Quel avenir pour les moteurs de recherche?

Nicolas Bonnel[⋆,†], **Fabienne Moreau**[†]

[⋆] France Telecom, Division R&D,
4, rue du Clos-Courtel, BP 91226, 35512 Cesson-Sévigné Cedex, France
nicolas.bonnel@francetelecom.com

[†] IRISA, Campus universitaire de Beaulieu, 35012 Rennes Cedex, France
{nicolas.bonnel,fabienne.moreau}@irisa.fr

Résumé : Cet article présente deux problématiques majeures liées au domaine de la recherche d'information textuelle et plus particulièrement à l'utilisation des moteurs de recherche sur Internet. La première concerne la pertinence des résultats retournés à l'utilisateur. La principale limite de ces outils est, en effet, de ne pas toujours retrouver l'information précise que recherche l'utilisateur, ce qui rend la tâche de recherche d'information particulièrement frustrante. La seconde problématique est liée à la restitution des résultats par les moteurs qui n'offrent pas à l'utilisateur la possibilité d'exploiter et de visualiser efficacement les informations retournées. Concernant ces deux aspects, cet article fait état des méthodes actuellement utilisées par les moteurs de recherche et de leurs limites. En prenant en compte ces critiques, des solutions sont proposées pour améliorer les performances de ces outils.

Mots-clés : moteurs de recherche, recherche d'information textuelle, systèmes d'information, restitution des résultats de recherche, traitement automatique des langues, *web mining*.

1 INTRODUCTION

Une étude réalisée par [Lyman, 2003] révèle qu'environ 800 Mo d'informations enregistrées sont produites par personne chaque année. Cette constante augmentation de la quantité de données n'échappe pas au Web et certains moteurs de recherche référencent déjà plus de 8 milliards de pages. Ces informations sont de nature multimédia, mais compte tenu des techniques spécifiques à chaque média, cet article ne traite que de l'information textuelle. La recherche d'information (RI) a donc de plus en plus besoin d'outils efficaces pour retrouver les documents recherchés par l'utilisateur. Parmi ces outils, les moteurs de recherche sont devenus incontournables. En effet, selon [Sullivan, 2003] 625 millions de requêtes sont effectuées par jour sur les principaux moteurs. Les utilisateurs rencontrent cependant deux difficultés majeures dans l'utilisation actuelle de ces outils. La première est liée à la pertinence des résultats retournés. Il est, en effet, fréquent de ne pas retrouver, parmi les réponses fournies, l'information recherchée. La seconde concerne la façon dont sont restitués les résultats. En effet, même si l'information recherchée est présente dans la liste des documents retournés par les systèmes, elle n'est pas toujours facilement accessible pour l'utilisateur. Ces deux problématiques sont essentielles pour l'avenir des moteurs de recherche et plus globalement de la RI.

Cet article propose, en section 2, une description simplifiée du fonctionnement des moteurs de recherche. La section 3 présente plus précisément les mécanismes de RI mis en œuvre. Les techniques utilisées étant responsables de la qualité des réponses fournies à l'utilisateur, l'accent est plus particulièrement mis sur la manière dont l'information textuelle est représentée et traitée. La section 4 s'intéresse à la restitution des résultats. L'objectif est d'exploiter au mieux ces résultats afin de proposer à l'utilisateur une organisation et visualisation qui lui permettent d'accéder immédiatement à l'information qu'il recherche. La dernière section synthétise les propositions faites et présente quelques directions de recherche, afin de s'orienter progressivement vers une recherche d'information « intelligente » .

2 DESCRIPTION DES MOTEURS DE RECHERCHE

Avant de décrire le fonctionnement des moteurs de recherche, nous définissons l'objectif de tout système de recherche d'information (SRI). Selon [Salton, 1983b], un SRI traite de la représentation, du stockage, de l'organisation et de l'accès aux éléments de l'information. En d'autres termes, un SRI est un outil informatique qui représente et stocke l'information pour que cette dernière puisse être retrouvée automatiquement. D'une manière simplifiée, le fonctionnement général d'un SRI est le suivant : l'utilisateur, qui recherche une information, accède au système en formulant une requête ; le système tente alors de retrouver les documents[1] pertinents pour cette requête et les retourne à l'utilisateur. Le processus de RI se décompose donc en deux tâches principales : la phase d'indexation automatique qui consiste à extraire et stocker sous une forme facilement exploitable le contenu sémantique des documents de la collection, et la phase

[1] Le terme document tel qu'il est utilisé dans cet article désigne l'unité textuelle qui est retournée à l'utilisateur, et correspond à une page Web ou plus généralement à un texte de longueur variable.

de recherche et d'interrogation qui concerne la formulation du besoin d'information de l'utilisateur sous la forme d'une requête, mais également la recherche de documents dans la collection indexée et la présentation des résultats à l'utilisateur.
Un moteur de recherche, comme ceux utilisés pour accéder aux informations du Web [Google], constitue un cas particulier de SRI. La caractéristique principale de ce type d'outils par rapport à un SRI classique est l'ajout d'une phase supplémentaire aux deux précédentes : la collecte des données issues du Web. Une autre particularité des moteurs de recherche est de traiter généralement des documents appartenant à un domaine sémantique ouvert (les pages Web peuvent traiter de n'importe quel type de sujets) contrairement aux SRI qui sont le plus souvent dédiés à des thématiques.
Un moteur de recherche est composé de trois modules principaux (voir figure 1). Le premier concerne la collecte automatique des données. Un robot logiciel (souvent appelé *crawler* ou *spider*) a pour mission de parcourir de liens en liens les milliards de pages du Web et de recenser les adresses des sites visités. Le rôle du second module est d'analyser les pages précédemment collectées et de stocker leur contenu (généralement représenté par un ensemble de mots-clés) et leur adresse dans un index. Il s'agit de l'étape d'indexation automatique. Le dernier module dit de recherche consiste, après que l'utilisateur ait formulé sa requête, à interroger l'index et à présenter les résultats à l'utilisateur.
La section suivante (section 3) s'intéresse tout d'abord aux mécanismes mis en œuvre par ces systèmes pour indexer et rechercher les documents. Les techniques utilisées concernant à la fois les moteurs de recherche mais également les SRI, nous nous replaçons donc pour cette description dans le cadre plus général des SRI. La section 4 décrit ensuite la phase de présentation des résultats à l'utilisateur et s'applique plus particulièrement aux cas des moteurs de recherche.

