes de recherche d'information

Karen Pinel-Sauvagnat, Josiane Mothe

IRIT, Equipe SIG
Université de Toulouse
118 route de Narbonne, 31062 Toulouse Cedex 9
{sauvagnat,mothe}@irit.fr

ABSTRACT. *In this paper we review approaches for evaluating information retrieval systems using test collections. We first give the definition of a test collection and present the main metrics used in literature to evaluate systems. We then show, thanks to three examples (search results clustering, automatic summarization and image retrieval), the variety of the existing evaluation frameworks.*

RÉSUMÉ. *L'objectif de cet article est de présenter un panorama de l'évaluation des systèmes de recherche d'information se basant sur des collections de référence. Nous détaillons dans un premier temps ce qu'est une collection de référence ainsi que les mesures d'évaluation associées. Nous développons ensuite les problématiques d'évaluation à travers trois cadres de recherche d'information spécifiques: le clustering de documents, le résumé automatique et la recherche d'images par l'exemple, et montrons la variété des mesures et collections de référence existantes.*

KEYWORDS: *Information System, Information Retrieval System, Evaluation, Metrics, Collection*

MOTS-CLÉS : *Système d'Information, Système de Recherche d'Information, Evaluation, Mesures, Collections*

DOI:10.3166/ISI.X.X.1-27

1. Introduction

Systèmes d'Information et masses d'informations à gérer sont aujourd'hui étroitement liés. Documents de travail, rapports, documentations, emails, etc. se doivent d'être accessibles au sein des organisations sous peine de perte de productivité (Gordon, 1997): pourquoi perdre du temps à rechercher des informations ? La recherche d'une information peut prendre plusieurs formes, allant de la recherche d'un document précis et connu à la collecte d'informations sur un thème particulier (Bonnel *et al.*, 2008). Pour arriver à ses fins, l'utilisateur peut naviguer dans des collections de documents, ou bien formuler son besoin sous forme de requête en interrogeant un moteur. Dans ce contexte, les Systèmes de Recherche d'Information (SRI) permet-

tent de restituer des documents susceptibles d'être pertinents par rapport à un besoin utilisateur. La Recherche d'Information (RI) est un domaine vaste, incluant des problématiques d'indexation de documents, d'interprétation des requêtes et d'appariement documents/requêtes (Baeza-Yates, Ribeiro-Neto, 2011).

Historiquement, le développement des différents modèles et systèmes de RI est allé de pair avec leur évaluation, et ce depuis le début des années 50. L'évaluation doit bien sur être menée en fonction de ce que l'on cherche à évaluer. Saracevic (1995) a proposé six classes d'évaluation qu'il faudrait mener dans le but d'évaluer la qualité d'un SRI:

– D'un point de vue *ingénierie*, les performances "système" peuvent être évaluées: vitesse, maintenance, intégrité, coût d'accès/stockage, ... On parle alors de *l'efficience* d'un système.

– Au niveau *entrée* du SRI, des questions se posent quant à l'exhaustivité des collections considérées dans le domaine pris en compte.

– Au niveau *traitement*, l'évaluation concerne la performance des algorithmes et techniques utilisés pour la recherche. Les SRI sont ici évalués par rapport à leur *efficacité*.

– Au niveau *sortie*, les questions à traiter concernent les interactions avec le SRI, les retours utilisateurs, ...

– D'un point de vue *utilisation et utilisateur*, l'évaluation concerne l'intérêt des systèmes pour les problèmes et tâches qu'ils cherchent à prendre en compte.

– Enfin au niveau *social*, il faut considérer les impacts des SRI sur leur environnement (effets sur la recherche, la productivité, le processus de décision,...).

Les trois premiers points concernent ce que l'on appelle généralement l'évaluation automatique, alors que les trois derniers concernent majoritairement l'évaluation orientée utilisateur. Ces classes, bien que non-disjointes, sont évaluées dans les faits de façon isolée, ce que Saracevic (1995) considère un défaut dans les procédures d'évaluation existantes. Dans cet article nous nous intéressons à l'évaluation de l'efficacité des systèmes (c'est à dire leur capacité à renvoyer des documents pertinents), en laissant de côté les problèmes d'efficience, ainsi que l'évaluation utilisateur (généralement basée sur des études d'usage). On trouvera dans (Kelly, 2009) un état de l'art complet sur ce dernier point.

Le scénario de recherche le plus connu est celui de la recherche *adhoc*: l'utilisateur envoie une requête sur une collection de documents (texte, images, vidéos) fixée et obtient en réponse des résultats provenant de la collection. Ces documents doivent être pertinents, c'est à dire répondre au besoin en information initial. Afin d'évaluer la capacité des SRI à renvoyer des documents pertinents, il est nécessaire de fixer des collections de documents, des requêtes sur ces collections, ainsi que l'ensemble des documents pertinents correspondant aux requêtes. C'est ce que l'on appelle en RI des collections de test (ou encore collections de référence). On associe à ces collections un ensemble de mesures, qui vont permettre d'évaluer les résultats d'un système.

Ce processus de construction de collections de test a été initié au début des années 1960 par Cleverdon dans le projet Cranfield (Cleverdon, 1967). Les bases de l'évaluation des SRI étant ainsi posées, de nombreuses campagnes d'évaluation ont vu le jour, parmi lesquelles on peut citer TREC (*Text Retrieval Conference*, http://trec.nist.-

gov), CLEF (*Conference and Labs of the Evaluation Forum*, http://www.clef-initiative.eu/) ou encore NTCIR (*NII Test Collection for IR Systems*, http://research.nii.ac.jp/-ntcir/index-en.html).

Cet article permet de poser les bases nécessaires à la compréhension de la problématique générale de l'évaluation en recherche d'information. Nous proposons également, à travers trois problématiques spécifiques (*clustering* de documents, résumés automatiques et recherche d'images), de montrer la diversité des cadres d'évaluation qu'il est nécessaire de mettre en œuvre. Dans ces cadres spécifiques, nous avons contribué à la définition de mesures et/ou de collections de test.

Cet article est organisé comme suit. Dans un premier temps, nous détaillons ce qu'est une collection de test et les différentes campagnes d'évaluation qui ont permis de construire ces collections (section 2). La section 3 présente les mesures d'évaluation majeures utilisées sur ces collections pour évaluer la qualité des systèmes. Nous détaillons dans les sections 4, 5, 6 les trois problématiques spécifiques que nous avons choisi de développer dans cet article: le *clustering* de documents, le résumé automatique et la recherche d'images par l'exemple. La section 7 conclut cet article.

2. Evaluation basée sur des collections de test

2.1. Cranfield

La première évaluation de SRI est le travail effectué par Cleverdon dans les années 60 sur le projet Cranfield (Cleverdon, 1967). Afin de comparer plusieurs systèmes, Cleverdon a proposé une collection de test (CRANFIELD II), composée de 1400 articles scientifiques. Les auteurs des articles ont, à la demande de Cleverdon, rédigé une requête résumant la problématique de leur article. Ils ont également évalué sur une échelle de 1 à 5 la pertinence par rapport à ces requêtes de tous les articles les référençant. Des évaluateurs indépendants ont ensuite évalué la pertinence de chaque document restant de la collection en se référant aux auteurs en cas de doute. Ceci a permis de collecter pour chaque requête l'ensemble des documents pertinents dans la collection.

Cleverdon a ainsi posé les bases de l'évaluation utilisant des collections de test : sur une collection de documents fixe, le ou les systèmes à évaluer exécutent une requête, et les résultats sont comparés à la réponse idéale, c'est à dire à l'ensemble des documents pertinents pour la requête.

2.2. Collection de test

La collection de test est donc le support de l'évaluation en RI. Outre l'évaluation de l'efficacité des systèmes, elle permet de comparer les systèmes entre eux et d'assurer la reproductibilité des évaluations. Chaque collection de test est composée des éléments suivants (Sanderson, 2010b; E. M. Voorhees, 2002):

– une *collection de documents*, chaque document possédant un identifiant unique;

– un ensemble de besoins d'informations (en anglais *topics*) sur cette collection, un besoin étant généralement exprimé sous forme de descriptif textuel, et traduit en une suite de mots-clés (formant ainsi la requête qui sera envoyée aux SRI);

– des *jugements de pertinence* (ou vérité terrain), permettant d'indiquer quels documents sont pertinents pour chacun des besoins. Ces jugements de pertinence sont généralement binaires (un document est pertinent ou non) et produits manuellement.

D'une manière générale, pour évaluer un SRI à l'aide d'une collection de test, on procède de la façon suivante: le système exécute les requêtes une par une sur la collection de documents et renvoie pour chacune une liste ordonnée de documents qu'il considère comme potentiellement pertinents. Ces listes mises bout à bout forment ce que l'on appelle une exécution (en anglais *run*). Chaque *run* est ensuite comparé aux jugements de pertinence et des mesures d'efficacité sont calculées. Nous présentons ces mesures dans la section 3.

Construire des collections de test n'est pas chose aisée: il faut obtenir des documents (qui doivent être libres de droits ou distribuables à la communauté sous certaines conditions), associer des requêtes "intéressantes", mais également et surtout, effectuer des jugements de pertinence pour chaque requête. La collection Cranfield II, bien que posant les bases de l'évaluation en laboratoire, est de petite taille comparée aux masses d'information que les SRI doivent être capables de gérer. Les collections de taille importante mises en place assez rapidement par la communauté de la RI ont rendu impossible la collecte de jugements de pertinence exhaustifs. Afin de contourner cette limitation du passage à l'échelle, Jones, Rijsbergen (1975) ont proposé l'utilisation du vote (en anglais *pooling*). Le principe est le suivant: seul un ensemble représentatif des documents potentiellement pertinents de la collection est jugé manuellement. Cet ensemble (le *pool*) est construit en utilisant les k premiers documents restitués par différents SRI (généralement ceux que l'on cherche à évaluer).

