Un nouveau passage à l'échelle en recherche d'information

Mohand Boughanem, Lynda Tamine-Lechani, José Martinez, S. Calabretto, Jena Pierre Chevallet

Un nouveau passage à l'échelle en recherche d'information

Mohand Boughanem[*] — Lynda Tamine-Lechani[*]
José Martinez[**] — Sylvie Calabretto[***] — Jean-Pierre Chevallet[****]

[*] *Équipe SIG, IRIT (UMR 5505), 118 route de Narbonne, F-31000 Toulouse cedex 9*

[**] *Équipe Atlas, INRIA & LINA (FRE CNRS 2729), École polytechnique de l'université de Nantes, La Chantrerie, Rue Christian Pauc, BP 50609, F-44306 Nantes cedex 03*

[***] *Áxe ADC, LIRIS (UMR 5205), Bâtiment Blaise Pascal, 7 avenue Jean Capelle, F-69621 Villeurbanne cedex*

[****] *Équipe MRIM, IMAG, Bâtiment B CLIPS-IMAG, BP 53, F-38041 Grenoble cedex 9*

{mohand.boughanem, tamine}@irit.fr
jose.martinez@univ-nantes.fr sylvie.calabretto@liris.cnrs.fr
jean-pierre.chevallet@imag.fr

RÉSUMÉ. *La quantité d'information numérique produite et consultée a considérablement augmenté et sa diversité s'est accrue. Or, le facteur d'échelle joue un rôle important dans la quantité et la qualité des traitements que l'on peut appliquer aux informations, aussi bien de manière intrinsèque que perçue par l'utilisateur. Cet article propose un panorama des problèmes qui découlent de ces évolutions ainsi que quelques pistes de recherche afin de répondre à ce qui semble bien être un* nouveau *défi de « passage à l'échelle ».*

ABSTRACT. *the quantity of numerised information that is being produced and consulted daily has considerably increased over the last few years; it is much more heteregeneous too. The quality and the quantity of processing that can be applied to sources of information is inherently dependent on their sizes as well as the expectations of the users. This paper surveys some of the main problems that are directly related to the increase in size and variety of the collections of data. It also indicates some research issues in order to answer a seemingly* new *"scale up" problem in information management.*

MOTS-CLÉS : *recherche d'information, Toile, passage à l'échelle, efficience, efficacité.*

KEYWORDS: *information retrieval, Worl Wide Web, multimedia, scale-up, efficiency, effectiveness.*

RSTI - ISI - 11/2006. Passage à l'échelle, pages 9 à 35

1. Introduction

Les systèmes d'information informatisés couvrent progressivement toute la production d'information, soit directement lors de la production (ex. : dessins animés en images de synthèse) soit après numérisation. Dans le même sens, l'explosion de l'utilisation de la Toile est à l'origine d'une croissance très significative des applications destinées à aider les utilisateurs à produire de l'information et à y accéder aisément. Ainsi que la Toile l'a démontré, parcourir des collections de documents par navigation est très populaire auprès des utilisateurs, notamment les néophytes. En intégrant, de surcroît, les effets de la mise en œuvre d'intranets et d'extranets d'entreprises ainsi que les bibliothèques numériques, on prend conscience du volume exponentiel d'information et du nombre grandissant d'utilisateurs auxquels est confronté tout service d'information. Rapportons plus précisément trois séquences de chiffres :

1) Le nombre d'internautes est passé de 61 millions en 1996 à 147 millions en 1998 pour dépasser le milliard en décembre 2005 (Source : SVM).

2) Le nombre de serveurs web double chaque année depuis 1993.

3) On estime également que la masse d'information sur la Toile est passée de quelques milliers de textes au tout début de la Toile, en 1993, à plus de trois milliards de pages en 2002 et 8 milliards de pages indexées en 2006 par Google ^{TM}. Certains observateurs s'accordent à dire que ces volumes doublent tous les vingt mois.

Ce changement d'échelle tant dans le volume d'information que dans le nombre d'utilisateurs amène de nouveaux problèmes et ouvre de nouvelles perspectives.

D'un côté, on peut affirmer que la communauté scientifique en recherche d'information a apporté des solutions avérées pour améliorer sans cesse l'ensemble des fonctions d'un système de recherche d'information. On cite plus particulièrement les résultats probants obtenus dans le domaine de la modélisation de l'information, de l'évaluation de requêtes et de la visualisation des résultats des requêtes. D'un autre côté, force est de constater que les solutions traditionnelles connues ne peuvent conserver leur degré d'efficacité, voire leur faisabilité, dans le contexte actuel où la problématique de la recherche d'information a pris une nouvelle dimension. Cette inadéquation mérite des investigations qui permettraient, d'une part, de transposer en partie des acquis considérables en la matière (modèles et mécanismes théoriques développés en recherche d'information) au contexte actuel et, d'autre part, de dresser une étude prospective qui éclairerait les voies de recherche à investir pour faire face à ce nouveau « passage à l'échelle ».

Dans ce cadre, nous visons deux objectifs principaux :

– identifier les verrous technologiques liés à la croissance et à la diversification des contenus des systèmes d'information informatisés et aux processus de recherche d'information associés ;

– dresser un bilan prospectif permettant de s'affranchir, en partie, des verrous posés par la problématique du passage à l'échelle.

Afin de répondre à cette double problématique, la section 2 présente les facteurs à l'origine de la nécessité du passage à l'échelle, une caractérisation sous l'angle des dimensions espace/temps ainsi que les verrous technologiques et scientifiques engendrés. La section 3 présente le problème du passage à l'échelle à travers sa projection sur le processus classique de recherche d'information. Plus précisément, nous abordons les contraintes imposées par le passage à l'échelle sur les phases de préparation des collections, d'évaluation des requêtes et de visualisation des résultats d'une recherche. La section 4 analyse la viabilité des protocoles d'évaluation de la recherche d'information dans un contexte de grande échelle. Enfin, la section 5 dresse un bilan des réflexions menées dans ce cadre. Ce bilan se décline essentiellement dans la définition de directions de recherche permettant de s'affranchir en partie des verrous liés au passage à l'échelle.

2. Problématique du passage à l'échelle

Le traitement de grands volumes d'informations est souvent désigné par l'expression « passage à l'échelle ». Plus précisément, le passage à l'échelle d'une technique ou d'un algorithme désigne sa capacité à traiter des volumes considérables d'informations tout en conservant une complexité du même ordre de grandeur réelle, c'est-à-dire chronométrée, que celle induite par le traitement des volumes antérieurs moins importants. Dans le domaine de la recherche d'information, la dimension du problème et le contexte d'utilisation ont changé dans des proportions considérables depuis quelques années. En effet, l'espace de stockage augmente continuellement puisque :

– la production de données, sous forme électronique, continue de croître ;

– les documents électroniques contiennent de plus en plus souvent des informations multimédias (images et graphiques sont courants, mais audio et vidéo tendent à se généraliser) ;

– des méta-informations sont générées et associées aux données de base afin de faciliter les accès ultérieurs ;

– les utilisateurs accèdent à des sources de plus en plus vastes et disséminées, le cas extrême étant la Toile.