3 INDEXATION ET RECHERCHE DE DOCUMENTS

Cette section met l'accent en premier lieu, sur la manière dont les systèmes stockent l'information textuelle contenue dans les documents et sur les mécanismes de modélisation généralement utilisés pour faciliter l'accès à cette information. Elle aborde ensuite les limites actuelles de ces méthodes. Enfin, compte tenu de ces critiques, elle s'achève par une description des perspectives proposées afin d'améliorer la qualité des résultats retournés par le système.

3.1 Présentation du mécanisme d'indexation

La principale difficulté pour les SRI est d'établir une correspondance entre l'information recherchée par l'utilisateur et l'ensemble des documents disponibles. Pour cela, l'étape d'indexation joue un rôle primordial puisqu'elle consiste à analyser au préalable les documents et la requête et à créer une représentation formelle de leur contenu. Une fonction de correspondance est alors définie afin de comparer les représentations internes des documents et de la requête. Les documents dont le contenu est le plus similaire à celui de la requête sont retournés à l'utilisateur. Le choix du cadre formel pour définir à la fois la représentation des documents et des requêtes caractérise le modèle de RI (*cf.* sous-section 3.2).
Il existe différentes techniques plus ou moins complexes pour représenter de façon formelle le contenu des documents[2]. L'approche généralement adoptée consiste à décrire le contenu sémantique des documents et requêtes en utilisant les mots qui les composent. Les mots d'un document ou d'une requête n'étant pas tous significatifs, le processus d'indexation revient alors à identifier et à extraire uniquement les mots les plus représentatifs de leur contenu. Pour cela, ce traitement, basé essentiellement sur des méthodes statistiques, s'appuie notamment sur la notion de fréquence[3] et consiste à admettre qu'un mot qui apparaît fréquemment dans un texte représente un concept important [Salton, 1983b]. Néanmoins, pour éviter le problème des mots fréquents mais non significatifs, une liste dite de mots vides (tels que les articles, les prépositions) est utilisée pour éliminer tous les mots non porteurs de sens. Une fois les termes les plus représentatifs extraits, une pondération leur est appliquée afin de prendre en compte deux critères essentiels : le premier doit refléter l'importance du terme dans le document, le second concerne son pouvoir de discrimination [Salton, 1983b] (*i.e.* la capacité du terme à différencier les documents de la collection)[4]. Cette pondération varie selon les modèles de RI utilisés, le poids le plus usité correspond au *tf.idf*, où *tf (term frequency)* désigne le nombre d'occurrences du terme dans le document, et *idf (inverse document frequency)* détermine sa fréquence documentaire inverse (*i.e.* la valeur inverse du nombre de documents dans lesquels le terme est présent). Comme résultat de la phase d'indexation, pour chaque document et requête, un ensemble de termes pondérés est obtenu et utilisé pour représenter leur contenu. Le stockage de ces termes peut varier selon le modèle de RI utilisé. Une des structures de stockage couramment utilisée est le fichier inversé qui contient tous les termes d'indexation, classés par ordre alphabétique, avec l'adresse précise de leurs occurrences dans les documents. Le but de cette structure est d'accélérer l'accès à l'information.

3.2 Modèles de RI

Toute méthode d'indexation automatique de documents repose sur le choix au préalable d'un modèle de RI qui permet de définir à la fois, le type de formalisation utilisé

[2] Seule l'indexation entièrement automatique est ici considérée. Nous limitons également notre définition à l'indexation dite en texte intégral (*full text*) par opposition aux méthodes qui n'indexent qu'une partie des documents (*e.g.* l'indexation des titres des pages).

[3] D'autres critères entrent également en jeu, *e.g.* la proximité des termes dans le document, leur position ou encore leur ordre d'apparition.

[4] En effet, un terme qui a une valeur de discrimination élevée doit apparaître seulement dans un petit nombre de documents. Et inversement, un terme contenu dans tous les documents de la collection n'est pas discriminant.

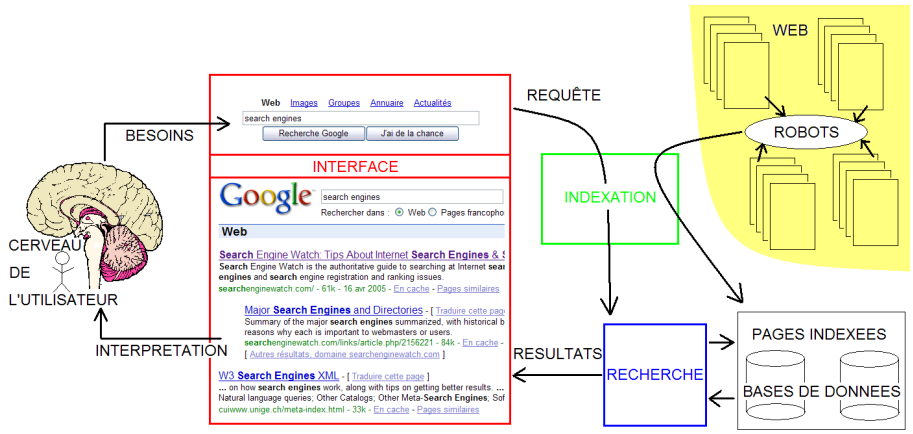


FIG. 1 – Description simplifiée du fonctionnement d'un moteur de recherche. Les parties **recherche des résultats pertinents** (en bleu) et **indexation** (en vert) sont abordées dans la section 3 et la partie **interface** (en rouge) est traitée dans la section 4.

pour la représentation du contenu des documents et des requêtes, les stratégies d'appariement à mettre en œuvre pour évaluer la pertinence des documents par rapport à la requête de l'utilisateur et les méthodes utilisées pour classer les documents.