2.3. Discussion

Cleverdon, dans ses expérimentations Cranfield, s'est basé sur trois hypothèses simplificatrices (E. M. Voorhees, 2002):

1. la pertinence d'un document est équivalente à sa similarité à la requête. Ceci implique que la pertinence d'un document est indépendante de la pertinence des autres documents de la collection, et que le besoin en information de l'utilisateur est statique.

2. les jugements de pertinence associés à une requête sont représentatifs de tous les besoins traduits par la requête (et donc de tous les besoins de l'ensemble de la population)

3. tous les documents pertinents pour une requête sont connus.

La première hypothèse est particulièrement forte: on considère la pertinence comme absolue. Les travaux cherchant à exploiter des jugements de pertinence relatifs sont relativement récents, ils impliquent de juger pour des paires de documents quel document est préféré, comme cela est fait dans l'ordonnancement automatique (Cao *et al.*, 2007). Une autre considération à prendre en compte est celle de la pertinence marginale: un document est-il encore utile si l'information qu'il porte est déjà connue de l'utilisateur grâce à d'autres documents consultés précédemment? Une façon de mesurer la pertinence marginale est d'évaluer la nouveauté et la diversité des réponses (Clarke *et al.*, 2008).

Concernant le deuxième point, afin de biaiser le moins possible l'évaluation, les besoins mis en place dans les collections de test sont généralement non ambigus: il ne peut y avoir plusieurs interprétations différentes d'un besoin. La pertinence reste cependant très subjective car dépendante de chaque personne. Certaines études menées dans le cadre de TREC ont montré un accord faible entre différents juges chargés d'évaluer une même requête (Manning *et al.*, 2008). Cependant, même si ces désaccords entrainent une différence dans les résultats de certaines mesures, le classement général des différents systèmes n'est que peu affecté.

Enfin, la troisième hypothèse a rapidement été remise en cause par l'utilisation des techniques de vote. Ces techniques ont été évaluées comme fiables (Zobel, 1998) : même si certains documents pertinents ne font pas partie de l'ensemble des documents jugés, le classement des systèmes ne s'en trouve pas affecté et les collections de tests peuvent être réutilisables (c'est à dire que des systèmes n'ayant pas servi à former l'ensemble de documents jugé peuvent tout de même être évalués sur la collection).

Le coût des jugements de pertinence reste très élevé, et d'autres solutions ont été envisagées dans la littérature. On peut par exemple se baser que les journaux de connexion (*logs*) des moteurs de recherche, en considérant qu'un document consulté ("cliqué") est pertinent (Liu *et al.*, 2007). Cette dernière technique, bien que basée sur une hypothèse très forte, permet de se passer complètement des jugements adhoc. Une autre voie, particulièrement à l'étude aujourd'hui, est celle du *crowdsourcing* (Alonso *et al.*, 2008), consistant à demander à de très nombreux utilisateurs en ligne, issus de communautés diverses, d'effectuer chacun une petite tâche d'évaluation.

Malgré ces limitations, les évaluations basées sur les collections de test sont largement moins coûteuses que les évaluations basées sur des études d'usage, et fournissent de nombreuses informations sur le comportement des systèmes (E. M. Voorhees, 2002).

2.4. Campagnes d'évaluation

Les collections de test sont généralement construites dans le cadre de campagnes d'évaluation. Ces campagnes fonctionnent toutes sur le même principe : tous les ans, des tâches de recherche à évaluer sont définies, une collection de documents et de besoins est distribuée aux participants qui doivent renvoyer les résultats d'exécution correspondants. Une conférence est ensuite organisée pour que les participants confrontent leurs résultats et points de vue, et discutent des tâches prioritaires de la campagne suivante.

La première campagne (et aussi la plus connue) à avoir vu le jour est la campagne TREC (*Text REtrieval Conference*) (Ellen M. Voorhees, 2005). Mise en place en 1992 avec une tâche de recherche adhoc, elle propose aujourd'hui de très nombreuses tâches de recherche, parmi lesquelles on peut citer pour 2012 la tâche de recherche Web, la tâche de RI médicale, la tâche de recherche dans des microblogs, ...

La campagne CLEF a été à l'origine construite pour l'évaluation de la RI multilingue. Elle est aujourd'hui devenue un forum pour de nombreuses évaluations spécialisées: RI dans le domaine de la chimie, des maths, ... Le pendant asiatique de la campagne CLEF, encore aujourd'hui dédié à la RI multilingue, est la campagne NTCIR.

D'autres campagnes spécifiques existent, parmi lesquelles on peut citer TRECVid pour la vidéo (http://trecvid.nist.gov/), CLEFImage pour la recherche d'images (http://-www.imageclef.org/) ou encore INEX (INitiative for the Evaluation of XML Retrieval, https://inex.mmci.uni-saarland.de/) pour la RI dans des documents semi-structurés, intégrée en 2012 dans la campagne CLEF.

Le lecteur intéressé par plus d'informations sur les collections de test et leur utilisation en RI pourra se référer à l'état de l'art complet proposé par Sanderson (2010b).

3. Mesures d'évaluation de la recherche adhoc

3.1. Mesures usuelles utilisées en recherche d'information

Dans le domaine documentaire, deux notions fondamentales sont utilisées pour qualifier la qualité de la réponse d'un système à un besoin d'information : le silence et le bruit documentaires. Le silence documentaire fait référence aux documents pertinents existants mais que le système n'a pas restitué à l'utilisateur tandis que le bruit documentaire fait référence aux documents restitués à l'utilisateur mais ne répondant pas à son besoin. En RI, deux mesures ont été définies pour quantifier ces notions. Le rappel mesure la proportion de documents pertinents restitués et la précision mesure la proportion de documents pertinents dans l'ensemble de documents restitués. Pour une requête q donnée, ces mesures sont définies par :

$$Rappel(q) = \frac{RestPert(q)}{Pert(q)} \tag{1}$$

$$Precision(q) = \frac{RestPert(q)}{Rest(q)} \tag{2}$$

Où $RestPert(q)$ correspond au nombre de documents restitués et pertinents par le système en réponse au besoin d'information q, $Pert(q)$ (resp. $Rest(q)$) correspond au nombre de documents pertinents (resp. restitués) pour ce même besoin d'information. Les valeurs de ces mesures sont comprises entre 0 et 1 et sont optimales pour 1. Ces deux mesures varient en sens inverse : les méthodes permettant d'augmenter la précision ont tendance à dégrader le rappel et vice versa. La mesure F_1 permet de combiner le rappel et la précision comme suit :

$$F_1 = 2\frac{Rappel(q).Precision(q)}{Rappel(q) + Precision(q)} \tag{3}$$

Cette mesure F_1 donne la même importance à la précision et au rappel. Des variantes (F_β) permettent de donner plus d'importance à l'un ou à l'autre.

$$F_\beta = (1+\beta)\frac{Rappel(q).Precision(q)}{\beta.Rappel(q) + Precision(q)} \tag{4}$$

Pour la même raison de complémentarité des mesures de rappel et de précision, les performances d'un système sont souvent reportées en termes de courbes Rappel/Préci-

sion. Ces courbes sont bien adaptées pour comparer de façon globale plusieurs systèmes (E. M. Voorhees, 2006). La précision est calculée pour différentes valeurs de rappel, généralement de 0 à 1 par pas de 0,1 et est interpolée comme suit :

$$Prec_{@r}(q) = \max\left(Prec_{@m}(q)\right) \ \ quelque\ soit\ \ m > r \tag{5}$$

r variant de 0 à 1 par pas de 0,1.

La précision peut également être calculée lorsque l'on considère les premiers documents retrouvés. Ce type de mesure est en particulier utilisé pour évaluer la haute précision. Par exemple, la $P@10$ correspond à la précision obtenue si seuls les 10 premiers documents de la liste des documents retrouvés étaient considérés.

La précision moyenne (AP pour *Average Precision*) est également largement utilisée pour évaluer un système ou pour comparer des systèmes. Elle est définie par:

$$AP(q) = \frac{\sum_{r=1}^{R}\left[P@r\ .rel\left(r\right)\right]}{Pert\left(q\right)} \tag{6}$$

avec $Pert(q)$ le nombre de documents pertinents pour la requête q, R le nombre de documents restitués, r le rang et $P@r$ la précision lorsque les r premiers documents retrouvés sont considérés. $rel\left(r\right)$ vaut 1 si le document au rang r est pertinent et 0 sinon.

Enfin, nous évoquerons le nDCG *(Normalized Discounted Cumulative Gain*) qui permet de considérer plus de deux niveaux de pertinence (Järvelin, Kekäläinen, 2002).

Le DCG au rang k est défini par :

$$DCG@k(q) = \sum_{r=1}^{k} \frac{2^{rel(r)} - 1}{r + 1} \tag{7}$$

Où $rel(r)$ est le niveau de pertinence du document au rang r.

Le NDCG normalise le DCG par le maximum du $DCG@k$.

Toutes ces mesures sont calculées pour une requête donnée. Cependant, les moteurs doivent être évalués sur un ensemble de requêtes; les mesures précédentes se déclinent alors en valeur moyenne. Par exemple, la moyenne de la précision moyenne (MAP pour *Mean Average Precision*) est la moyenne pour un ensemble de requêtes de l'AP obtenue pour chaque requête.

Le lecteur intéressé trouvera la présentation d'autres mesures dans différentes références (Rijsbergen, 1979) (Ellen M. Voorhees, 2005) (Baeza-Yates, Ribeiro-Neto, 2011). $Trec\_eval$ (http://trec.nist.gov/trec_eval/) dans sa version 8.1 calcule 135 mesures, le $readme$ de cet outil présente les mesures implantées.