Cette croissance du volume se constate également à plusieurs niveaux de granularité :

– nombre de collections ;

– nombre de documents au sein d'une collection ;

– nombre des granules au sein des documents (semi) structurés ;

– nombre de descriptions ou métadonnées associées aux granules ;

– nombre d'index associés aux données et métadonnées.

Ainsi, le volume d'informations ne se mesure plus en giga-octets (10^9) mais en téra-octets (10^{12}), voire en péta-octets (10^{15}). La recherche d'information étant le processus permettant de renvoyer à des utilisateurs le sous-ensemble des documents pertinents contenus dans une collection, il est ainsi clair qu'aujourd'hui apparaissent de nouveaux enjeux. Cette section présente les principaux facteurs qui nécessitent un passage à l'échelle puis tente de les projeter sur les dimensions d'espace et de temps.

2.1. *Facteurs de l'échelle*

L'analyse de tout phénomène nécessite une étude préliminaire de ses origines, d'où notre souci d'identifier les facteurs de l'échelle. L'analyse du contexte documentaire met en avant deux sources privilégiées :

– les *bibliothèques numériques* constituent des serveurs de collections de documents « fermées » accessibles par un public relativement averti ;

– la *Toile* est une source d'information accessible à un très large public et tendant à devenir la première référence.

La recherche d'information sur la Toile se singularise vis-à-vis d'une recherche d'information classique par un degré plus accentué de nombreux paramètres : taille, hétérogénéité et dynamique des collections, variété des langues utilisées, interconnexion des documents, nombre et variété des profils des utilisateurs (Huang, 2000). Il est ainsi clair que le passage à l'échelle se pose avec davantage d'acuité dans le cas de la recherche sur la Toile, si bien que des techniques proposées pour traiter de très larges collections de documents ont déjà été proposées dans ce contexte (Hawking *et al.*, 1999).

Cela dit, abstraction faite du type de source d'information, on peut dégager les deux principaux facteurs suivants du passage à l'échelle :

– croissance de la taille des collections ;

– croissance du nombre des utilisateurs.

2.1.1. *Croissance de la taille des collections*

Nous identifions la croissance de la taille des collections comme le facteur dominant. En effet, nous sommes encore actuellement dans une phase de croissance *exponentielle* de la taille des collections interrogées en termes de nombre de documents accessibles. La taille de certains documents, ou plutôt multidocuments, comme les encyclopédies en ligne, nécessitent de faire apparaître de très nombreux granules d'informations (phrases, paragraphes, etc., mais aussi concepts, thèmes, etc.) qui ne se ramènent pas nécessairement à la page. Enfin, le nombre et le volume des lexiques, *thesaurus* ou ontologies manipulés augmentent également.

Au passage, notons les incidences sur les coûts en espaces de stockage occupés par de tels volumes d'information, qui demeurent d'actualité malgré la baisse des coûts

des technologies de stockage, ainsi que les coûts de fonctionnement (ex. : consommation électrique).

Une incidence immédiate porte sur le temps d'évaluation des requêtes des utilisateurs puisqu'elle dépend en grande partie de la qualité et de la taille des index mis en œuvre. Dans ce contexte, on rapporte qu'un compromis généralement adopté entre la taille des collections indexées et la qualité des réponses en termes de délais de réponse a conduit de nombreux concepteurs de moteurs de recherche sur la Toile à limiter l'espace d'indexation à quelques pour cents seulement de la totalité de la Toile, de l'ordre de 20 %.

2.1.2. *Croissance du nombre des utilisateurs*

Ensuite, on estime que plus de deux millions de requêtes sont soumises quotidiennement sur la Toile à des moteurs de recherche. Le nombre d'utilisateurs de ce service est passé de 25 millions en 1997 à plus de 200 millions en 2000 avec des proportions disparates en fonction des secteurs géographiques (Kobayashi *et al.*, 2000). Une étude réalisée sur 211 063 utilisateurs (Spink *et al.*, 2001) révèle une tendance à l'interrogation à l'aide de requêtes courtes ; plus précisément plus de 48,4 % de requêtes se limitent à un unique terme, 20,8 % ont deux termes et 31 % ont trois termes ou plus. Cette étude rapporte en outre que les utilisateurs consultent rarement les résultats de recherche figurant au-delà de la deuxième page. Le facteur de croissance du nombre d'utilisateurs pose alors un double impératif :

– sur les délais des réponses, issues d'interrogations *simultanées* ;

– sur la qualité de la fonction d'évaluation des scores de pertinence des documents puisque seuls les quelques premiers documents sont consultés par les utilisateurs alors que des centaines et milliers, voire des millions, de documents peuvent s'apparier avec des requêtes *peu expressives*.

2.2. *Passage à l'échelle : une réalité*

Dans la section précédente, deux des trois facteurs identifiés concernent des problèmes de performances, liés aux volumes et aux temps de traitement respectivement. Le passage à l'échelle est une réalité qui peut être effectivement appréhendée de ces façons-là. Elles ne s'excluent pas ; bien au contraire, la première implique la seconde.

On peut considérer, naïvement, que le problème du passage à l'échelle est seulement un problème de volume. Or, ce problème entraîne un problème de temps de calcul : temps pour indexer l'information et temps pour retrouver l'information. Formellement, il existe une relation fonctionnelle entre le temps et l'espace :

$$t = f(e)$$

Les temps de traitement dépendent directement du volume des données. Il est évident qu'un algorithme lent ne peut pas être appliqué sur une collection aussi vaste que celle affectée à un algorithme rapide. Un algorithme rapide rencontrera lui-même une frontière. Cette dernière pourra être repoussée si l'algorithme est parallélisable (et les données réparties). Elle pourra même être repoussée indéfiniment, du moins en théorie, si l'algorithme est extensible. Mais n'oublions pas qu'il doit aussi être incrémental, de nouvelles données venant s'ajouter régulièrement à la collection d'information. Enfin, il nous faut lever la supposition toute simple et implicite de l'existence même d'un algorithme ! Certaines applications nécessitent une indexation en grande partie manuelle.

Le passage à l'échelle se révèle donc être tout autant un problème de volume que de temps de calcul, le premier influant directement et fortement sur le second mais n'étant pas le seul facteur de complexité.