Il existe une grande variété de modèles de RI [Baeza-Yates, 1999], principalement répartis autour de trois familles : les modèles booléen, vectoriel et probabiliste. Nous présentons dans cet article le principe général des deux premiers modèles. Pour cette description, nous utilisons la représentation suivante : soit $D = \{d_1, \ldots, d_i, \ldots, d_m\}$ un ensemble de documents d'une collection (qui peut correspondre à un ensemble de pages Web pour le cas d'un moteur de recherche, ou à un ensemble de textes d'un domaine donné pour le cas d'un SRI) et $T = \{t_1, \ldots, t_j, \ldots, t_n\}$ un ensemble de termes indexant ces documents.

Pour le modèle booléen, chaque document d_i est représenté par un ensemble de termes non pondérés sous la forme d'une expression logique : une conjonction des termes qu'il contient. La requête q est une expression booléenne dont les termes sont reliés par les opérateurs de conjonction, disjonction ou de négation. Un document d_i correspond à une requête q s'il vérifie l'implication logique : $d_i \rightarrow q$. L'inconvénient majeur de ce modèle est de considérer les documents qui ne contiennent pas tous les termes de la requête comme non pertinents (une requête composée des termes t_1, t_2 et t_3 ne retournera pas, par exemple, des textes contenant uniquement t_1 et t_2). Un document est donc soit pertinent, soit non pertinent, et par conséquent les réponses ne sont pas ordonnées. Le modèle booléen tel qu'il est actuellement utilisé au sein de certains moteurs de recherche est une extension de ce modèle classique [Salton, 1983a].

Le modèle vectoriel [Salton, 1983b] représente un document d_i et une requête q par un vecteur dans un espace à n dimensions : $\vec{d_i} = (\omega_{1i}, \ldots, \omega_{ji}, \ldots, \omega_{ni})$ et $\vec{q} = (\omega_{1q}, \ldots, \omega_{jq}, \ldots, \omega_{nq})$, où ω_{ji} est le poids du terme t_j dans le document d_i et ω_{jq} est le poids du terme t_j dans la requête q. La formule la plus utilisée pour calculer le poids des termes est le *tf.idf* décrit précédemment. Chaque document et requête étant représentés par un vecteur, il est alors possible de calculer un coefficient de similarité qui est généralement donné par la formule du cosinus (produit scalaire des vecteurs normalisés). Ainsi, si les termes d'indexation d'un document sont identiques à ceux utilisés dans la requête, l'angle entre le vecteur du document et celui de la requête est nul et la mesure de similarité est maximale. Un atout principal de ce modèle est de retourner en réponse à l'utilisateur une liste ordonnée de documents, classés dans l'ordre décroissant de leur degré de similarité avec la requête. La stratégie d'appariement partiel utilisée offre également l'avantage de présenter comme résultat des documents ne contenant pas nécessairement tous les termes de la requête. Les points négatifs du modèle vectoriel concernent principalement la représentation du contenu des documents : cette représentation dite en « sac de mots » (les documents sont transformés en vecteurs dont chaque composante représente un terme) présente, en effet, le désavantage d'ignorer l'ordre des mots[5] et de ne pas rendre compte des dépendances entre les termes[6].

Bien que tous les modèles de RI ne soient pas détaillés dans cet article, la présentation de ces deux exemples est suffisante pour pointer leurs faiblesses quant à la représentation qu'ils offrent du contenu informationnel des documents et aux stratégies d'appariement qu'ils utilisent. Ces insuffisances ont un impact direct sur la qualité des résultats fournis par les SRI à l'utilisateur, limites sur lesquelles nous nous étendons quelque peu ci-dessous.

[5] Par exemple, l'indexation des deux requêtes suivantes : *la voile du bateau* et *le bateau à voile* donne le même vecteur : *Q= voile, bateau= bateau, voile*.

[6] Ainsi, dans un exemple emprunté à Strzalkowski [Strzalkowski, 2000], les expressions suivantes : *information retrieval, retrieval of information, retrieve more information* et *information that is retrieved* entretiennent toutes la même relation de dépendance entre les termes : *retrieve* représente l'élément dominant (la tête de l'expression) et *information* est l'argument (le modifieur) de *retrieve*. L'intérêt de mettre en valeur ces dépendances est d'aboutir à la normalisation de ces différentes variantes syntaxiques en une seule et même forme : $retrieve + information$.

3.3 Limites des SRI

Compte tenu du mécanisme de mise en correspondance des documents et de la requête basé sur une simple comparaison de chaînes de caractères (pour le modèle booléen) ou de vecteurs de termes (pour le modèle vectoriel), les SRI se trouvent rapidement confrontés à plusieurs limites. La première est liée au fait qu'une même idée, un même concept peuvent être exprimés de différentes manières. La principale conséquence est de ne pas pouvoir retourner à l'utilisateur un document pertinent qui contient des termes « sémantiquement proches » de sa requête mais toutefois différents tels que des synonymes ou des hyperonymes[7]. Ainsi, une requête de l'utilisateur contenant par exemple le terme *voiture* ne pourra pas retrouver un document contenant le mot *automobile*. Ce mécanisme provoque la baisse des performances des systèmes qui ne peuvent pas proposer à l'utilisateur certains documents intéressants. Lorsque l'on cherche à évaluer les performances des SRI, ce phénomène se mesure par le biais du rappel (*i.e.* le rapport entre le nombre de documents pertinents trouvés par le système et le nombre total de documents pertinents). La seconde critique de ces systèmes concerne l'absence de traitement du phénomène de polysémie (*i.e.* prendre en compte le fait qu'un mot peut avoir plusieurs sens). Ainsi, le terme *serveur* présent dans la requête de l'utilisateur peut à la fois renvoyer à des documents parlant d'*informatique* ou de *restauration*. L'ambiguïté des termes conduit également à diminuer les performances des systèmes puisqu'elle entraîne la récupération de documents non pertinents. Ce phénomène s'évalue à l'aide de la mesure de précision (*i.e.* le rapport entre le nombre de documents pertinents retrouvés par le système et le nombre total de documents sélectionnés).