3.2. Tests statistiques

En RI, il est fréquent de comparer les résultats de différents moteurs de recherche. Dans le cas de la définition d'une nouvelle approche, celle-ci doit être comparée à une approche reconnue de la littérature (le système témoin). Par ailleurs, il est fréquent de devoir mesurer l'impact de la valeur d'un paramètre d'une méthode de recherche. Les mesures de performances citées dans la section précédente permettent ces comparaisons. Cependant, il est important de pouvoir détecter des améliorations significatives

(c'est-à-dire non liées à la chance) même lorsque ces améliorations sont faibles. Les tests statistiques définis en mathématique ont cet objectif. Plusieurs tests statistiques peuvent être utilisés dans ce cadre. Différents travaux se sont intéressés à étudier les plus adaptés dans le cadre de la RI. Sanderson, Zobel (2005) ont comparé le test de Student (ou t-test), de Wilcoxon et de signe. Ils ont conclu que les tests de Student et de Wilcoxon étaient plus pertinents. Smucker *et al.* (2007) ont quant à eux étudié 6 tests (t-test, Wilcoxon, signe, bootstrap, randomization et permutation). Ils préconisent l'utilisation des tests de Student, le bootstrap et le randomization pour lesquels les auteurs montrent qu'ils sont par ailleurs généralement concordants.

Dans cette section, nous illustrerons l'utilisation des tests statistiques dans l'évaluation en RI par le t-test. Le t-test permet de comparer les moyennes de deux échantillons statistiques ou de comparer la moyenne observée d'un échantillon à une valeur donnée. Il s'applique sur des données qui suivent une distribution normale ce qui n'est pas le cas dans les données résultat produites en RI (Rijsbergen, 1979). Cependant, Hull (1993) indique que ce test peut être utilisé lorsque le nombre d'observations est suffisamment grand et qu'il reste efficace même lorsque la distribution normale des données n'est pas observée.

Le t-test de Student peut permettre de comparer les valeurs obtenues par deux systèmes pour une mesure de performance donnée (par exemple AP) pour un ensemble de requêtes. Ceci implique de connaitre pour les deux systèmes le détail des résultats pour chaque requête. Supposons que l'on ait les valeurs de AP pour une configuration de système B que l'on veut évaluer et celles du système de référence A auquel le système B doit être comparé. On espère que la valeur de MAP pour le système B est meilleure que la MAP du système de référence A. La question est de savoir si la différence observée est suffisante pour rejeter l'hypothèse nulle $H0$ que la méthode sous-jacente au système B n'apporte rien. La p-valeur associée est le seuil auquel on rejetterait l'hypothèse nulle compte tenu des valeurs observées. Le rejet de l'hypothèse nulle se fait généralement pour une valeur $p - value < 0,05$ en dessous de laquelle les deux systèmes comparés sont considérées comme statistiquement différents.

3.3. Discussion sur les mesures

La variabilité des systèmes (c'est-à-dire le fait qu'un système obtienne de bonnes performances sur une requête et de mauvaises sur une autre alors qu'un autre système fonctionnera inversement) implique de réaliser des évaluations sur un nombre important de cas. Les campagnes d'évaluation utilisent généralement 50 *topics*. Ce choix a été confirmé par les travaux de Buckley, Voorhees (2000) dans lequel les auteurs montrent que le comportement de la plupart des mesures est stable lorsque le nombre de cas diminue (de 50 à 5). Ils montrent cependant que cela n'est pas le cas pour certaines mesures; par exemple pour les mesures de haute précision ($P@10$ à $P@1$), le nombre de cas doit être augmenté pour avoir une confiance suffisante dans les résultats.

Les calculs des mesures de performance des systèmes de RI sont souvent basés sur l'outil $trec\_eval$. Cabanac *et al.* (2010) se sont intéressés à un biais dans l'utilisation de l'outil : lors du calcul des performances d'un système, les scores obtenus par les documents sont utilisés pour créer des listes ordonnées de documents « restitués » en réponse à un besoin d'information. Dans le cas de documents ex-equo l'ordre choisi

est celui des noms des documents. Les auteurs ont montré que ce choix pouvait avoir un impact sur le classement des systèmes.

Par ailleurs, le choix des mesures de performance pour comparer plusieurs systèmes dépend de l'objet de la comparaison. Par exemple, certaines tâches, comme la collecte de documents en vue de réaliser un état de l'art dans un domaine, doivent favoriser un rappel fort alors que d'autres doivent plutôt favoriser la précision. Si les mesures que nous avons citées au début de cette section sont les plus utilisées, nous avons rappelé que $trec\_eval$ en calculait un nombre important. Baccini *et al.* (2012) se sont intéressés à la redondance des mesures en s'appuyant sur une analyse grande échelle de résultats de moteurs sur différentes collections. Ils ont montré que les mesures pouvaient être regroupées en 6 groupes fortement corrélés. Les méthodes utilisées indiquent que même si les mesures ne mesurent pas les mêmes choses, leurs résultats sont fortement corrélés, ainsi une seule mesure de chaque groupe permet d'avoir une vue globale des performances des systèmes.

L'outil $trec\_eval$ a été développé en particulier pour évaluer la recherche adhoc, de nombreuses autres mesures ont été proposées dans la littérature afin de répondre à des tâches spécifiques. Dans les trois sections suivantes, nous nous centrons sur trois cas particuliers et indiquons les mesures de performances et les campagnes associées.

4. Clustering de documents

4.1. Description du problème

L'approche classique des SRI est de renvoyer aux utilisateurs en réponse à une requête une liste de documents triés par pertinence système décroissante. Certaines alternatives ont cependant été proposées afin de faciliter l'accès à l'information aux utilisateurs. Parmi elles, nous pouvons citer le *clustering* des résultats de recherche[1], qui a pour but de former des groupes de documents similaires (généralement par rapport à leur contenu) en réponse à une requête utilisateur (Manning *et al.*, 2008). On trouvera dans (Carpineto *et al.*, 2009) un panorama complet de ces approches de clustering. Elles sont présentes dans des systèmes aussi bien académiques que grand public tels que Vivisimo (Koshman *et al.*, 2006) ou Exalead (http://www.exalead.fr).

Historiquement, l'intérêt du clustering en RI a été souligné par Jardine, Rijsbergen (1971) au travers de la définition de la *Cluster hypothesis*. Cette dernière stipule que l'appartenance de documents à un même groupe donne une indication quant à la pertinence de ces documents à une même requête. Autrement dit, les documents proches (concentrés dans un même groupe) tendront à être pertinents pour les mêmes requêtes. Dans le même temps, la *Cluster hypothesis* a suscité des critiques (Tombros, 2002; Lamprier, 2008; Lamprier *et al.*, 2010). En effet, l'hypothèse qu'un système arrive à concentrer tous les documents pertinents au sein d'un même groupe est optimiste et lorsqu'un seul groupe est restitué, l'utilisateur doit tout de même le par-

1. Dans la littérature française, plusieurs termes sont employés pour désigner le clustering: catégorisation, classification non supervisée ou classification automatique de documents. Le vocabulaire employé ne faisant pas consensus, nous préférons conserver le terme anglais dans cet article, à l'instar des auteurs de (Lamprier, 2008) et (Denoyer, 2004).

courir. L'utilisateur gagnerait peut-être à obtenir plusieurs groupes, dans lesquels les différents aspects de la requête seraient distribués.

L'évaluation de ces techniques de clustering a été abordée, comme dans la RI ad-hoc, selon deux grands types d'approches : les approches dites "automatiques" (sans sollicitation humaine directe) et les études d'usage qui consistent à observer la performance d'individus confrontés à la réalisation d'une tâche de recherche (Julien *et al.*, 2008). Nous nous focalisons dans cette section sur les approches automatiques.

4.2. Mesures

La validité des résultats renvoyés peut être évaluée de façon interne ou externe (Halkidi *et al.*, 2001). Les mesures de validité interne vérifient certaines propriétés structurelles des groupes comme leur homogénéité, la détection des "individus" aberrants,....(Carpineto *et al.*, 2009). D'une manière générale, ces mesures cherchent à évaluer si le SRI atteint une forte similarité intra-cluster (les documents d'un même groupe sont similaires) et une faible similarité inter-clusters (les documents provenant de groupes différents sont différents). On trouvera un état de l'art détaillé sur la validité interne dans (Ingaramo *et al.*, 2008).

La plupart des mesures de validité externe se basent sur l'existence d'une vérité terrain binaire (pertinent/non pertinent). Parmi elles on peut citer

– celles cherchant à évaluer le groupe optimal (Jardine, Rijsbergen, 1971; Tombros, 2002)

– celles cherchant à évaluer la qualité globale des groupes (Nayak *et al.*, 2009; Vries *et al.*, 2011)

– celles reconstruisant des listes de résultats par simulation d'usagers, pour ensuite effectuer une évaluation "traditionnelle" (Lamprier *et al.*, 2010; Leuski, 2001).

Dans le premier cas (évaluation du groupe optimal), le meilleur groupe dans l'ensemble des groupes est sélectionné en fonction de la précision, du rappel ou de la F-mesure. En sélectionnant le groupe optimal, il est ensuite possible de comparer les approches avec celles proposant des listes triées de résultats, en utilisant la mesure du Mk-1k (Jardine, Rijsbergen, 1971; Tombros, 2002). Le principe est le suivant: soit l la liste des documents renvoyés par un moteur de recherche et utilisée en entrée de la technique de clustering évaluée. e_c et e_l sont deux ensembles de documents composés respectivement des documents du meilleur groupe de l'approche de clustering et des documents de l, avec l coupée à N documents, N étant le nombre de documents du meilleur groupe. Les ensembles e_c et e_l sont alors comparables grâce aux mesures rappel , précision et F_1 définies comme suit, pour un système s et une requête q :

$$Rappel(q) = \frac{RestPert(e_s, q)}{Pert(l, q)} \tag{8}$$

$$Precision(q) = \frac{RestPert(e_s, q)}{Rest(e_s, q)} \tag{9}$$

$$F_1(q) = \frac{2 \cdot Rappel(q) \cdot Precision(q)}{Rappel(q) + Precision(q)} \tag{10}$$

$Rest(e_s, q)$ et $RestPert(e_s, q)$ sont respectivement le nombre de documents restitués et le nombre de documents pertinents restitués par e_s pour q, avec $e_s \in \{e_c, e_l\}$. $Pert(l, q)$ est le nombre de documents pertinents dans l pour q. Un système idéal aura un rappel et une précision de 1. La mesure Mk-1k d'effectivité optimale est complétée par d'autres mesures, MK3 et MK4, dont on pourra trouver la description dans (Tombros, 2002).