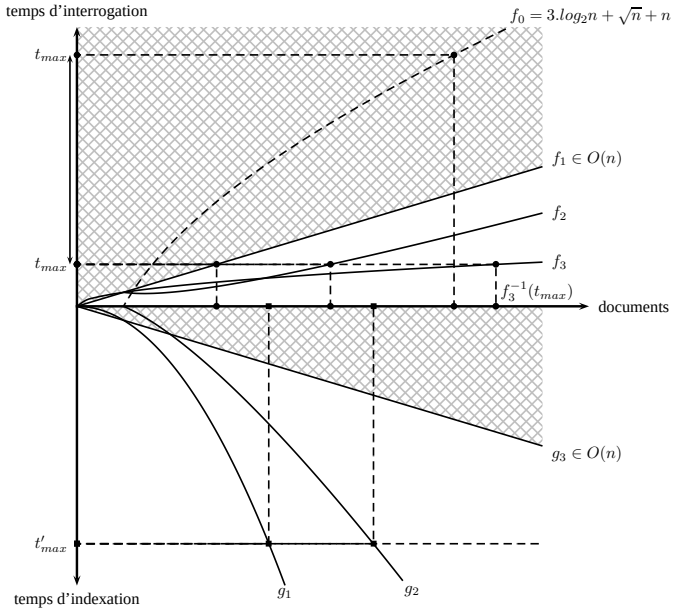

Figure 1 – *Compromis temps/espace et temps d'indexation / temps d'interrogation*

Cela est traduit sur la figure 1 où l'on distingue :

– un axe horizontal correspondant à l'espace, c'est-à-dire à la taille des données (collections et objets) ;

– un axe vertical qui porte :

- dans sa partie supérieure le temps d'interrogation ;
- dans sa partie inférieure le temps d'indexation.

La fonction f qui lie le temps et l'espace n'est rien d'autre qu'une courbe de complexité asymptotique. Sur l'axe du temps, il est possible de placer des contraintes $t_{\max}$ plus ou moins fortes :

– pour des réponses en ligne, la durée des traitements peut varier de moins d'une seconde pour un système local à quelques secondes sur un système largement réparti et chargé comme la Toile ;

– pour des systèmes de recherche très précis, la durée de traitement peut être de plusieurs heures, notamment si le traitement est lancé en fin de journée pour obtenir les résultats le matin suivant.

Pour l'interrogation, la complexité *maximale* envisageable est linéaire, $f_1 = O(n)$, tandis que pour l'indexation c'est la complexité *minimale*, g_3, qui est linéaire. Les parties hachurées ne sont donc pas accessibles. Si cela est évident pour l'indexation, en revanche, on peut envisager des traitements *légèrement* plus coûteux pour l'interrogation. Mais il devient évident que le traitement plus coûteux ne peut s'appliquer qu'à un *sous-ensemble* de la base, présélectionné par un traitement plus rapide [1].

Fixons les idées en nous appuyant sur un exemple. La fonction f_0 pourrait être égale à $3.\log_2 n + \sqrt{n} + \sqrt{n}^2$ où :

– le premier terme serait le temps de sélection des trois listes inverses associées à trois termes d'une requête ;

– le second serait un nombre raisonnable de documents suffisamment pertinents retrouvés par intersection des débuts des listes [2] ;

– le dernier serait un traitement coûteux, quadratique en l'occurrence, sur le sous-ensemble des documents retenus comme *a priori* pertinents.

Soulignons que $f_0 \in O(n)$. Sa complexité asymptotique montre qu'un tel algorithme n'est pas envisageable sur une très grande collection. Par exemple, fin 2005 Google ™ affirme indexer huit milliards d'URL (*uniform resource locator*) ; on pourrait donc être amenés à appliquer un traitement quadratique sur quelques 90 000 pages [3]. Comme d'habitude, un traitement coûteux sur un volume de données très important ne doit

1. On pourra remarquer que dans les systèmes de gestion de bases de données relationnels, certains traitements sont d'une complexité plus que linéaire sur l'ensemble des données. Plus précisément, les opérations de jointure et de tris (ainsi que les opérations dérivées : agrégations, opérations ensemblistes) ont une complexité en $O(n.\log_2 n)$. Dans la pratique, l'unité de mesure est le nombre de blocs manipulés sur le disque et non de n-uplets. Dans ce cas, le facteur logarithmique est ramené à une constante, dans l'intervalle $[2, 5]$, qui correspond à la hauteur des arbres-B qui indexent les relations.

2. Le pourcentage du nombre de documents retenus par rapport au nombre de documents indexés se réduit rapidement avec l'augmentation de ces derniers : $\left(\frac{\sqrt{10}}{10}\right)^k$ pour des puissances de dix, soit une suite géométrique décroissante.

3. Le 26 avril 2006, les requêtes « Java », « langage Java » et « langage objet Java » sélectionnent respectivement (et très approximativement) 1 milliard, 5 millions et 1 million de pages. Les sujets « danse Java » et « île Java » renvoient respectivement à plus de 1 et 6 millions de pages.

pas dépasser une complexité en $O(m.\log m)$ (Skiena, 1998) ; dans le cas de f_0, cela aboutit à une complexité en $O(\sqrt{n}.\log_2 n)$ et, dans le cas de Google, $\log_2 8.10^9 \simeq 33$ seulement. Inversement, si l'on ne dispose pas d'un tel algorithme, alors l'étape de présélection doit être plus restrictive, ne retenant que $\sqrt[4]{n}$ documents pour un post-traitement quadratique, par exemple, voire un nombre constant de documents fixés *a priori*.

Cet exemple montre combien les performances doivent être traitées au cas par cas, en fonction des structures d'index et des stratégies de recherche. Sur la figure 1, le rectangle est délimité par l'origine, la contrainte de temps $t_{\max}$ et $f^{-1}(t_{\max})$ est la zone d'applicabilité d'un système de recherche d'information.

L'amélioration progressive des performances de l'algorithme de recherche, $f_3 < f_2 < f_1$, permet d'interroger des bases de plus en plus grandes. A la limite, un algorithme en $O(1)$ permet d'accéder à une collection de taille infinie. C'est actuellement le cas ! En effet, nous avons vu que les utilisateurs de moteurs de recherche interrogent souvent la Toile en utilisant un unique terme. Le nombre de termes est une constante, très grande, peut-être de l'ordre de 10^7, mais un fichier inverse permet de retrouver la liste – ordonnée – des documents contenant le terme indiqué en un maximum de $\lceil \log_2 10^7 \rceil = 24$ comparaisons.