Ces deux limites sont directement liées à la complexité du langage naturel. Une solution souvent évoquée est donc d'intégrer, au sein de ces systèmes, une analyse linguistique qui présente l'avantage de ne plus considérer les mots comme de simples chaînes de caractères mais comme des entités linguistiques à part entière. Les traitements linguistiques en RI, effectués par le biais de techniques du traitement automatique des langues (TAL), extraient automatiquement des informations linguistiques des documents et des requêtes. Ces connaissances ont pour ambition de permettre aux systèmes de mieux comprendre les contenus textuels et d'avoir par conséquent un impact sur leurs performances. Les traitements linguistiques peuvent intervenir à différents niveaux d'un SRI. Pour ce qui est de l'indexation, ils contribuent, en exploitant les connaissances linguistiques extraites des textes, à créer une représentation plus riche de leur contenu ; cette représentation vise à obtenir un appariement plus pertinent entre l'information recherchée par l'utilisateur et les documents de la collection.

Nous revenons à présent plus en détail sur les différents types d'informations linguistiques qui peuvent être exploités lors de l'indexation des documents et requêtes et tentons de montrer comment nos travaux s'intègrent dans cette perspective.

3.4 Perspectives d'amélioration

Intégrer des connaissances linguistiques dans les SRI pour améliorer leurs performances n'est pas un phénomène nouveau [Moreau, 2005]. Parmi les travaux qui s'intéressent au couplage TAL/RI, trois types de connaissances linguistiques susceptibles d'être exploitées par les SRI sont traditionnellement distingués.

Les informations d'ordre morphologique peuvent tout d'abord être considérées. Leur objectif est de permettre aux systèmes de reconnaître, au sein des documents et requêtes, les différentes formes d'un même mot et de pouvoir les apparier, limitant ainsi la baisse de rappel due à cette variation morphologique[8]. Pour procéder à l'analyse morphologique des documents et des requêtes, les techniques issues du TAL sont variées, mettant en œuvre des outils plus ou moins complexes[9].

Le second type de connaissances linguistiques exploitables par un SRI appartient au niveau syntaxique de la langue. L'intégration de la syntaxe dans l'analyse des documents et requêtes a pour ambition d'extraire des entités linguistiques plus pertinentes car moins ambiguës. Il s'agit plus précisément de termes complexes ou structurés. Ainsi par exemple, l'expression *effet de serre*, considérée comme une seule et même unité, est plus spécifique dans son ensemble que chacun de ses termes pris isolément (*effet, serre*). L'analyse syntaxique offre aussi l'avantage de prendre en compte les différentes relations et dépendances qu'entretiennent les termes entre eux (*e.g.* dans l'exemple de *information retrieval* cité précédemment), allant ainsi plus loin qu'une simple représentation en « sac de mots » comme dans le cas du modèle vectoriel.

Enfin, des informations d'ordre sémantique peuvent également être intégrées dans un SRI. Ces connaissances sont utilisées, tout d'abord, pour réduire le problème de la formulation différente d'un même concept. L'objectif à atteindre est de ne plus considérer uniquement les termes des documents mais de prendre également en compte par exemple leurs synonymes ou des termes sémantiquement proches afin d'assouplir l'appariement entre documents et requête. L'exploitation d'informations sémantiques doit également permettre d'atténuer le phénomène de polysémie. L'idéal serait de s'orienter progressivement vers une indexation dite « sémantique » basée non plus sur les termes mais sur leur sens. Ainsi, pour reprendre l'exemple de *serveur* (*cf.* section précédente), une telle indexation consisterait à préciser

[7] L'hyperonyme est un incluant (*e.g. oiseau* par rapport à *rouge-gorge*).

[8] Un même mot peut en effet avoir plusieurs formes : *e.g. transformer, transforme, transformateur, transformation...* Ainsi, si aucune analyse morphologique n'est effectuée, une requête contenant la forme *transforme* ne peut pas être appariée avec un document employant par exemple la forme infinitive *transformer*.

[9] Les lemmatiseurs en sont un exemple. Par le biais d'une analyse morphologique sophistiquée, ils cherchent à reconnaître la forme de base (le lemme) de chaque mot en supprimant ses flexions (la forme de base d'un verbe est l'infinitif, celle d'un nom est sa forme au singulier...). Ainsi, *transformait, transforme, transforment* sont normalisés en une seule et même forme *transformer*.

que le terme apparaissant dans tel document particulier est utilisé dans un contexte informatique… Néanmoins, les méthodes actuellement utilisées pour intégrer ces informations sémantiques ne permettent pas d'atteindre de tels résultats. Des améliorations sont donc encore nécessaires (notamment à travers la mise en œuvre de traitements de désambiguïsation plus efficaces) pour que de telles connaissances puissent contribuer de manière significative à une hausse des performances des SRI.

Ces connaissances linguistiques, qu'elles soient d'ordre morphologique, syntaxique ou sémantique, ont toutes déjà fait l'objet de nombreuses expérimentations et ont été intégrées au moins partiellement au sein d'un modèle de RI particulier (généralement le modèle vectoriel ou probabiliste). Bien qu'il soit évident que ces informations permettent une représentation du contenu informationnel et du besoin de l'utilisateur plus riche que de simples mots-clés, les résultats de ces expériences sont cependant mitigés et ne mettent pas suffisamment en valeur l'intérêt d'intégrer de telles connaissances au sein d'un SRI. Nous avons donc cherché à comprendre les raisons des résultats plutôt décevants obtenus jusqu'à présent par les différentes études expérimentant ce couplage des techniques du TAL et de la RI. Une de nos hypothèses justifie ces résultats mitigés par le fait que les connaissances prises en compte au sein d'un SRI appartiennent généralement à un seul niveau de la langue (morphologique, syntaxique ou sémantique). En effet, très peu de systèmes ont cherché à combiner des informations appartenant aux trois niveaux. Nos travaux s'intéressent donc à la possibilité de coupler ces différentes informations au sein d'un même modèle de RI afin d'obtenir une caractérisation plus riche de l'information. Le principal obstacle rencontré est lié au modèle de RI utilisé pour représenter les documents et requêtes. Qu'ils soient booléens, vectoriels ou probabilistes, la plupart des modèles traitent l'élément textuel comme un ensemble de mots indépendants, inaptes à prendre en compte les relations entre les termes et difficilement ouverts à accueillir des informations aussi riches que celles fournies par les analyses linguistiques multi-niveaux des documents et des requêtes. Il est donc nécessaire d'adapter ou de modifier ces modèles afin d'exploiter au mieux les informations linguistiques que nous souhaitons intégrer. L'objectif final est de montrer que ces connaissances linguistiques, dès lors qu'elles sont bien couplées et intégrées efficacement au sein des modèles de RI, constituent de nouveaux critères de pertinence (au même titre que la fréquence ou la position des termes par exemple) susceptibles d'améliorer les performances actuelles des outils de recherche.