Pour évaluer la qualité globale des groupes, il est possible d'utiliser la mesure NCCG (*Normalized Cumulative Cluster Gain*) (Nayak *et al.*, 2009). Le gain cumulatif d'un groupe (CCG - Cumulative Gain of a Cluster) est calculé en comptant le nombre de documents pertinents du groupe. Pour une approche donnée, un vecteur trié CG est créé représentant chaque groupe en fonction de sa valeur CCG. Les groupes ne contenant pas de document pertinent ont une valeur de 0. Un vecteur de gain cumulé pour le vecteur CG est ensuite calculé et normalisé par le vecteur de gain idéal. Chaque solution de clustering cs est ensuite évaluée sur sa capacité à répartir l'information pertinente sur les groupes pour une requête q. On a ainsi:

$$SplitScore(q, cs) = \sum^{|CG|} \frac{cumsum(CG)}{Pert(q)^2} \tag{11}$$

avec $Pert(q)$ le nombre de documents pertinents dans l'ensemble total de résultats pour la requête q. Le pire scénario est celui dans lequel chaque document pertinent est dans un groupe différent. Soit CG1 le vecteur qui contient le gain cumulatif de chaque groupe.

$$MinSplitScore(q, cs) = \sum^{|CG|} \frac{cumsum(CG1)}{Pert(q)^2} \tag{12}$$

Le gain cumulatif normalisé est alors:

$$nCCG(q, cs) = \frac{SplitScore(q, cs) - MinSplitScore(q, cs)}{1 - MinSplitScore(q, cs)} \tag{13}$$

La mesure nCCG est comprise entre 0 et 1: une bonne solution de clustering a une valeur élevée, ce qui traduit le fait qu'un grand nombre de documents pertinents sont regroupés dans un même groupe.

Les mesures précédentes se basent sur le fait que les documents répondant à une même requête sont dans un même groupe (respect de la *cluster hypothesis*). Or, lorsqu'une requête revêt plusieurs aspects, les groupes peuvent être vus comme complémentaires, et donc ces mesures ne sont plus adaptées. Pour solutionner le problème, plusieurs approches permettent de simuler le comportement de l'usager dans le parcours des groupes. Cette simulation, basée sur le fait que les groupes sont triés, ainsi que les documents qu'ils contiennent, aboutit à la reconstruction d'une liste de résultats, qui sera ensuite évaluée selon des méthodes "traditionnelles". Les parcours les plus simples sont les parcours en profondeur et en largeur. Le parcours en profondeur examine les groupes les uns après les autres, en considérant tous les documents d'un groupe avant de passer au suivant. Le parcours en largeur considère le premier document non lu par groupe, en bouclant sur l'ensemble des groupes, jusqu'à ce que tous les documents aient été lus. Leuski (2001) propose de simuler le comportement de l'utilisateur qui explorerait un groupe séquentiellement et en changerait dès qu'il trouve plus de documents non pertinents que de pertinents. Les travaux de Lamprier

(Lamprier *et al.*, 2010; Lamprier, 2008) proposent égalements plusieurs parcours: par exemple, le parcours orienté par la pertinence des documents modélise un usager qui prend en compte le ratio de documents pertinents trouvés dans chaque groupe, et le parcours orienté par la proximité des documents pertinents modélise un usager qui oriente sa recherche vers des groupes dont le contenu des documents examinés semble correspondre aux informations portées par les documents pertinents.

Dès lors que les jugements de pertinence reflètent la variété des résultats (c'est à dire dès lors que la pertinence n'est plus binaire mais montre les différents aspects de la requête), ou que l'on connaît a priori les catégories des documents, les groupes résultats peuvent être comparés à la classification idéale (*Gold Classification*). La mesure la plus utilisée dans ce cadre est probablement la pureté. La pureté mesure à quel point un groupe contient des documents majoritairement d'une seule classe/catégorie. Chaque groupe c se voit assigner le label correspondant à la majorité des documents qu'il possède.

$$\text{Pureté}(c) = \frac{\text{nombre de documents avec le label majoritaire dans c}}{\text{nombre de documents dans c}} \tag{14}$$

Puisqu'il y a plusieurs documents renvoyés pour chaque requête (pour une solution de clustering cs), les valeurs de pureté peuvent être agrégée selon une micro- ou macro-moyenne.

$$\text{micro-pureté}(cs) = \frac{\sum_{k=0}^{n} \text{pureté(k)} \cdot \text{nombre de documents dans k}}{\text{nombre de documents renvoyés dans cs}} \tag{15}$$

$$\text{macro-pureté}(cs) = \frac{\sum_{k=0}^{n} \text{pureté(k)}}{n} \tag{16}$$

Dans un contexte de recherche d'information, il est très facile d'obtenir un micro et une macro-pureté élevées, il suffit de renvoyer autant de groupes que de documents (c'est à dire un seul document par groupe). Pour limiter ce biais, le nombre de groupes peut être comparé au nombre de catégories de la classification idéale, ou on peut encore considérer le nombre de catégories qui ont été correctement identifiées.

Parmi les autres mesures servant à comparer avec une classification idéale, on peut citer la mesure F1, ou encore l'entropie et l'information mutuelle (Crabtree *et al.*, 2005). Ces mesures sont cependant seulement informatives puisqu'elles ne permettent pas vraiment de comparer deux méthodes entre elles. De plus, elles dépendent beaucoup du nombre de groupes renvoyés par les différents modèles (nombre qui n'est pas forcément identique). Enfin toutes ces mesures ne peuvent pas être utilisées lorsque les documents peuvent faire partie de plusieurs catégories. Par exemple, la mesure d'information mutuelle nécessite qu'il n'y ait pas d'intersection entre les groupes (Crabtree *et al.*, 2005) .

4.3. Campagnes d'évaluation

A notre connaissance, il existe peu de campagnes d'évaluation permettant d'évaluer des techniques de clustering des résultats de recherche. La plupart des évaluations existantes, parmi lesquelles on peut citer celles se basant sur la collection Reuters Lewis *et al.* (2004), évaluent la capacité des systèmes à faire des groupes de documents correspondant à des classes pré-établies mais indépendamment de toute requête.

Si l'on s'intéresse précisément à l'évaluation du clustering des résultats, nous pouvons citer la tâche XML Mining de la campagne d'évaluation INEX qui en 2009 et 2010 a explicitement concerné l'évaluation de la *Cluster hypothesis*. Elle s'est basée sur les requêtes et jugements de pertinence utilisés dans le cadre des tâches de recherche adhoc proposées les mêmes années. Il s'agissait d'évaluer la qualité des groupes dans le but de sélectionner le groupe de documents optimal pour chaque requête. La mesure NCCG présentée précédemment a été utilisée pour comparer les systèmes.

Dans le cadre du projet européen Quaero (*Projet européen Quaero*, 2012), nous avons également mis en place deux campagnes d'évaluation pour les approches de clustering de documents de recherche.

La première visait à évaluer la *Cluster hypothesis*, comme dans le cadre des campagnes INEX 2009 et 2010. Nous avons utilisé une collection composée de 2,6 millions de pages Web issues du domaine français (.fr) et aspirées par le moteur de recherche Exalead en 2008. Nous avons associé à cette collection 25 besoins, comprenant une requête sous forme de mots-clés et un texte explicitant le besoin en information. Ces besoins (que nous nommons mono-aspect) sont des besoins réels qui ont été soumis par des utilisateurs d'Exalead (extraits du log de ce moteur de recherche). Les jugements de pertinence ont été recueillis suivant une procédure similaire à celle utilisée dans TREC. Nous avons constitué un *pool* de résultats issus de 144 configurations différentes du moteur de recherche Terrier (Ounis *et al.*, 2005). Ces configurations ont été construites en utilisant différentes formes d'indexation, différents modèles de recherche, et en réalisant ou non l'expansion de requêtes. Ce *pool* a ensuite été évalué manuellement pour identifier les documents pertinents de chaque requête (pertinence binaire : pertinent/non pertinent). L'évaluation a ensuite été faite selon la mesure Mk-1k décrite dans la section précédente (Navarro *et al.*, 2011).

La seconde campagne avait un double objectif. Le premier était de nouveau d'évaluer des approches suivant la *Cluster hypothesis*, avec des requêtes "mono-aspect". Le second était d'évaluer des approches remettant en cause la *Cluster hypothesis*, et formant des groupes représentatifs des différents aspects d'une requête. Pour ce faire, nous avons utilisé un corpus de 10 millions de documents collectés par Exalead sur une période de 3 mois (100 premiers résultats de toutes les requêtes sur le moteur pendant 3 mois). A l'aide des logs associés, nous avons sélectionné 25 requêtes mono-aspect comme pour la première campagne d'évaluation, mais également 25 requêtes multi-aspects pour lesquelles la pertinence pouvait être définie selon plusieurs aspects (par exemple, la requête "bol d'or" peut concerner une course de moto ou bien une course de voile). Ces aspects, déterminés a priori en fonction de documents du corpus, n'ont pas été fourni aux participants de la campagne, mais ont servi pour les jugements de pertinence. Là encore, pour former le *pool* de documents utilisé pour les jugements de pertinence, nous avons utilisé plusieurs configurations du moteur de recherche Terrier (112). Les jugements de pertinence pour les requêtes mono-aspects ont été effectués, comme pour la première campagne, de façon binaire. Pour les requêtes multi-aspects, les personnes effectuant les jugements de pertinence ont dû déterminer pour chaque document du *pool* l'aspect principal du document ainsi éventuellement qu'un

ou plusieurs aspects secondaires. L'évaluation a ensuite été faite en fonction de la classification idéale (calcul de la pureté des groupes). A notre connaissance, cette dernière campagne est la seule campagne de RI permettant d'évaluer des techniques de clustering des résultats de recherche ne respectant pas la *Cluster hypothesis*.