Il faut distinguer le rectangle correspondant au temps d'interrogation de celui correspondant au temps d'indexation. Le premier est généralement bien plus contraint que le second, $t_{\max} \ll t'_{\max}$, les traitements d'indexation se déroulant hors-ligne.

Ils entretiennent cependant un rapport. Le temps d'indexation est unitaire alors que le temps d'interrogation s'applique à chaque requête sur la collection indexée. Ainsi, leur rapport, c'est-à-dire le *nombre de requêtes*, est un élément important.

Le problème du « passage à l'échelle » est que l'on ne contrôle pas le système par rapport au temps maximal autorisé mais seulement par la taille des données et que ces dernières augmentent. Il faut donc trouver des solutions pour maintenir le temps d'indexation dans des limites *raisonnables* et maintenir les temps de réponse dans des limites *inextensibles*. Du seul point de vue algorithmique, le « passage à l'échelle » s'avère déjà une problématique de recherche importante.

2.3. *Verrous technologiques et scientifiques*

Au de-là des problèmes algorithmiques, le passage à l'échelle se heurte aussi à la « malédiction de la dimensionnalité » (Berrani *et al.*, 2002). Ce terme traduit l'extrême difficulté, et à la limite l'impossibilité, de gérer efficacement et simultanément un très grand nombre de paramètres aussi bien en interrogation qu'en indexation. Si l'interrogation textuelle a pu éviter ce problème dans toute son ampleur, notamment en limitant arbitrairement le nombre maximum de termes intervenant dans une requête, ce n'est pas le cas de la recherche de données multimédias par le contenu où,

de par le manque de sémantique associée aux métadonnées, il est nécessaire de fournir systématiquement un très grand nombre de valeurs de descripteurs.

Dans les deux cas, cela entraîne à aborder peu ou prou non seulement les questions d'efficience et mais aussi d'efficacité d'un processus de recherche d'information (Frieder *et al.*, 2000; Newby, 2000).

2.3.1. *Efficience*

> « Efficience : capacité de rendement, performance » (Larousse)

Sous l'angle de l'*efficience* (c'est-à-dire de la rapidité), des travaux montrent que le temps d'indexation moyen de la collection et le temps de traitement moyen des requêtes augmentent de manière très significative en fonction de la taille des collections (Voorhees *et al.*, 1999; Hawking *et al.*, 1999). Cela est essentiellement dû, d'une part, à l'accroissement des index et donc de l'espace de stockage face à une évolution quasi-linéaire des ressources matérielles et, d'autre part, à une complexité trop importante des algorithmes d'indexation et d'évaluation de requêtes. Ce verrou a été abordé en proposant des solutions qui portent essentiellement sur la compression physique des informations, améliorant donc le premier point et par contre-coup le second (Moffat *et al.*, 1996; Scholer *et al.*, 2002; Heinz *et al.*, 2003). Cependant, le problème de l'échelle n'étant pas abordé dans sa globalité et dans sa diversité, l'impact sur l'efficience de la recherche reste peu significatif ; l'amélioration des performances est réelle mais ne change pas la complexité asymptotique.

2.3.2. *Efficacité*

> « Efficace : qui produit l'effet attendu [. . .] dont l'action aboutit à des résultats utiles » (Larousse)

Sous l'angle de l'*efficacité* (c'est-à-dire de la qualité des résultats), force est de constater que les travaux en recherche d'information ont peu considéré le paramètre du volume. La mesure classique d'évaluation des performances moyennant les taux de rappel et précision est en effet compromise dans un contexte où la procédure de sélection et de tri des documents pertinents agit sur un volume considérable d'informations. Cela signifie que l'ancien objectif « idéal » d'équilibre entre rappel et précision devrait s'orienter davantage vers de la *haute précision* de manière à cibler plus précisément les documents plus pertinents parmi les milliers et plus de documents candidats. Par ailleurs, les protocoles d'évaluation de la pertinence des résultats de recherche demeurent à ce jour indépendants de la taille et de la diversité des *corpus* de test. Cela engendre des biais d'évaluation qui peuvent être relativement conséquents lors de la comparaison des performances entre différents systèmes (Zobel, 1998). Ce verrou constitue l'enjeu d'une tâche récente, *TeraByte*, introduite à juste titre dans la campagne d'évaluation TREC (*text retrieval conference*, `http://trec.nist.gov`).

La prise en compte du volume d'information lors de l'évaluation des systèmes en termes d'efficience et d'efficacité a été la principale motivation de ce travail. Dans

cette perspective, l'objectif est tout d'abord de cerner les différentes facettes de cette problématique puis de dresser des pistes de réflexion qui abordent les verrous évoqués précédemment.

3. Le processus de recherche d'information et le passage à l'échelle

Le processus de base, dit en « U », de la recherche d'information est établi de longue date. Après la présentation de ses principales phases, nous tentons de l'observer à présent sous l'angle du passage à l'échelle afin d'en mesurer les conséquences.

3.1. *Le processus de base en recherche d'information*

Un processus de recherche d'information a pour objectif de répondre à un besoin en information exprimé par les utilisateurs *via* des requêtes, en présentant une liste de documents sélectionnés à partir d'une collection donnée appelée *corpus* du système. Cette liste est généralement triée selon un score de pertinence calculé sur la base d'un modèle précis.

La figure 2 rappelle le processus de base de la recherche d'information. L'étape de préparation des collections au sens large, qui a pour objectif d'aboutir à la fois à :

– une représentation des documents eux-mêmes, répartie entre :

- informations *locales* à chaque document ;
- informations *globales* à la collection ;

– une représentation physique incluant, entre autres :

- le modèle de données (relationnel, propriétaire…) ;
- des index à proprement parler (fichiers inverses, fichiers de signature…) ;
- des techniques de compression des données ;

– une éventuelle classification des documents (*clusters* de documents similaires…) ;

Lors de l'interrogation, l'utilisateur fournit une description dans un langage simplifié. Le système se charge de traduire cette formulation en une formulation interne qualifiée de requête.

La requête est alors mise en correspondance avec l'index de la base documentaire. Du point de vue logique, le but de cette étape, qualifiée d'évaluation de requête, est de *sélectionner* les documents pertinents relativement au besoin en information de l'utilisateur. Mais cette étape est cruciale du point de vue des performances du système ; elle doit *élaguer* aussi rapidement que possible le domaine de la recherche.

Les (premiers) résultats sont alors ordonnés et présentés à l'utilisateur. Contrairement à une requête dans un système de gestion de bases de données, cette liste de

résultats n'est pas nécessairement « la » réponse à la requête. Les imprécisions de la requête, tout autant que la difficulté à indexer les documents, entraînent la présence de bruit. L'utilisateur peut alors procéder à une évaluation de la pertinence, c'est-à-dire indiquer au système quels documents sont pour lui plus particulièrement pertinents, ou non pertinents, parmi ceux présentés. Le système procède à une rétroaction qui consiste à réécrire la requête en tenant compte de ces indications. La rétroaction peut également passer par l'utilisateur, notamment lorsqu'il doit lui-même réécrire sa requête ou par une approche alternative basée sur un système de recherche par navigation.