L'amélioration des mécanismes d'indexation et de recherche, telle que nous venons de l'évoquer, a pour principal objectif d'offrir des réponses plus pertinentes à l'utilisateur. Néanmoins, même si nous parvenons à retrouver davantage de documents pertinents, la présentation de ces résultats reste problématique. L'utilisateur rencontre, en effet, des difficultés à exploiter efficacement la liste des réponses retournées par les systèmes. Nous nous intéressons donc plus précisément dans la section suivante à cette phase de restitution des résultats. Cette étape étant particulièrement importante pour les utilisateurs du Web, nous limitons notre analyse au cas particulier des moteurs de recherche.

4 RESTITUTION DES RÉSULTATS

La restitution des résultats est la dernière étape d'un moteur de recherche (et plus généralement d'un SRI). Elle définit la façon dont les résultats sont présentés à l'utilisateur. Son objectif n'est pas de remettre en cause les réponses obtenues lors du processus de recherche (*cf.* section 3), mais de proposer une interface adaptée qui permette d'exploiter efficacement ces résultats. La restitution doit donc prendre en compte l'organisation et la visualisation des résultats ainsi que les aspects navigation et interaction.

4.1 Motivations

Le nombre de résultats issus d'une requête est devenu si important qu'il est nécessaire d'améliorer le processus de restitution. Actuellement la majorité des recherches sur le Web se font *via* des moteurs qui se contentent de retourner les résultats sous la forme d'une succession de listes ordonnées selon un critère représentant généralement le rang du résultat. Cette approche, bien que largement utilisée, ne permet pas à l'utilisateur d'exploiter efficacement les réponses qui sont devenus trop nombreuses. Cela est confirmé par une étude [iProspect, 2004] qui révèle que 81,7% des utilisateurs ne dépassent pas la troisième page de résultats (soit 30 documents). Ce comportement est paradoxal compte tenu de la grande diversité des résultats. Une solution pour améliorer la restitution (et donc la consultation) des résultats consiste à proposer une organisation efficace et une visualisation adaptée des documents. Concernant ces deux points, un bref aperçu de certaines méthodes est proposé dans le paragraphe suivant.

L'organisation des résultats est un aspect essentiel pour aider l'utilisateur à exploiter au mieux les réponses d'un moteur de recherche. Ce besoin d'organisation commence à apparaître dans certains moteurs tel que [Vivísimo] qui propose de regrouper les résultats similaires dans des répertoires grâce à une technique de *clustering* à la volée (calculée sur les premières réponses). Un autre exemple est [Grokker] qui présente les résultats dans des catégories afin d'éliminer les listes désorganisées de résultats qui rendent la recherche frustrante. La visualisation des résultats de recherche commence également à être prise en compte, sous forme 2D et plus rarement sous forme 3D. Pour ce type de visualisation, il existe de nombreuses techniques dont une taxonomie est proposée dans [Bonnel, 2005b]. Ces nouvelles propositions d'interface permettent alors de visualiser graphiquement certains attributs sur les documents mais surtout de visualiser les relations entre

les documents. La visualisation de ces relations se fait essentiellement sous forme de graphes ou de cartes. *Cat-a-Cone* [Hearst, 1997] est un exemple utilisant une visualisation 3D d'un arbre afin d'afficher simultanément les résultats obtenus et une hiérarchie de catégories. Concernant les cartes et plus généralement les approches géographiques [Skupin, 2003], il existe de nombreux exemples de visualisation tirant profit de l'aspect cognitif. Bien que les approches cartographiques en 2D soient les plus présentes [Kartoo, MapStan Search, Map.net], certains systèmes proposent des interfaces 3D [Boyack, 2002] [Sparacino, 2002].

Après cette courte introduction sur l'état actuel des méthodes de restitution, ce paragraphe permet de préciser le contexte dans lequel notre approche s'inscrit. Devant un grand nombre de résultats, deux directions sont possibles : réduire ou organiser les résultats. La première possibilité consiste à raffiner la requête afin de réduire le nombre de résultats. Cependant cette solution possède certains inconvénients tels que la discontinuité du mécanisme qui ne permet pas de comparer deux raffinements ou encore l'effort demandé à l'utilisateur pour préciser ses besoins. La deuxième possibilité consiste à organiser les résultats, ce qui nécessite aussi de rassembler les résultats similaires et donc de considérer les techniques de *clustering*. Cette deuxième approche est celle retenue dans cet article. Les documents à traiter sont les résultats issus d'une requête sur le Web, et seule la partie textuelle de ces documents est utilisée. Il est également important de préciser le nombre de résultats à considérer car ce choix influe directement sur les méthodes d'organisation et de visualisation. Bien que l'utilisateur semble actuellement se contenter des 30 premiers résultats, aucune étude n'a été réalisée sur des interfaces plus riches (cartes, graphes, univers 3D) et donc propices à l'exploitation d'un plus grand nombre de résultats. Pour cette première étude, seuls les 100 premiers résultats sont pris en considération.

4.2 Interface de restitution proposée

Cette partie présente les deux phases essentielles de l'interface de restitution : l'organisation des résultats et leur visualisation. Ensuite, les parties suivantes proposent une évaluation ainsi qu'une discussion sur différents aspects de l'interface.

4.2.1 Organisation des résultats

Chaque document est représenté par un vecteur de mots et de leurs fréquences. L'objectif est alors de regrouper les documents similaires mais surtout de pouvoir placer les documents dans l'espace des résultats en fonction de leur similarité. Ainsi deux documents proches dans l'espace des résultats doivent être relativement similaires. C'est dans ce contexte que la méthode des cartes auto-organisatrices [Kohonen, 1995] a été envisagée. Il s'agit d'une méthode de *clustering* qui respecte la notion de voisinage. Cette approche a déjà été utilisée avec succès sur des données textuelles dans le project Websom [Websom project]. L'organisation des résultats proposée peut être décomposée en trois étapes, détaillées dans les paragraphes suivants : le pré-traitement, l'algorithme d'organisation et le post-traitement.