5. Résumés automatiques

5.1. Description du problème

La production de résumés automatiques date des années 50. Luhn (1958) proposait alors d'utiliser la fréquence des mots et leur distribution pour estimer l'importance des mots puis des phrases afin de sélectionner celles qui apparaissaient dans le résumé produit automatiquement. La production de résumé peut être réalisée à partir d'un seul ou de plusieurs documents. Par ailleurs, certaines approches s'intéressent à produire des résumés à partir de documents en lien avec une requête particulière (Zhao *et al.*, 2009), (Alguliev *et al.*, 2011).

Vivaldi *et al.* (2011) distingue clairement deux sortes de résumés, en fonction de la façon dont ils sont générés : les extraits et les résumés (*extracts/abstracts* en anglais). Les premiers sont construits à partir de passages du texte original alors que les seconds paraphrasent les textes initiaux, en général en générant des phrases, selon des approches issues du traitement du langage naturel, à partir de passages ou de groupes de mots considérés comme importants. Dans les méthodes basées sur l'extraction des phrases à partir du texte initial, la cohérence des phrases est en général garantie, même si le problème de résolution des anaphores soit être résolu. Cette cohérence est en revanche un problème en soit pour les méthodes basées sur la génération automatique de phrases en langage naturel. Enfin, l'ordonnancement des phrases est un autre challenge du résumé automatique.

L'évaluation d'un résumé peut être basée sur le fait qu'il corresponde à un texte cohérent et qu'il contienne les concepts clefs (Jones, 2007). Pour identifier les concepts clefs, une solution consiste à déterminer un ensemble de questions trouvant leur réponse dans le texte d'origine ; le résumé sera considéré pertinent s'il permet de répondre à ces questions. Ces questions peuvent correspondre à des questions de compréhension de texte comme dans (Morris *et al.*, 1992) ou à un ensemble de questions liées au thème du texte initial comme dans SUMMAC (Mani *et al.*, 2002).

L'évaluation pyramidale (Nenkova, Passonneau, 2004a) s'appuie sur un ensemble d'annotateurs qui définissent des unités de contenu (*Summary Content Units*) dont le poids dépend du nombre d'annotateur les ayant identifiés. Un résumé automatique sera d'autant bien noté qu'il contient des unités de fort poids.

Il s'agit là d'approches manuelles qui ne permettent pas un passage à l'échelle aisé. Une autre approche consiste à comparer le résumé produit automatiquement avec un résumé de référence. Les résumés de référence peuvent être produits par des humains. Ce type d'évaluation a été largement utilisé en particulier dans les campagnes d'évaluation DUC (*Document Understanding Conference*, http://duc.nist.gov) et TAC *(Text Analysis Conference*, http://www.nist.gov/tac/). La comparaison entre le résumé de référence et le résumé produit automatiquement peut se faire sur la base

des phrases les composant ou sur la base de n-grammes comme dans ROUGE (Lin, 2004a) mais d'autres mesures ont également été proposées dans le domaine.

5.2. Mesures

5.2.1. Evaluation pyramidale

Il s'agit d'un modèle prédisant la distribution du contenu de l'information dans les résumés. Cette méthode est basée sur la notion de SCU (*Summary Content Unit*), c'est-à-dire des unités regroupant différentes expressions correspondant au même contenu. Ces expressions sont généralement issues de différents résumés de référence. Une expression, lorsqu'elle est sélectionnée pour faire partie d'un SCU y « contribue ». A chaque SCU est associé de façon manuelle un libellé qui exprime son contenu. Par ailleurs, à chaque SCU est associé un poids qui dépend du nombre de résumés de référence qui ont contribué au SCU. La construction des SCU est manuelle ; cependant des travaux s'intéressent à leur création automatique (Hennig *et al.*, 2010).

5.2.2. ROUGE

ROUGE, pour *Recall-Oriented Understudy for Gisting Evaluation* est un outil qui intègre plusieurs variantes de la mesure portant le même nom (Lin, 2004a) : ROUGE-n, ROUGE-L, ROUGE-W, ROUGE-S and ROUGE-SU. Ces métriques comparent un résumé produit de façon automatique avec un résumé de référence généralement produit manuellement.

Le principe utilisé dans ROUGE est de compter le nombre d'items, par exemple des n-grammes, communs entre le résumé automatique et les résumés de référence. Les mesures *ROUGE-n* comparent les n-grammes des résumés. La mesure peut s'écrire comme suit (Lin, 2004b) :

$$ROUGE - n = \frac{\sum_{S \in \{ResumesDeReference\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ResumesDeReference\}} \sum_{gram_n \in S} Count(gram_n)} \quad (17)$$

Où n correspond à la longueur des n-grammes considérés (généralement ROUGE-1 et ROUGE-2 sont choisis), $Count_{match}(gram_n)$ est le nombre maximum de n-grammes qui apparaissent à la fois dans le résumé produit automatiquement et dans les résumés de référence, $Count(gram_n)$ est le nombre de n-grammes qui apparaissent dans les résumés de référence. ROUGE-n est une mesure orientée rappel. ROUGE-1 est semblable à la mesure cosinus mais pour n>1 l'ordre des mots étant pris en compte, ROUGE est plus stricte que la mesure Cosinus (Alguliev *et al.*, 2011). La valeur du dénominateur augmente avec le nombre de résumés de référence. Ainsi, l'ajout de références différentes favorise des résumés automatiques variés. Par ailleurs un résumé automatique qui contient des n-grammes qui apparaissent dans plusieurs références sera favorisé.

Les variantes de ROUGE-n permettent d'affiner la capacité de la mesure à considérer les similarités des résumés par rapport à leurs n-grammes communs. Par exemple, la mesure ROUGE-L (*Longest Commun Subsequence*) considère la chaine de mots communs la plus longue entre le résumé automatique et le résumé de référence.

$$ROUGE - L = \frac{(1 + \beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \quad (18)$$

Où

$$R_{LCS} = \frac{LCS(X,Y)}{m} \text{ et } P_{LCS} = \frac{LCS(X,Y)}{n} \quad (19)$$

Avec $LCS(X,Y)$ la longueur de la chaîne de mots commune entre X et Y, m (resp. n) la longueur de X (resp. Y). ROUGE-L vaut 1 lorsque les deux chaînes sont identiques et 0 lorsque les chaines n'ont aucun mot en commun. La mesure ROUGE-W (*weighted LCS*) favorise la proximité des mots communs dans les deux chaines. Quant à elle, la mesure ROUGE-S (*Skip-Bigram Co-Occurrence Statistics*) permet de considérer des sauts entre les mots pour construire les n-grammes à comparer. Ces méthodes de construction des bi-grammes peuvent pénaliser des chaînes qui contiennent des mots communs mais dans des ordres différents, ROUGE-SU étant ROUGE-S en ajoutant les mots unitaires. ROUGE-SU4 par exemple considèrera les bi-grammes avec un maximum de 4 mots entre les mots ainsi que les uni-grammes lors de la comparaison des résumés de référence avec le résumé automatique à évaluer. Le détail des variantes ROUGE sont décrites dans (Lin, 2004b). La corrélation entre les évaluations manuelles et automatiques avec ROUGE varie de 0.80 à 0.94 (Louis, Nenkova, 2009a). Par ailleurs Owczarzak, Dang (2011) rapporte des corrélations de Pearson supérieures à 0.8 entre l'évaluation Pyramidale et ROUGE-SU4.

5.2.3. Divergence de Kullback-Leibler et de Jenssen-Shanon

Les mesures de divergence de Kullback-Leibler (KL) et de Jenssen-Shanon (JS) permettent de mesurer la similarité entre deux distributions de probabilité de termes $P(i)$ et $Q(i)$, représentant respectivement le résumé à évaluer et le résumé de référence. La mesure de KL est définie par :

$$D_{KL}(P||Q) = \sum_i \ln \frac{P(i)}{Q(i)} P(i) \quad (20)$$

La mesure de JS est une variation symétrique et lissée de KL et est définie par :

$$D_{JS}(P||Q) = \frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M) \quad (21)$$

Avec $M = (P+Q)/2$

Jenssen-Shanon est la mesure qui a été montrée comme ayant la plus forte correlation avec ROUGE (Louis, Nenkova, 2009b) mais cette mesure est sensible à la taille des résumés.

5.2.4. Mesure LogSim

Cette mesure a été introduite dans (Saggion *et al.*, 2010) afin de mesurer la similarité de contenu d'un résumé avec un résumé de référence; la mesure ne devant pas être sensible à la taille des résumés (de référence et à évaluer). Elle est définie par :

$$LogSim(T,S) = \sum_{t \in T} P(t|T) \times \frac{\min(\log(1+P(t|T)), \log(1+P(t|S)))}{\max(\log(1+P(t|T)), \log(1+P(t|S)))} \quad (22)$$

Où T est un ensemble de termes de référence et S l'ensemble de termes issus du résumé à évaluer. $P(t|X)$ est la probabilité conditionnelle $\frac{f_X(t)}{f_X}$ avec X étant T ou S. $f_E(t)$ correspond à la fréquence du terme t dans l'ensemble E.

Cette mesure a des propriétés identiques aux mesures de précision interpolée si la précision est définie comme le nombre de n-grammes dans le résumé de référence.

5.3. Campagnes d'évaluation

DUC (*Document Understanding Conference*, http://duc.nist.gov) est le premier programme d'envergure d'évaluation des résumés automatiques. Ce forum d'évaluation a débuté en 2001 à partir des initiatives développées dans le programme TIDES (*Translingual Information Detection Extraction and Summarization* du DARPA (*Defense Advanced Research Projects Agency*); il s'est terminé en 2007 pour être remplacé par la compagne TAC *(Text Analysis Conference*, www.nist.gov/tac/).

En 2001, le NIST proposait 60 ensembles de référence, la moitié pour l'apprentissage, l'autre moitié pour les tests. Chaque ensemble contenait les documents originaux en anglais, des résumés pour chaque document et des résumés multi-documents. Deux tâches de création de résumés totalement automatiques étaient proposées aux participants : d'une part pour des documents considérés individuellement et d'autre part pour un ensemble de documents. L'évaluation était manuelle : l'évaluateur qui avait créé l'ensemble de référence avait pour tâche d'évaluer les résumés créés automatiquement.