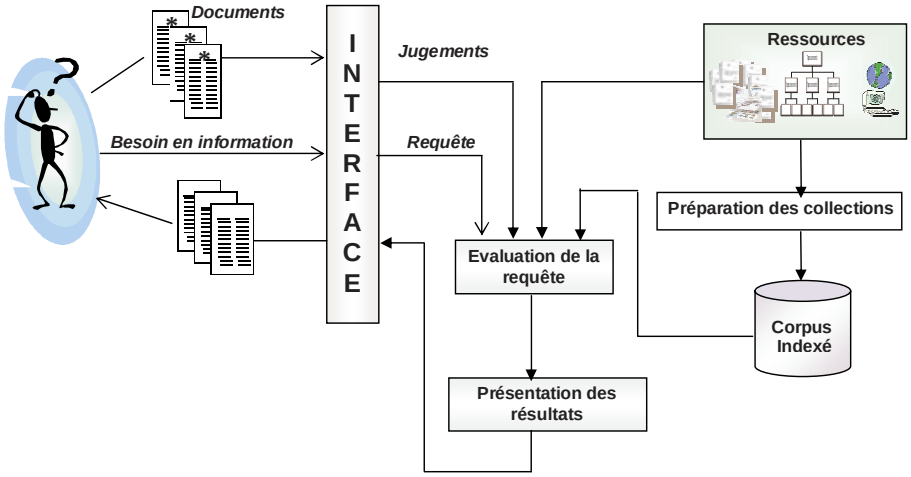

Figure 2 – *Processus de base de la recherche d'information*

Les sections suivantes tentent d'identifier les conséquences du passage à l'échelle sur les différentes phases du processus de base de la recherche d'information : préparation des collections, évaluation de requêtes, présentation des résultats, et d'en dégager *in fine* les principaux enjeux. Ces phases constituent les dimensions selon lesquelles nous décrivons la problématique, des solutions (partielles) existantes ainsi que les perspectives envisagées.

3.2. *Préparation des collections*

La préparation de collections volumineuses pose des problèmes liés fondamentalement à la grande dimension de l'espace de représentation. nous présentons dans cette section, un aperçu global des problèmes posés par la préparation de grands volumes d'informations puis synthétisons les solutions apportées dans les travaux de recherche du domaine.

3.2.1. *Problématique*

Un modèle de représentation de l'information peut être perçu sous deux angles. Le point de vue conceptuel induit la définition d'un support théorique comme base de représentation des unités d'information et de formalisation de la fonction de pertinence du système. De très nombreux modèles ont été proposés dans la littérature : booléen, vectoriel, probabiliste, etc. Sous cet angle, le représentation de grands volumes d'informations pose des problèmes liés fondamentalement à la grande dimension de l'espace de représentation. Nous avons déjà évoqué la nature intrinsèquement multidimenssionnelle des métadonnées multimédias. Pour les données textuelles, il a été constaté que l'accroissement du volume engendre des problèmes similaires :

– une diminution de la pertinence de la représentativité locale et globale des documents ;

– un accroissement des index conceptuels (vocabulaire des *corpus*) qui accentue les difficultés liées à la synonymie et à la polysémie (Berry *et al.*, 1999) ;

– une difficulté à maîtriser l'évolution des vocabulaires en fonction de la taille des collections (Williams *et al.*, 2005) ;

– une multiplication des zones textuelles accessibles telles que titre, résumé, paragraphe et par conséquent une gestion plus complexe des index (Kowalski, 1997).

Du point de vue physique, les problèmes qui en découlent sont principalement :

– l'augmentation considérable de l'espace de stockage des fichiers inverses ;

– le débordement du cache de l'index de la mémoire vive vers le disque ce qui nécessite un nombre supplémentaire d'accès disque pour l'évaluation des requêtes et par conséquent une augmentation des délais de réponse ;

– la difficulté des mises à jour fréquentes des index, ces derniers étant généralement compressés.

3.2.2. *Etat de l'art*

La préparation de collections volumineuses pose des problèmes liés fondamentalement à la grande dimension de l'espace de représentation. Plus précisément, le passage à l'échelle engendre, d'une part, un accroissement du coût de stockage et du maintien des index et, d'autre part, une diminution de la pertinence des représentations locales et globales. Cela a suscité des travaux de recherche qui s'inscrivent dans deux directions principales :

– la représentation conceptuelle de l'information ;

– la compression physique des documents.

3.2.2.1. Représentation conceptuelle de l'information

La représentation conceptuelle de l'information traduit la réduction de l'espace de représentation d'une collection en s'orientant davantage vers la représentation des documents à l'aide de concepts agrégés que sur des unités d'information plus fines

telles que les mots ou n-grammes. En effet, dans le cas de collections croissantes, la matrice de base termes/documents est de dimension très élevée et de surcroît creuse. Une étude réalisée sur la distribution des termes dans un échantillon de collections volumineuses révèle que ces matrices ont un taux de densité moyen évalué à seulement 1 % (Berry, 1992). L'un des premiers modèles proposés dans une perspective de réduction de l'espace de représentation est le modèle LSI (*latent semantic indexing*) (Dumais, 1993). La base mathématique du modèle LSI est la décomposition en valeurs singulières de la matrice termes / documents. Cette décomposition identifie une base vectorielle de dimension plus réduite qui couvre le même espace vectoriel associé que celui de la représentation initiale des documents, obtenue par application d'opérations algébriques (rotation, factorisation) sur les vecteurs d'origine. On trouve des extensions au modèle de base (Kokiapoulou *et al.*, 2004; Tang *et al.*, 2004), notamment des techniques appropriées à la gestion de matrices creuses non-régulières, traduisant une disparité dans l'occurrence des termes d'indexation dans les collections ; elles sont essentiellement basées sur la révision du processus de décomposition (Lehoucq *et al.*, 1996) et de minimisation de la trace (Sameh *et al.*, 1982). D'autres méthodes pour la réduction de l'espace de représentation basées sur la factorisation F de la matrice termes / documents ont été développées (Berry *et al.*, 1999).