Pré-traitement Considérant toute l'information textuelle contenue dans les pages Web, le nombre de mots extraits de ces pages afin de les représenter devient rapidement très important. Une sélection est alors réalisée. Elle consiste à ne conserver que les mots les plus fréquents dans l'ensemble des résultats. Chaque mot ainsi sélectionné est alors représenté par une variable. Comme précisé dans la section précédente, tous les mots n'ont pas la même importance. La pondération *tf.idf* est donc appliquée afin d'augmenter ou de diminuer l'importance de certains mots. Il est donc possible d'utiliser l'information présente dans les index (*cf.* sous-section 3.1). La sélection des mots ainsi que la nature des documents présents sur le Web (certaines pages contiennent très peu d'information textuelle) font que certains documents ne contiennent que très peu de variables non nulles. Ces documents peuvent être considérés comme mal représentés et risquent alors de biaiser l'organisation. C'est pourquoi une pondération sur les documents est introduite afin de donner plus de poids aux documents bien définis et moins d'importance à ceux qui sont mal définis. Elle peut, par exemple, être définie de façon proportionnelle au nombre de variables non nulles.

Algorithme L'organisation des résultats est basée sur un algorithme d'intelligence artificielle : les cartes auto-organisatrices (qui sont en fait un réseau de neurones particulier). La carte choisie dans cet article peut être assimilée à une grille 2D carrée avec une notion de 4-voisinage (voir figure 2a), et chaque unité de la carte représente un neurone. Dans notre cas, le calcul de la carte se fait avec les pondérations particulières introduites dans l'étape de pré-traitement. La carte permet alors d'organiser les documents en se basant uniquement sur la distribution des mots. Ainsi, les résultats similaires sont regroupés dans les mêmes neurones (ou unités de la carte). La carte possède aussi une notion de voisinage qui permet de placer les résultats proches dans des unités voisines de la carte. Les neurones sont alors étiquetés par les mots les plus représentatifs des documents qu'ils contiennent.

Post-traitement Afin de proposer un aperçu plus global des différentes thématiques de la recherche, un regroupement des neurones est réalisé. Il est basé sur un algorithme de classification ascendante hiérarchique et permet de regrouper les neurones similaires au sein de mêmes classes. Du fait de la proximité sémantique des neurones sur l'espace de projection, ces classes définissent différentes zones sur la carte qui correspondent alors à des « thèmes » différents (définis à partir des mots étiquetant les neurones).

Une fois les résultats organisés, il faut définir une

métaphore[10] de visualisation adaptée à l'organisation. Il s'agit donc de proposer une représentation graphique des résultats ou groupes de résultats, ainsi que des interactions.

4.2.2 Métaphore de visualisation

La métaphore de visualisation proposée est celle d'une ville (voir figure 2). Les résultats sont représentés par des bâtiments dont la texture correspond au contenu de la page Web. La pertinence des résultats est associée à la hauteur des bâtiments. Contrairement aux moteurs traditionnels, une représentation du rang basée sur des intervalles est utilisée. Par exemple, les 10 premiers résultats possèdent la même hauteur de bâtiment. La disposition au sol des bâtiments respecte l'organisation proposée en 4.2.1. Ainsi, la grille issue de l'algorithme d'organisation des résultats est plaquée sur le sol de la ville, permettant un découpage en quartiers où chaque quartier représente un neurone (ou groupe de résultats). D'après les propriétés de l'algorithme d'organisation, les résultats d'un même quartier sont similaires et deux quartiers voisins contiennent des résultats aussi proches que possible. De plus, l'étape de post-traitement permet de créer des zones « thématiques » sur la carte qui sont matérialisées par différentes couleurs sur le sol des quartiers. La métaphore 3D est accompagnée de parties 2D telles que le plan de la ville (vue aérienne), la zone d'expression de la requête (et des différents paramètres) ou encore l'affichage contextuel des informations sur les résultats (URL, *snippet*, mots-clés). D'ailleurs l'affichage contextuel des informations sur le résultat sélectionné est accompagné de l'affichage d'informations sur des résultats voisins et donc sémantiquement proches. La navigation dans la scène 3D n'est pas contrainte, l'utilisateur est donc libre de ses mouvements. En revanche, le plan 2D est interactif afin de faciliter les déplacements vers les différents quartiers.

4.3 Évaluation

Les propositions précédentes sur l'organisation et la visualisation des résultats sont intégrées dans un prototype [Bonnel, 2005a] dont des éléments d'évaluation sont présentés dans cette partie. La tâche d'évaluation est particulièrement importante dans un processus de restitution de l'information à l'utilisateur. Dans notre cas, cette évaluation est double étant donné la nécessité d'évaluer la qualité de l'organisation des données et la métaphore de visualisation. Concernant la métaphore de visualisation, un test utilisateur, basé sur les propositions de [Shneiderman, 1998] en matière d'évaluation d'interfaces, a été réalisé. Il faut cependant préciser que ce test porte sur une version antérieure de la métaphore. Les différences majeures par rapport à la version présentée sur la figure 2 sont les absences de la texture des bâtiments, de l'affichage contextuel des quatre résultats voisins et de l'interactivité du plan, ainsi qu'une organisation des résultats basée sur un placement selon deux axes (noms de domaine des résultats et mots-clés les plus fréquents). Le tableau 1 et la figure 3 présentent un extrait représentatif de ce test. Pour chaque question, l'utilisateur doit donner une note comprise entre 1 et 5, ou éventuellement ne pas se prononcer. Les résultats sont toutefois difficilement interprétables car les utilisateurs ont tendance à éviter les notes extrêmes au détriment des notes centrales. Il serait alors peut-être plus adapté de contraindre le système de notation ou de proposer des scénarios d'utilisation. Ce test révèle cependant que les utilisateurs ne trouvent pas d'inconvénient majeur à utiliser notre interface 3D. Le choix de la métaphore de la ville est d'ailleurs plutôt bien accepté (figure 3(f)) malgré la nécessité d'améliorer certains points tels que la navigation (figure 3(g)) ou la clarté du prototype (figure 3(a)). D'après la figure 3(d), les capacités de l'utilisateur doivent mieux être prises en compte ou alors il faut mieux cibler le public visé par cette métaphore. Enfin il semble que les utilisateurs soient prêts à utiliser des degrés d'abstraction sur les résultats tels que le *clustering* (figure 3(h)), ce qui est pris en compte dans la métaphore proposée sur la figure 2. Une autre évaluation plus complète est également envisagée afin notamment de comparer notre proposition avec d'autres interfaces. Concernant les interfaces de restitution des résultats de recherche, un projet intéressant serait de mettre en place une campagne d'évaluation similaire à celles existantes pour la phase de recherche du moteur [TREC, CLEF, INEX]. Cependant, cela nécessite de définir préalablement les critères