En 2002, les premières tentatives d'évaluation automatiques ont été introduites. En 2003, la procédure d'évaluation a été améliorée en s'intéressant à la fois à des résumés de références produits manuellement et constitués automatiquement ; ces résumés étant considérés comme des résumés « témoins ». L'évaluation était basée sur l'outil SEE (http://www.isi.edu/licensed-sw/see/) qui estime le pourcentage d'information des résumés de référence présent dans le résumé contruit automatiquement.

A partir de DUC 2004, les mesures ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin, 2004b) ont été utilisées. Au départ uniquement en anglais, les documents utilisés ont ensuite été également dans d'autres langues comme l'arabe. Les documents étaient issus de d'autres tâches comme la tâche TDT *(Topic Detection and Tracking*) en 2003 et 2004, *Novelty* en 2004.

A partir de DUC 2005, l'objectif a été de proposer une tâche plus focalisée et correspondant plus à un besoin d'utilisateurs (*topic*). Il s'agissait de synthétiser un ensemble de 25-50 documents correspondants à un besoin de réponse à une question en un résumé de moins de 250 mots. Des ensembles de 45 à 50 besoins d'information étaient considérés. Les résumés étaient alors évalués sur leur contenu mais également par rapport à leur lisibilité. L'évaluation manuelle se basait sur 5 qualités linguistiques (grammaticale, de non redondance, de compréhension des références, de focus et de cohérence) que devait posséder le résumé avec une échelle de valeur ainsi que sur la qualité de la réponse à la question initiale. L'évaluation du contenu quant à lui était évalué par les mesures ROUGE-2 et ROUGE-SU4 après radicalisation des mots. H. T. Dang (2005) rapporte une corrélation (Spearman) de 0,872 à 0,951 entre la valeur de ROUGE et l'évaluation manuelle de réponse à la question.

La campagne DUC a attiré de nombreux participants (en 2007, 32 équipes ont participé de 11 pays différents). La tâche de résumé automatique a été prolongée dans le programme TAC (*Text Analysis Conference*) qui vise plus généralement l'évaluation de technologies de traitement du langage naturel (H. Dang, 2008).

En 2008, une tâche de résumé d'opinion a été introduite, basée sur des besoins d'information type question/réponse et évaluée par l'évaluation pyramidale (Nenkova, Passonneau, 2004b) ainsi que par un score compris entre 1 et 10 sur la capacité du résumé à répondre à la question. En 2010, la notion de catégories des besoins des utilisateurs a été introduite (accident, santé, ?) pour favoriser des approches plus linguistiques.

La campagne d'évaluation INEX s'est également intéressée aux résumés automatiques. Nous pouvons citer en particulier les tâches *snippets* et *Tweet Contextualization* (Crouch *et al.*, 2012) (SanJuan *et al.*, 2012). Cette dernière a pour objet d'aider à la compréhension de messages très courts que sont les tweets. Le système doit proposer aux lecteurs d'un tweet un résumé de 500 mots constitué à partir de wikipédia. Cette tâche a été initiée en 2011. Elle est évaluée sur le caractère informatif du résumé via la mesure LogSim et sur la lisibilité à partir de questions fermées sur les aspects grammaticaux, résolution d'anaphores, etc.

6. Recherche d'images par l'exemple

6.1. Description du problème

Les systèmes de recherche d'images ont des objectifs variés, qui vont de la recherche d'images spécifiques à la navigation dans une collection d'images. Un utilisateur peut par exemple rechercher une image particulière (le tableau de la Joconde de Leonard de Vinci), une catégorie d'images ou un concept (un paysage avec des montagnes) ou alors de façon plus vague faire une recherche exploratoire parce qu'il n'a pas d'idée précise de ce qu'il recherche. Alors que l'expression des besoins en RI textuelle se fait généralement sous forme de mots-clés, en recherche d'images la requête peut également être une image exemple ("trouve moi toutes les images qui ressemblent à cette image, qui ont le même paysage, le même objet,...").

L'évaluation dans ce contexte se doit bien entendu de correspondre à la tâche de recherche mise en place par le SRI: on peut chercher à évaluer une recherche adhoc comme dans le cas de la recherche textuelle (que ce soit avec une requête textuelle ou une requête image), la détection d'objets, la classification d'images ou l'annotation automatique (comment extraire automatiquement de "bons" descripteurs textuels pour une image), ...(Clough *et al.*, 2010; Moëllic, Fluhr, 2006; *Challenge Pascal*, 2012).

Nous nous intéressons principalement dans cette section aux problématiques d'évaluation lorsque la requête utilisateur est une image (éventuellement complétée par des mots-clés). Deux tâches de recherche ressortent dans ce contexte:

– la recherche adhoc (trouve les images similaires à cette image);

– la détection de copies (cette image est-elle la copie d'une autre image ? y a-t-il des copies de mon image dans la collection ?). Cette tâche est cruciale dans un contexte de protection des droits d'auteur.

Les SRI capables de traiter ce genre de requêtes utilisent des approches de recherche d'images par le contenu (*Content-based Information Retrieval*): il s'agit d'utiliser des caractéristiques visuelles des images telles que la couleur, la texture, les formes... D'une manière générale, les approches proposées calculent des descripteurs basés sur

le contenu sur des caractéristiques visuelles globales (concernant l'image entière) et locales (concernant les objets de l'image) (Datta *et al.*, 2005).

6.2. Campagnes d'évaluation

La campagne d'évaluation la plus connue en recherche d'images est ImageCLEF. ImageCLEF fonctionne sur le même principe que la campagne TREC, avec des tâches de recherche évaluées variant d'une année sur l'autre. ImageCLEF a été lancée pour la première fois en 2003, dans le cadre de la campagne d'évaluation CLEF (Cross-Language Evaluation Forum). Parmi les tâches d'évaluation partant d'une image requête, on peut citer:

– la tâche de recherche d'images photographiques (2003-2009) (Paramita, Grubinger, 2010), dans laquelle les requêtes sont constituées d'images exemple et d'un descriptif textuel;

– la tâche de recherche d'images dans Wikipedia (2008-2011) (Tsikrika, Kludas, 2010), pour laquelle là encore les requêtes sont consituées d'images exemple et d'un descriptif textuel. Le but ici est de se confronter à des collections d'images plus grandes (230 000 images pour la collection Wikipedia);

– la tâche de recherche d'images médicales (2004-2012) (Müller, Kalpathy-Cramer, 2010) : là encore les requêtes sont constituées d'images exemple et d'un descriptif, mais les problématiques évaluées sont propres à la recherche médicale (recherche d'images suivant une certaine modalité (radios, scanners, images provenant de microscopes,...), recherche de cas cliniques, ...);

Une autre initiative a également vu le jour en 2006 en France: la campagne ImagEval (Moëllic, Fluhr, 2006), fondée par le programme français "Techno-Vision". La campagne s'est intéressée à la recherche adhoc, la détection d'objets mais aussi la détection d'images transformées. Faute de financements, la campagne n'a malheureusement pas été poursuivie les années suivantes.

Toutes les campagnes précédemment citées travaillent sur des collections d'images de relativement petite taille. Dans le cadre du *Projet européen Quaero* (2012), nous avons participé, en partenariat avec Exalead http://www.exalead.fr) et l'équipe TeXMeX de l'INRIA de Rennes, à la construction de collections de tests de plus grande ampleur, utilisées pour des tâches de détection de copies :

– une collection de 1 128 000 d'images en deux résolutions (512 et 150 pixels sur leur plus grande largeur). Les images de cette collection proviennent principalement de photos de Flickr auxquelles ont été ajoutées des collections utilisées traditionnellement en recherche d'images (Caltech (Fei-Fei *et al.*, 2004), INRIA holidays (Jégou *et al.*, 2008)...). 1000 images sélectionnées aléatoirement ont subi 49 transformations différentes (rotation, compression, dégradation, ...) afin de servir de requêtes (Fabien A.P. Petitcolas, 1998).

– une collection de 100 million d'images (98 064 203 exactement) collectées sur le Web par Exalead. Ces images font sur leur plus grand côté 150 pixels, et le volume occupé par la collection est de plus de 2 To. Les requêtes utilisées sont les mêmes que pour la collection précédente.

6.3. Mesures et jugements de pertinence

En recherche d'images, l'efficacité et l'efficience sont très étroitement liées, de façon peut-être plus importante qu'en RI textuelle. En effet, le calcul des descripteurs

des images peut être très coûteux (de nombreuses approches cherchant à réduire leur dimension (Datta *et al.*, 2005)), et il faut alors trouver le bon équilibre entre qualité des résultats et temps d'exécution. C'est pour cette raison que souvent les mesures d'efficacité sont corrélées avec des mesures d'efficience (Moëllic, Fluhr, 2006), (*Projet européen Quaero*, 2012). Parmi les mesures d'efficience pouvant être prises en compte, on peut citer le temps de calcul des descripteurs, le nombre d'accès disque, le temps moyen de traitement d'une requête, la complexité de l'approche,...

Les mesures utilisées dans le cadre d'une recherche adhoc sont très similaires à celles de la recherche adhoc textuelle: on peut citer des mesures telles que la MAP, BPref, la R-Précision, P@r, nDCG, ... (Sanderson, 2010a), (Müller *et al.*, 2001). Une exception notable est la tâche de recherche d'images photographiques de ClefImage qui a cherché en 2009 à évaluer la diversité des résultats renvoyés. Chaque requête était associée à un certains nombres de catégories, et les participants ont été évalués en fonction du nombre de catégories présentes dans les 10 premiers résultats de recherche (mesure CR@10 - *Cluster recall* (Zhai *et al.*, 2003)).