En ce qui concerne les données multimédias, la « malédiction de la dimensionnalité » est particulièrement présente. Elle est liée, d'une part, au très grand nombre de descripteurs de bas niveau que l'on peut extraire des images, bandes audio et vidéos et, d'autre part, aux dimensions importantes de la plupart de ces descripteurs (histogrammes, transformées de Fourier, attributs cepstraux, etc.). Leur taille pose des problèmes difficiles de réduction de la dimensionnalité : conservation des métriques, perte d'information et apparition d'artefacts (ce qui peut entraîner un mauvais regroupement des objets sur de pseudocaractéristiques communes importantes). Ces problèmes ont été largement explorés (ACP – analyse en composantes principales –, SOM – *self-organising maps* –, SVM – *support vector machine* –, etc.) mais sont réactualisés par la dimension croissante des objets mesurés. Ces problèmes de coût de calcul et de précision (densité des données, espace creux) sont partagés par les techniques de classification et d'indexation. Les données multimédias posent également des problèmes difficiles de conceptualisation de l'information de bas niveau en des termes intelligibles par un être humain – autrement que par la visualisation et l'interactivité. Par exemple, si les notions liées à la couleur sont assez facilement transposables dans le langage naturel (« rouge », « bleu », « clair », « vif », etc.), en revanche les nombreux descripteurs de texture, visuelle ou sonore, ne sont sont pas aisés à conceptualiser. Enfin, il est utile de définir précisément ce que l'on entend par recherche d'information multimédia ; les descripteurs et les métriques varient beaucoup en fonction des applications visées (notamment dans le domaine de la reconnaissance des formes).

3.2.2.2. Compression physique de l'information

La compression physique de l'information comprend l'ensemble des techniques permettant d'améliorer l'efficience en recherche d'information en se basant sur la réduction de l'espace de stockage des structures physiques représentatives des do-

cuments, plus particulièrement les fichiers inverses. Les approches standard pour la construction de l'index conviennent davantage pour de petites collections où le vocabulaire et les fichiers inverses résident en mémoire vive, permettant ainsi une construction rapide de l'index (Heinz *et al.*, 2003). Le passage à l'échelle engendre l'accroissement du nombre de termes et donc de l'index, l'allongement des listes inverses et l'augmentation de l'espace de stockage. Partant, l'index n'étant plus en mémoire vive, le temps d'évaluation des requêtes devient bien plus important en raison des accès disque induits. L'une des solutions apportées est la compression des fichiers inverses qui a pour principaux avantages (Scholer *et al.*, 2002) :

– la réduction de l'espace de stockage ;

– une meilleure rentabilité des outils de communication (pour des schémas de compression efficaces, le temps de compression ajouté aux temps de transfert et de décompression est inférieur au temps de transfert des données non compressées) ;

– la réduction du nombre moyen d'accès disque puisque la probabilité d'avoir une partie de l'index en mémoire vive est plus importante.

La plupart des techniques de compression se basent sur la représentation des différences, appelées *gaps*, entre séquences ascendantes d'entiers représentant les adresses de documents où apparaissent chaque terme de l'index. En effet, les différences sont plus petites que les adresses absolues et donc représentables sur de plus petits formats. On distingue principalement trois classes de méthodes (Frieder *et al.*, 2000) :

– le principe le plus simple est celui des méthodes de compression à longueur fixe qui associent à chaque plage de valeurs fixes le nombre de bits nécessaires pour sa représentation ;

– les méthodes de compression à longueurs variables consistent à déterminer la longueur du code comme une fraction de sa valeur. La méthode la plus représentative est le code de Elias (Elias, 1977) qui représente un entier x avec $2\lfloor \log_2 x \rfloor + 1$ bits ;

– les méthodes basées sur la longueur de la liste inverse consistent à retenir un vecteur de référence $V = (v_1, v_2, \ldots, v_n)$ contenant des entiers positifs. Pour coder un entier $x > 1$, il faut trouver un entier k tel que :

$$\sum_{j=1}^{k-1} v_j < x \leq \sum_{j=1}^{k} v_j$$

puis coder :

$$d = x - \sum_{j=1}^{k-1} v_j - 1$$

Le vecteur V peut être modifié pour avoir différents schémas de compression ajustables en fonction des collections. Une solution élégante (Zobel *et al.*, 1992) consiste à faire varier le vecteur V pour chaque liste inverse en posant : $V = (b, 2 \times b, 4 \times b, \ldots, 64 \times b)$ où b est la valeur médiane dans la liste inverse.

En pratique, on montre que des taux de compression appréciables sont atteints en adoptant le code de Golomb pour la représentation des d-*gaps* et des fréquences des termes dans les documents, et le code de Elias pour les positions des termes dans les documents (Heinz *et al.*, 2003). Une méthode de compression basée non plus sur l'ordre des identifiants de documents mais sur leur similarité a été proposée récemment (Shieh *et al.*, 2003). L'algorithme génère un ordre de réaffectation des identifiants de documents en fonction de leurs similitudes puis calcule les d-*gaps* qui sont ainsi plus faibles puisque associés à des séquences de documents de similitudes graduelles. Les résultats obtenus sur de larges collections montrent que la valeur moyenne du *gap* peut être réduite de 30 % et le taux de compression amélioré de 15 %.

3.2.3. *Perspectives*

Les solutions proposées au problème de la représentation de grands volumes d'informations sont, à notre connaissance, parcellaires : peu de travaux couvrent le niveau conceptuel tandis que la majorité des travaux se sont focalisés sur des aspects d'ordre physique (espace de stockage, nombre d'E/S, etc.). Ces travaux ont pu améliorer les performances de systèmes de recherche d'information mais ils sont insuffisants pour faire face à la croissance des collections.

Il serait intéressant d'aller vers des solutions de représentation des documents qui regroupent les différents niveaux d'abstraction depuis le niveau conceptuel jusqu'aux détails physiques en passant par un contexte d'architecture fonctionnelle. L'objectif visé est d'intégrer dans un même modèle les contraintes des différents niveaux afin d'en déduire des règles de conception d'unités d'information. A notre sens, une première réflexion à envisager dans les travaux futurs porterait sur la granularité de l'information. En effet, la granularité d'un document joue certainement un rôle, mais la recherche d'information textuelle manipule maintenant depuis longtemps les documents en texte intégral. Un document n'est donc qu'une très importante séquence (ou seulement ensemble) de termes.

Mais chercher à « granulariser » des documents multimédias (images, audio et vidéo) présente de nombreux problèmes :

1) Le premier d'entre eux est qu'il est difficile, voire impossible, de cerner le granule. Dans le cas idéal, encore lointain, les éléments isolés feront sens pour l'être humain. Pour des images, ce sera l'extraction de chacun des objets du monde réel (voiture, humain, animal, végétal, etc.). Ce problème n'est pas inexistant dans le document textuel mais la liaison entre le signifiant et le signifié est considérablement réduite puisque le langage, écrit ou parlé, a été conçu dans ce but.