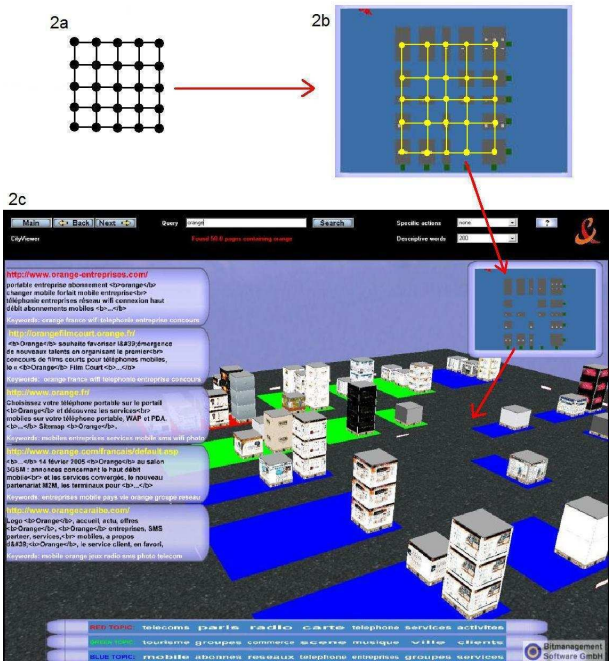


FIG. 2 – Métaphore de visualisation de type ville (2c). La grille issue de la carte auto-organisatrice (2a) est plaquée sur le sol de la ville pour obtenir une organisation sous forme de quartier. Cette superposition est visible sur la vue de dessus (2b) où la grille est représentée en jaune.

[10] Dans ce contexte, une métaphore est la réalisation d'une association entre des paramètres graphiques de la visualisation et des informations sur les résultats.

	NOTE (n)						
	1	**2**	**3**	**4**	**5**	**NE**	$\bar{n}$
QUESTION							
Clarté d'utilisation du prototype	5	17	22	8	7	1	2.9
Facilité d'utilisation du prototype	3	12	21	13	9	2	3.2
Apprendre à utiliser le prototype	2	11	12	17	17	1	3.6
Destiné à tous les niveaux d'utilisateurs	5	16	18	4	15	2	3.1
Compréhension de la métaphore ville	1	10	11	21	15	2	3.7
Pertinence de la métaphore ville	3	6	13	24	12	2	3.6
Navigation dans la métaphore ville	4	11	16	17	10	2	3.3
Utilité d'une visualisation regroupant certaines pages Web	1	1	17	11	23	7	4

TAB. 1 – Extrait du test utilisateur réalisé sur 60 personnes d'âges et de professions différents. La valeur **1** correspond à la plus mauvaise note et la valeur **5** à la meilleure note. La colonne **NE** représente les réponses non exprimées et la colonne $\bar{n}$ correspond à la note moyenne.

d'évaluation [Chevalier, 2005].

4.4 Discussion

Le regroupement des résultats similaires (du point de vue de leur sens) au sein de mêmes classes offre de multiples avantages. Il est notamment censé permettre à l'utilisateur de retrouver plus rapidement l'information pertinente, ce qui doit être vérifié par un nouveau test utilisateur. Grâce à ce regroupement, il est aussi possible d'avoir un classement distinct des résultats dans les différentes classes. Ainsi, il n'est plus nécessaire de comparer des résultats traitant de thèmes différents, ce qui est encore le cas de nombreux moteurs de recherche. Un autre avantage consiste à ne plus favoriser, dans l'espace des résultats, un thème par rapport aux autres. En plus de l'aspect classification, l'organisation des résultats (ou groupes de résultats) doit désormais être prise en compte. Cette organisation visuelle des résultats est un atout des interfaces 2D et 3D (par rapport à une simple exploitation linéaire d'une interface). L'idée consiste à diminuer les efforts de l'utilisateur en proposant de placer les informations sémantiquement proches dans des espaces voisins sur l'interface de visualisation.

Concernant l'interface de visualisation proposée, une particularité importante est l'utilisation de la 3D qui est cependant volontairement sous-exploitée dans cet exemple. Il est évident que la majorité des utilisateurs sont peu familier des interfaces 3D et notamment par la navigation dans de telles interfaces. Cependant l'émergence de la 3D dans certains produits, tels que les futurs environnements de bureau annoncés par *Sun Microsystems* (projet *Looking Glass*) ou *Microsoft* (librairie graphique *Avalon* de son prochain système d'exploitation baptisé *Longhorn*), risque de modifier les habitudes des utilisateurs face à la 3D.

5 CONCLUSION

Cet article présente deux problématiques essentielles liées à la RI sur le Web : le problème de la pertinence des résultats d'une requête et celui de la présentation des résultats à l'utilisateur. À travers la description des mécanismes utilisés par les outils de recherche pour traiter ces deux problèmes, certaines de leurs limites sont identifiées et de nouvelles perspectives sont proposées afin d'améliorer les résultats de ces systèmes. Ces deux problématiques mettent en œuvre des traitements différents : l'intégration d'informations linguistiques issues des techniques du TAL pour la phase de recherche et l'exploitation des résultats par le biais de techniques issues du *web mining* ou de la visualisation pour la phase de restitution. Cependant ces traitements peuvent et doivent être couplés dans un même moteur de recherche. En effet, l'objectif est identique ; il concerne la nécessité de traiter plus finement l'information textuelle à tous les niveaux d'un moteur de recherche. Il est donc nécessaire de mieux représenter, stocker, organiser et exploiter l'information pour concevoir des outils de recherche de plus en plus efficaces et performants, capables de traiter ces masses de données en constante évolution.