Dans le cadre de la détection de copies, deux scénarios principaux de recherche peuvent être envisagés: (S1) trouver toutes les images étant des transformations d'une image requête (l'auteur d'une image veut savoir si elle a été utilisée dans d'autres contextes), (S2): étant donnée une image transformée, trouver l'image source (cette image semble être une copie, est-ce le cas?). Les transformations que peuvent subir une image sont variées: rotations, translations, qualité JPEG, ajout de bruit, ajout de texte,..(Fabien A.P. Petitcolas, 1998). Le premier scénario peut être évalué avec des mesures usuelles basées sur le rappel et la précision. Le second scénario est quelque peu différent: il se peut qu'il n'y ait pas d'image réponse dans la collection, et s'il y en a, il n'y en a qu'une. Les mesures proposées se focalisent donc sur les tous premiers résultats de recherche et/ou sur le fait que certaines requêtes n'ont pas de résultats dans la collection: (Moëllic, Fluhr, 2006; *Projet européen Quaero*, 2012):

– MRR (*Mean Reciprocal Rank*) (E. Voorhees, 1999). Il s'agit du rang de la première bonne réponse d'un système.

– NDCR (*Normalized Detection Cost Rate*) (Over *et al.*, 2010). Cette mesure, initialement utilisée pour la campagne TRECVID (évaluation de la recherche de vidéos) peut être utilisée afin de mesurer le coût associé à un système de détection de copies dans un scénario particulier. Par exemple, il est possible de prendre un scénario dans lequel la plupart des images requête ne sont pas des copies et pour lequel le coût de manquer une copie est plus grand que celui d'un faux positif. Chaque transformation d'images est évaluées séparément. Pour calculer cette mesure, tous les résultats de toutes les requêtes, pour chaque transformation d'images testée, doivent être concaténés et triés en fonction de leur score. Chaque fichier correspondant à une transformation particulière sur toutes les requêtes d'un *run*, va être utilisé pour calculer la probabilité de manquer une détection et le taux de faux positifs ((P_{miss} et R_{FA}). On a, pour un seuil de détection θ donné:

$$P_{Miss} = \frac{FN}{N_{target}} \tag{23}$$

$$R_{FA} = \frac{FP}{T_{queries}} \tag{24}$$

FN est le nombre de faux négatifs (c'est à dire le nombre de détections manquées), FP est le nombre total de faux positifs, N_{target} est le nombre de requêtes contenant une copie, et $T_{queries}$ est le nombre de requêtes. On a ensuite:

$$NCDR = P_{Miss} + \beta.R_{FA} \tag{25}$$

avec

$$\beta = \frac{Cost_{FA}}{Cost_{Miss} * R_{target}} \tag{26}$$

$Cost_{Miss}$ est le coût des détections manquées, $Cost_{FA}$ est le coût des fausses alarmes, et R_{Target} est une constante.

– Averaged MAP : nous avons proposé cette mesure dans le contexte de la campagne d'évaluation mise en place dans le cadre du *Projet européen Quaero* (2012). Pour chaque *run*, tous les résultats, y compris les résultats des requêtes pour lesquelles il n'y a pas de correspondant dans la collection sont concaténés et triés en fonction de leur score (ce qui implique que les scores sont comparables d'une requête à l'autre). La précision est ensuite évaluée à chaque point de rappel (c'est à dire à chaque image pertinente retrouvée). Nous avons ainsi:

$$Averaged - MAP = \frac{1}{|Pert|} \sum_{j=1}^{|Pert|} Precision(j) \tag{27}$$

où $|Pert|$ est le nombre total d'images pertinentes dans la collection.

Concernant les jugements de pertinenceutilisés pour le calcul de ces mesures, ils sont parfois "de fait" (c'est à dire liés à la collection), ou alors réalisés par des experts du domaine. Dans ce dernier cas, comme en RI textuelle, des techniques de *pooling* peuvent être utilisées.

7. Conclusion

Dans cet article, nous avons présenté le cadre général de l'évaluation des systèmes de RI. La RI s'intéresse à proposer des modèles et des systèmes permettant d'accèder à des informations numériques sous forme documentaire. Cet accès couvre en réalité des tâches très variées. La tâche de recherche *adhoc*, qui à partir d'une requête exprimée par l'utilisateur, vise à lui restituer les seuls documents qui répondent à son besoin, est celle qui a permis de poser les cadres de l'évaluation des systèmes de RI. Dans cet article, nous nous sommes donc attachés à présenter cette tâche ainsi que les mesures utilisées pour l'évaluer. Les autres tâches de RI se sont inspirées de la recherche *adhoc* lors de la définition des cadres d'évaluation.

Nous avons également présenté trois cadres d'évaluation spécifiques qui ne sauraient prétendre donner une vue exhaustive de l'évaluation en RI mais qui donnent un

Table 1. Résumé des cadres d'évaluation présentés

Type de recherche	Mesures principales	Campagnes d'évaluation
Adhoc	Rappel, précision par rapport à des listes de documents	TREC, CLEF
Images	Rappel, précision par rapport à des listes d'images	ImageCLEF, ImagEval, Quaero
Clustering	Rappel, précision par rapport à des groupes	INEX, Quaero
Résumé automatique	Précision au niveau des termes	DUC, TAC, INEX

aperçu de la variété à la fois des cadres, des mesures et des collections de référence existantes. Ces cadres d'évaluation sont résumés dans la Table 1.

L'existence de ces cadres d'évaluation est depuis l'origine de la RI un élément central du domaine. Nombreux sont les chercheurs qui, comme nous, pensons que l'existence de ces cadres et de ces collections de référence partagés ont permis des avancées notables dans le domaine. Ils assurent par ailleurs la reproductivité des approches et méthodes proposées ainsi que la comparaison des nouvelles propositions avec des méthodes de la littérature.

Avec l'arrivée de nouveaux médias, en particulier les médias sociaux, ainsi qu'avec la production toujours plus importante d'information, les besoins d'accès à l'information sont en train d'évoluer et de nouvelles tâches de RI apparaissent. Les campagnes d'évaluation et les collections qu'elles produisent sont, au contraire de la réalité de ces sources, figées. Fixer les contenus confère l'avantage de la reproductibilité ainsi que de la possibilité de comparer de nouvelles approches avec des résultats obtenus précédemment. Cependant, de nombreux aspects liés à l'évolution des informations ne peuvent pas être évalués sur ce type de collection. La temporalité ou la fraîcheur sont des éléments qui doivent être pris en compte dans l'évaluation de la pertinence et qui ne le sont pas ou peu actuellement. Par ailleurs, les collections produites jusqu'ici sont assez homogènes alors que dans la réalité, certains besoins d'information ne peuvent être réellement satisfaits que par l'utilisation conjointe de plusieurs sources variées. Par exemple connaitre l'image numérique d'une entreprise nécessite de consulter à la fois des blogs, des flux RSS, des tweets, ... Des collections d'évaluation multi-sources sont très difficiles à mettre en place, spécifiquement lorsqu'il s'agit d'évaluation de la pertinence. Enfin, face à la masse d'information, les vues synthétiques de l'information deviennent nécessaires. Elles se développent dans des activités de veille, d'intelligence économique. Dans ce contexte, il est difficile de s'appuyer sur des collections dans la mesure où l'objectif est souvent de prédire des tendances futures à partir des informations analysées.

Les études basées sur des utilisateurs sont une réponse qui ne permet pas de passer à l'échelle et qui n'assure pas une reproductibilité des résultats obtenus. La définition de ce que pourraient être des collections de référence dans ces cadres (qui ne sauraient être exhaustifs) reste à définir.

References

Alguliev R. M., Aliguliyev R. M., Hajirahimova M. S., Mehdiyev C. A. (2011). Mcmr: Maximum coverage and minimum redundant text summarization model. *Expert Syst. Appl.*, Vol. 38, No. 12, pp. 14514-14522.

Alonso O., Rose D. E., Stewart B. (2008). Crowdsourcing for relevance evaluation. *SIGIR Forum*, Vol. 42, No. 2, pp. 9–15.

Baccini A., Déjean S., Lafage L., Mothe J. (2012). How many performance measures to evaluate information retrieval systems? *Knowl. Inf. Syst.*, Vol. 30, No. 3, pp. 693-713.

Baeza-Yates R. A., Ribeiro-Neto B. A. (2011). *Modern information retrieval - the concepts and technology behind search, second edition*.

Bonnel N., Chevalier M., Dousset B., Hubert G. (2008). Visualisation et recherche d'information. In *Information & visualisation – enjeux, recherches et applications.*

Buckley C., Voorhees E. M. (2000). Evaluating evaluation measure stability. In *Acm sigir conference on research and development in information retrieval*, pp. 33–40.

Cabanac G., Hubert G., Boughanem M., Chrisment C. (2010). Tie-breaking bias: Effect of an uncontrolled parameter on information retrieval evaluation. In *Multilingual and multimodal information access evaluation*, Vol. 6360, p. 112-123.

Cao Z., Qin T., Liu T.-Y., Tsai M.-F., Li H. (2007). Learning to rank: from pairwise approach to listwise approach. In *International conference on machine learning*, pp. 129–136. ACM.

Carpineto C., Osiński S., Romano G., Weiss D. (2009). A survey of web clustering engines. *ACM Comput. Surv.*, Vol. 41, No. 3, pp. 1–38.

Challenge pascal. (2012). Retrieved from http://pascallin.ecs.soton.ac.uk/challenges/VOC/

Clarke C. L., Kolla M., Cormack G. V., Vechtomova O., Ashkan A., *et al.* B. (2008). Novelty and diversity in information retrieval evaluation. In *International acm sigir conference on research and development in information retrieval*, pp. 659–666.

Cleverdon C. W. (1967). The cranfield tests on index languages devices. In *Aslib proceedings, volume 19, pages 173-192.*

Clough P., Müller H., Sanderson M. (2010). Seven years of image retrieval evaluation. In *Imageclef, experimental evaluation in visual information retrieval*, chap. 1.

Crabtree D., Gao X., Andreae P. (2005). Standardized evaluation method for web clustering results. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 280–283.