2) Le second problème, lié au premier, est celui de l'emboîtement des descriptions. Ce problème est analogue aux clauses dans les données textuelles et plus généralement aux *thesaurus*. Un élément identifié et nommé peut être associé, le plus souvent de manière hiérarchique mais pas seulement, à d'autres éléments. Par exemple, sur une image on pourra apercevoir un être humain, vu de face, et plus précisément son buste, c'est-à-dire tête, poitrine et partie supérieure des bras. Dans une vidéo, cette information dépend du temps, les éléments apparaissant, se modifiant et finissant par

disparaître, brutalement ou progressivement.

3) Le premier problème entraîne aussi celui de l'agrégation de descriptions de bas-niveau. Etant actuellement incapables, dans toutes les situations, de trouver du sens au signal, on tente de le cerner par une combinaison de descripteurs plus ou moins fiables, plus ou moins corrélés.

4) Enfin, les descripteurs proposés et retenus sont bien plus coûteux que les termes d'un document textuel, aussi bien en termes de temps d'extraction, de stockage, que de temps de recherche.

Pour toutes ces raisons, il semble préférable d'examiner des pistes de recherche autour de la classification, le « granule » documentaire étant alors la classe.

3.3. *Evaluation des requêtes*

Après l'indexation d'une collection, dont nous rappelons qu'elle peut s'effectuer en arrière-plan, il faut se soucier de l'évaluation des requêtes, qui obéissent le plus souvent à des contraintes de temps réel et doivent être envisagées en parallélisme massif.

3.3.1. *Problématique*

Le traitement d'une requête passe par l'évaluation d'une fonction de comparaison (ou fonction d'appariement). Cette fonction identifie les documents correspondant à la requête. Pour cela, elle compare la représentation de la requête à celle des documents dans un espace de représentation donné. Les techniques mises en jeu sont donc fortement corrélées aux représentations. La fonction de comparaison est binaire ou multivaluée : soit elle sépare le *corpus* en deux groupes – documents pertinents ou non –, soit elle ordonne le *corpus* – du document le plus pertinent au moins pertinent. Par exemple, dans un système booléen où la requête est exprimée par une combinaison de mots-clés reliés par des opérateurs booléens, chaque document va être également représenté par une combinaison de mot-clés (simple ensemble ou association avec l'ensemble des positions dans un texte, par exemple). Le premier type de fonction récupère les documents qui correspondent exactement à la formule booléenne de la requête, l'autre type de fonction va ordonner les documents par rapport à leur degré de pertinence vis-à-vis de la requête. (Le modèle booléen conduit à séparer le traitement de ces deux fonctions.)

L'accroissement des volumes des collections a des conséquences évidentes sur l'étape d'évaluation des requêtes tant pour l'efficience que pour l'efficacité.

L'impact sur l'efficience est lié à la complexité en temps de la fonction d'appariement. Cette complexité est cruciale pour le bon fonctionnement d'un système de recherche d'information. En effet, le temps de réponse à une requête doit toujours rester très court (quelques secondes au maximum), *quelle que soit* la taille du *corpus*. Or, évaluer une requête sur l'ensemble des documents engendre un traitement qui

pourrait devenir quadratique (c'est-à-dire impossible à réaliser en pratique), et cela indépendamment du modèle d'appariement deux à deux qui peut lui aussi prendre un temps considérable (comme la projection du modèle des graphes conceptuels). De même, l'étape finale de l'appariement, qui est le classement suivant le RSV (*rank score value*) décroissant, peut être problématique. Il est indispensable de limiter très rapidement le sous-ensemble de documents sur lequel va porter un tri. Par exemple, le modèle vectoriel évite tout tri explicite : il exploite une heuristique de classement décroissant des documents vis-à-vis de chaque terme pour en dériver un classement décroissant des documents vis-à-vis de la conjonction des termes, réputé suffisamment fiable. De manière générale, il convient de transférer aussi largement que possible la complexité en temps dans la phase d'indexation de la base.

L'impact sur l'efficacité est lié à la quantité de documents qu'il faut classifier selon leur pertinence vis-à-vis de la requête. Intuitivement, il est plus facile d'identifier les documents pertinents dans une collection comportant quelques centaines de documents plutôt que dans une collection comprenant des milliards de documents, car le risque d'erreur de classification est statistiquement plus faible dans le premier cas. Les résultats obtenus dans les campagnes TREC confirment ce point (Hawking *et al.*, 1999). L'hétérogénéité est plus importante dans les collections volumineuses et les descriptions statistiques ne sont plus aussi discriminantes. Par exemple, la conjecture de Luhn suppose que le pouvoir de résolution d'un terme, c'est-à-dire sa capacité de discrimination entre un document pertinent et un document non pertinent, est relatif à une fréquence documentaire moyenne. Cette conjecture est à revoir dans le cas de *corpus* de très grandes dimensions. De même, il semble que l'hypothèse qui prévaut à l'usage de la fréquence documentaire inverse, ne soit pas valide sur de grandes collections car l'influence du facteur idf (*inverse-document frequency*), correspondant au nombre de documents où un terme apparaît, va en diminuant avec la taille des collections (Beigbeder *et al.*, 2003). On peut également noter l'absence du modèle vectoriel dans les moteurs de recherche sur la Toile, et son omniprésence dans les expérimentations de recherche sur des *corpus* de taille plus faible. Néanmoins, il faut aussi noter que la valeur exacte de la fréquence des termes semble peu influencer les résultats (Franz *et al.*, 2002). L'objectif de *haute* précision que nous avançons, qui revient à privilégier la précision des premiers documents sélectionnés, implique des traitements plus sophistiqués pour mieux discriminer les (parties de) documents. De nouveaux algorithmes d'appariement, et de préappariement dans la phase d'indexation, restent encore à concevoir.

Pour ce qui est des données multimédias, aux problèmes déjà évoqués s'ajoutent deux autres difficultés. Tout d'abord, (i) la diversité des métadonnées extraites, (ii) l'apparition de nouveaux descripteurs ainsi que (iii) les différences de descriptions dues à des sources variées de données multimédias ou à la non-application de certains traitements coûteux sur certaines données qui rendent plus délicat le processus d'appariement et peuvent donc détériorer la qualité des réponses. Deuxièmement, les techniques d'appariements qui ont été utilisées jusqu'à maintenant (en l'absence de collections volumineuses) s'appuient essentiellement sur des calculs de distances qui se traduisent en substance par des parcours complets des collections de métadonnées.

Même en ayant procédé à une classification préalable des métadonnées, des sélections heuristiques semblent d'ores et déjà nécessaires ou, de manière plus ambitieuse, la définition de nouvelles métriques plus économiques en temps de calcul au moment de l'interrogation.