BIBLIOGRAPHIE

[Baeza-Yates, 1999] Baeza-Yates R. et Ribeiro-Neto B., *Modern Information Retrieval*. ACM Press / Addison-Wesley, New York, États-Unis.

[Bonnel, 2005a] Bonnel N., Cotarmanac'h A. et Morin A., Gestion et visualisation des résultats d'une requête. *Actes de l'atelier Visualisation et Extraction de Connaissances - EGC'05*, pages 37–47.

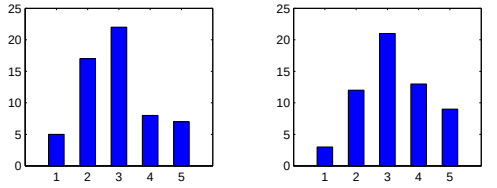
[Bonnel, 2005b] Bonnel N., Cotarmanac'h A. et Morin A., Meaning Metaphor for Visualizing Search Results. *Proceedings of Int. Conf. on Information Visualisation*, pages 467–472. IEEE Computer Society.

[Boyack, 2002] Boyack K. W., Wylie B. N. et Davidson G. S., Domain Visualization Using VxInsight for Science and Technology Management. *JASIST*, 53(9) :764–774.

[Chevalier, 2005] Chevalier M. et Hubert G., Evaluation d'une interface de restitution de recherche : Quelles conclusions en tirer ? *Actes de l'atelier Visualisation et Extraction de Connaissances - EGC'05*, pages 15–27.

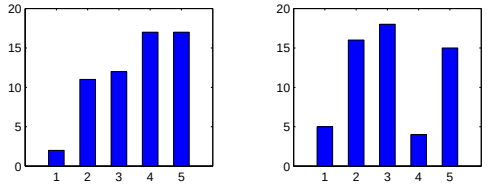
[CLEF] CLEF, Cross-Language Evaluation Forum. http ://clef.isti.cnr.it.

[Google] Google, http ://www.google.com.



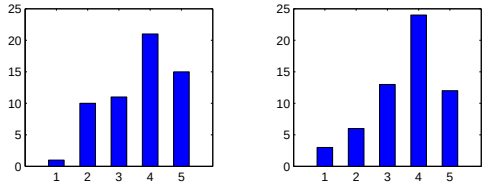
(a) Clarté d'utilisation du prototype

(b) Facilité d'utilisation du prototype



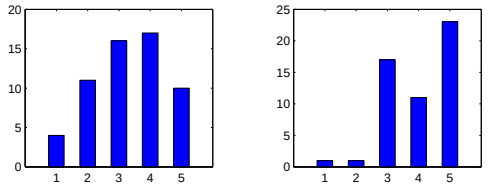
(c) Apprendre à utiliser le prototype

(d) Destiné à tous les niveaux d'utilisateurs



(e) Compréhension de la métaphore ville

(f) Pertinence de la métaphore ville



(g) Navigation dans la métaphore ville

(h) Utilité d'une visualisation regroupant certaines pages Web

FIG. 3 – Représentation graphique de l'extrait du test utilisateur proposé dans le tableau 1.

[Grokker] Grokker, A New Way to Look at Search. http ://www.grokker.com.

[Hearst, 1997] Hearst M. et Karadi C., Cat-a-Cone : An Interactive Interface for Specifying Searches and Viewing Retrieval Results using a Large Category Hierarchy. *Proceedings of Int. ACM/SIGIR Conference*, pages 246–255.

[INEX] INEX, Initiative for the Evaluation of XML Retrieval. http ://inex.is.informatik.uni-duisburg.de/2005/.

[iProspect, 2004] iProspect, iProspect's Search Engine User Attitudes Survey Results. White paper.

[Kartoo] Kartoo, meta search engine. http ://www.kartoo.com.

[Kohonen, 1995] Kohonen T., *Self-Organizing Maps*. Springer.

[Lyman, 2003] Lyman P. et Varian H., "How Much Information". Retrieved from http ://www.sims.berkeley.edu/how-much-info-2003.

[Map.net] Map.net, développé par Antartica. http ://map.net.

[MapStan Search] MapStan Search, meta search engine. http ://search.mapstan.net.

[Moreau, 2005] Moreau F. et Sébillot P., Contributions des techniques du traitement automatique des langues à la recherche d'information. Rapport Technique No 1690, IRISA.

[Salton, 1983a] Salton G., Fox E. et Wu H., Extended Boolean Information Retrieval. *Communications of the ACM*, 26(11) :1022–1036.

[Salton, 1983b] Salton G. et McGill M., *Introduction to Modern Information Retrieval*. Mac Graw Hill, New York, États-Unis.

[Shneiderman, 1998] Shneiderman B., *Designing the User Interface*. Addison-Wesley.

[Skupin, 2003] Skupin A. et Fabrikant S., Spatialization Methods : A Cartographic Research Agenda for Non-Geographic Information Visualization. *Cartography and Geographic Information Science*, 30(2) :99–119.

[Sparacino, 2002] Sparacino F., Wren C., Azarbayejani A. et Pentland A., Browsing 3-D spaces with 3-D vision : body-driven navigation through the Internet city. *Proceedings of Int. Symp. of 3DPVT*, pages 224–233.

[Strzalkowski, 2000] Strzalkowski T. et Perez-Carballo J., Natural Language Information Retrieval : Progress Report. *Information Processing & Management*, 36(1) :155–178.

[Sullivan, 2003] Sullivan D., Searches Per Day. Search Engine Watch, February 2003, http ://searchenginewatch.com/reports/article.php/2156461.

[TREC] TREC, Text REtrieval Conference. http ://trec.nist.gov.

[Vivísimo] Vivísimo, clustering engine. http ://www.vivisimo.com.

[Websom project] Websom project, http ://websom.hut.fi/websom/.