Crouch C. J., Crouch D. B., Chittilla S., Nagalla S., Kulkarni S., Nawale S. (2012). The 2012 inex snippet and tweet contextualization tasks. In *Clef (online working notes/labs/workshop).*

Dang H. (2008). Overview of the TAC 2008 Opinion Question Answering and Summarization Tasks. In *Proc. of the first text analysis conference.*

Dang H. T. (2005). Overview of DUC 2005. In *Document understanding conference.*

Datta R., Li J., Wang J. Z. (2005). Content-based image retrieval: approaches and trends of the new age. In *ACM SIGMM international workshop on Multimedia information retrieval*, pp. 253–262.

Denoyer L. (2004). *Apprentissage et inférence statistique dans les bases de documents structurés*. Thèse de doctorat, University of Paris VI.

Ellen M. Voorhees D. K. H. (2005). *Trec: Experiment and evaluation in information retrieval (digital libraries and electronic publishing)*. The Mit Press.

Fabien A.P. Petitcolas M. G. K., Ross J. Anderson. (1998). Attacks on copyright marking systems. In *Second International Workshop IH-98, Portland, Oregon, USA*.

Fei-Fei L., Fergus R., Perona P. (2004). Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *Ieee workshop on generative-model based vision.*

Gordon M. D. (1997). It's 10 a.m. do you know where your documents are? the nature and scope of information retrieval problems in business. *Information Processing and Management*, Vol. 33, No. 1, pp. 107 - 122.

Halkidi M., Batistakis Y., Vazirgiannis M. (2001). On clustering validation techniques. *J. Intell. Inf. Syst.*, Vol. 17, No. 2-3, pp. 107–145.

Hennig L., De Luca E. W., Albayrak S. (2010). Learning summary content units with topic modeling. In *Proc. coling: 2010*, pp. 391–399.

Hull D. (1993). Using statistical testing in the evaluation of retrieval experiments. In *Acm sigir conference on research and development in information retrieval*, pp. 329–338.

Ingaramo D., Pinto D., Rosso P., Errecalde M. (2008). Evaluation of internal validity measures in short-text corpora. In *Computational linguistics and intelligent text processing*, pp. 555–567.

Jardine N., Rijsbergen C. J. van. (1971). The use of hierarchic clustering in information retrieval. *Inform. Stor. Retr.*, Vol. 7, No. 5, pp. 217–240.

Järvelin K., Kekäläinen J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, Vol. 20, No. 4, pp. 422–446.

Jégou H., Douze M., Schmid C. (2008). Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of eccv.*

Jones K. S. (2007). Automatic summarising: The state of the art. *Information Processing and Management*, Vol. 43, No. 6, pp. 1449 - 1481.

Jones K. S., Rijsbergen C. V. (1975). *Report on the need for and provision for an 'ideal' information retrieval test collection*. Technical report.

Julien C.-A., Leide J. E., Bouthillier F. (2008). Controlled user evaluations of information visualization interfaces for text retrieval: Literature review and meta-analysis. *J. Am. Soc. Inf. Sci. Technol.*, Vol. 59, No. 6, pp. 1012–1024.

Kelly D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, Vol. 3, No. 1–2, pp. 1–224.

Koshman S., Spink A., Jansen B. J. (2006). Web searching on the Vivisimo search engine. *JASIST*, Vol. 57, No. 14, pp. 1875-1887.

Lamprier S. (2008). *Vers la conception de documents composites : Extraction et organisation de l'information pertinente*. Unpublished doctoral dissertation, Université d'Angers.

Lamprier S., Amghar T., Levrat B., Saubion F. (2010). Organiser les résultats d'une recherche d'information – clustering, répartition de l'information et facilité d'accès. *Document Numérique*, Vol. 13, No. 1, pp. 9–39.

Leuski A. (2001). Evaluating Document Clustering for Interactive Information Retrieval. In *CIKM'01*, pp. 33–40. ACM.

Lewis D. D., Yang Y., Rose T. G., Li F. (2004). Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, Vol. 5, pp. 361–397.

Lin C. Y. (2004a). ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the workshop on text summarization branches out (was).* Barcelona, Spain.

Lin C. Y. (2004b). ROUGE: A package for automatic evaluation of summaries. In *Workshop on text summarization branches out (was).* Barcelona, Spain.

Liu Y., Fu Y., Zhang M., Ma S., Ru L. (2007). Automatic search engine performance evaluation with click-through data analysis. In *World wide web conference*, pp. 1133–1134.

Louis A., Nenkova A. (2009a). Automatically evaluating content selection in summarization without human models. In *Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 1 - volume 1*, pp. 306–314.

Louis A., Nenkova A. (2009b). Performance confidence estimation for automatic summarization. In *Eacl*, p. 541-548.

Luhn H. (1958). The automatic creation of literature abstracts. *IBM Journal*, Vol. 2, pp. 159-165.

Mani I., Klein G., House D., Hirschman L., Firmin T., Sundheim B. (2002). Summac: a text summarization evaluation. *Nat. Lang. Eng.*, Vol. 8, No. 1, pp. 43–68.

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

Müller H., Kalpathy-Cramer J. (2010). The medical image retrieval task. In *Imageclef, experimental evaluation in visual information retrieval*, chap. 13.

Müller H., Müller W., Squire D. M., Marchand-Maillet S., Pun T. (2001). Performance evaluation in content-based image retrieval: overview and proposals. *Pattern Recognition Letters*, Vol. 22, No. 5, pp. 593 - 601.

Moëllic P.-A., Fluhr C. (2006). *Imageval 2006 official campaign, http://www.imageval.org.* Voir aussi: http://cmm.ensmp.fr/ marcoteg/ImagEval_e.htm.

Morris A. H., Kasper G. M., Adams D. A. (1992). The effects and limitations of automated text condensing on reading comprehension performance. *Information Systems Research*, Vol. 3, No. 1, pp. 17-35.

Navarro E., Chudy Y., Gaume B., Cabanac G., Pinel-Sauvagnat K. (2011). Kodex ou comment organiser les résultats d'une recherche d'information par détection de communautés sur un graphe biparti ? . In *Conférence CORIA*, pp. 25–40.

Nayak R., Vries C. M. D., Kutty S., Geva S., Denoyer L., Gallinari P. (2009). Overview of the inex 2009 xml mining track: Clustering and classification of xml documents. In *Inex*, p. 366-378.

Nenkova A., Passonneau R. (2004a). Evaluating content selection in summarization: The Pyramid method. In *Proceedings of hlt/naacl2004.*

Nenkova A., Passonneau R. (2004b). Evaluating content selection in summarization: The pyramid method. In *Hlt-naacl*, Vol. 2004.

Ounis I., Amati G., Plachouras V., He B., Macdonald C., Johnson D. (2005). Terrier Information Retrieval Platform. In *European Conference on IR Research*, pp. 517–519.

Over P., Awad G. M., Fiscus J., Michel M., Smeaton A. F., Kraaij W. (2010). TRECVID 2009 - goals, tasks, data, evaluation mechanisms and metrics. In *TRECVid Workshop 2009.*

Owczarzak K., Dang H. T. (2011). Overview of the TAC 2011 summarization track: Guided task and AESOP task. In *Text analysis conference (tac 2011).*

Paramita M. L., Grubinger M. (2010). Photographic image retrieval. In *Imageclef, experimental evaluation in visual information retrieval*, chap. 7.

Projet européen quaero. (2012). http://www.quaero.org.

Rijsbergen C. J. V. (1979). *Information retrieval* (2nd ed.). Newton, MA, USA, Butterworth-Heinemann.

Saggion H., Torres-Moreno J.-M., Cunha I. da, SanJuan E., Velázquez-Morales P. (2010). Multilingual summarization evaluation without human models. In *Coling (posters)*, p. 1059-1067.

Sanderson M. (2010a). Performance measures used in image information retrieval. In *Imageclef, experimental evaluation in visual information retrieval*, chap. 5.

Sanderson M. (2010b). Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, Vol. 4, No. 4, pp. 247–375.

Sanderson M., Zobel J. (2005). Information retrieval system evaluation: effort, sensitivity, and reliability. In *Acm sigir conference on research and development in information retrieval*, pp. 162–169.

SanJuan E., Moriceau V., Tannier X., Bellot P., Mothe J. (2012). Overview of the inex 2011 question answering track - tweet contextualization track. *Lecture Notes in Computer Science*, Vol. 7424, pp. 188-206.

Saracevic T. (1995). Evaluation of evaluation in information retrieval. In *Proc. ACM SIGIR conference on Research and development in information retrieval*, pp. 138–146.

Smucker M. D., Allan J., Carterette B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Acm conference on conference on information and knowledge management*, pp. 623–632.

Tombros A. (2002). *The effectiveness of hierarchic query-based clustering of documents for information retrieval*. Thèse de doctorat, University of Glasgow, UK.

Tsikrika T., Kludas J. (2010). The wikipedia image retrieval task. In *Imageclef, experimental evaluation in visual information retrieval*, chap. 9.

Vivaldi J., Da Cunha I., Ramírez J. (2011). The reg summarization system with question reformulation at qa@inex track 2010. In *International conference on initiative for the evaluation of xml retrieval: comparative evaluation of focused retrieval*, pp. 295–302.

Voorhees E. (1999). The trec-8 question answering track report. In *Proceedings of the 8th Text Retrieval Conference (TREC-8). NIST, Gaithersburg, MD*, p. 77-82.

Voorhees E. M. (2002). The philosophy of information retrieval evaluation. In *Evaluation of cross-language information retrieval systems*, Vol. 2406, p. 355-370.

Voorhees E. M. (2006). Overview of the trec 2006. In *Trec.*

Vries C. M. D., Nayak R., Kutty S., Geva S., Tagarelli A. (2011). Overview of the inex 2010 xml mining track : clustering and classification of xml documents. In *Initiative for the Evaluation of XML Retrieval (INEX) 2010.* Amsterdam, Springer.

Zhai C. X., Cohen W. W., Lafferty J. (2003). Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Acm sigir conference on research and development in informaion retrieval*, pp. 10–17.

Zhao L., Wu L., Huang X. (2009). Using query expansion in graph-based approach for query-focused multi-document summarization. *Inf. Process. Manage.*, Vol. 45, No. 1, pp. 35–41.

Zobel J. (1998). How reliable are the results of large-scale information retrieval experiments? In *ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 307-314.