3.3.2. *Etat de l'art*

Les propositions de réduction de l'échelle ont porté essentiellement sur la parallélisation des algorithmes de recherche d'information (Bayley *et al.*, 1996; Jain *et al.*, 2002), l'optimisation des accès aux fichiers inverses (Moffat *et al.*, 1996) et la réduction de la complexité des algorithmes proposés dans les modèles classiques de recherche d'information (Lee *et al.*, 1996).

Le traitement parallèle de grands volumes de données en recherche d'information possède de nombreux avantages comme l'amélioration du temps de réponse et la capacité à effectuer des recherches dans des collections volumineuses. L'index inversé est l'approche la plus populaire pour les moteurs de recherche textuels. Bayley propose une approche basée sur les fichiers inverses et les dictionnaires. Il a testé son approche sur le *corpus* TREC (3 giga-octets). Cependant, la parallélisation des index inversés n'est pas une tâche aisée. Ainsi, le traitement parallèle de Jain ne s'appuie pas sur ces fichiers inverses mais sur l'approche CSR (*compressed sparse row format*).

Les fichiers inverses constituent le cœur de la majorité des algorithmes de recherche d'information. Dans un fichier inverse, un terme est associé à un ensemble de documents qui contiennent le terme. Lorsque le nombre de documents augmente, il en est de même pour la liste des termes uniques et les fichiers inverses. Ainsi, le temps de traitement d'une requête croît approximativement linéairement avec la taille de l'ensemble de documents (et croît approximativement linéairement avec la taille de la collection de données indexées). Dans la plupart des méthodes de recherche d'information, la première étape consiste à générer l'ensemble de tous les documents qui contiennent les termes recherchés. Malheureusement, de nombreuses approches en recherche d'information croissent plutôt de manière exponentielle. Une solution pour réduire l'échelle consiste donc à optimiser les accès à ces fichiers inverses. En effet, la taille des fichiers inverses peut être aussi grande que les textes qu'ils indexent. La compression des fichiers inverses permet de réduire la taille de 80 %.

La réduction de la complexité des algorithmes est principalement basée sur l'approche à base de fichiers de signatures. Il s'agit d'un mécanisme de filtrage qui élimine les blocs de textes qui ne peuvent pas répondrent à une requête.

3.3.3. *Perspectives*

En conclusion, les éléments de solution liés au traitement des requêtes sont de plusieurs ordres :

– déporter autant que possible les traitements coûteux dans la phase d'indexation ;

– remplacer les traitements coûteux qui persisteraient dans la phase d'appariement par des heuristiques ;

– élaguer les traitements en tenant compte d'autres facteurs comme l'utilisateur, l'usage et le niveau d'abstraction des informations ;

– revoir les modèles de RI et notamment établir de nouvelles mesures, les descripteurs statistiques « standard » devenant trop peu discriminants dans des collections volumineuses hétérogènes, incomplètes et multimédias.

3.4. *Présentation des résultats*

3.4.1. *Problématique*

Lorsque l'on envisage l'accès à de grandes masses d'informations, il faut également traiter de la présentation des résultats. On peut alors légitimement se demander si le passage à l'échelle influence les modes de présentations des données. De manière générale, les travaux sur la visualisation de grandes masses de données ont été réalisés principalement dans le domaine des interfaces homme-machine (Hendley *et al.*, 1995; Nigay *et al.*, 1998) même s'il y a des travaux spécifiques aux systèmes de recherche d'information (Chalmers *et al.*, 1992; Hearst *et al.*, 1997). Les problèmes du passage à l'échelle dans la taille des *corpus* dans un système de recherche d'information rendent plus difficile l'évaluation et donc la perception des bonnes réponses dans un espace devenu plus vaste. Les études liées à la présentation de grandes masses de données doivent donc s'adapter à la recherche d'information.

3.4.2. *Etat de l'art*

Nous présentons dans ce qui suit une revue basée sur la taxonomie classique en visualisation : 1D, 2D et 3D, de la présentation des résultats d'une recherche. De manière générale, il s'agit en premier lieu de présenter les documents proposés par le système en même temps qu'un ordre de pertinence. D'autres éléments annexes peuvent être proposés en réponse comme une partie du contenu de la représentation interne des documents. Cette analyse est issue d'une étude spécifique sur les interfaces pour la recherche d'informations (Chevallet *et al.*, 2002).

3.4.2.1. 1D

La forme la plus simple et aussi la plus courante de présentation des résultats est la liste, en une dimension. Cette dimension correspond généralement à l'ordre de pertinence calculé par le système. Cette présentation, souvent textuelle, est celle employée par tous les moteurs de recherche avec des variantes consistant à regrouper sous une arborescence, donc à l'aide d'une deuxième dimension, les documents appartenant à un même site.

Cette présentation est celle adoptée par la quasi-totalité des moteurs de recherche sur la Toile. Néanmoins, elle limite fortement la quantité d'informations présentables en même temps. Lorsque l'on veut afficher plus d'informations simultanément, il faut utiliser deux dimensions.

3.4.2.2. 2D

La représentation en deux dimensions permet de représenter beaucoup plus d'informations que l'ordre de pertinence entre les documents. La représentation peut être l'occasion de laisser percevoir le *corpus*. On peut, par exemple, donner l'impression de la quantité de réponses, donc la taille du *corpus* pour un terme (*starfield*). Cette perception s'apparente à la notion de rappel. Contrairement au simple chiffre exprimant l'ordre de grandeur du nombre de réponse, une interface comme TIAPRI, permet de percevoir l'importance d'un terme dans l'ensemble des documents en réponse. Il y a également un couplage direct entre la requête et les réponses car l'importance d'un terme dans la requête est visible dans l'interface et a une influence sur la visualisation des résultats.

Dans ce genre de représentation, il y a perte de l'information textuelle (titre, auteur, aperçu du contenu, etc.), au profit d'une perception plus globale. Il existe néanmoins des techniques visuelles permettant d'accéder momentanément au contenu du texte associé à la réponse comme les techniques des lentilles magiques.

Le type de média influence également les possibilités de représentations efficaces de grandes quantités d'informations. Dans le cas des images, l'affichage des résultats en deux dimensions dans une matrice d'images est plus efficace que le texte, car il est plus facile de percevoir le contenu d'une image, même réduite, qu'un texte en petit format.

Pour augmenter le nombre d'images perçues, il existe des techniques de compressions visuelles comme la vue « en oeil de poisson », qui couplée avec la navigation, permet de percevoir et de manipuler un très grand nombre de résultats simultanément. Par exemple dans le système « ostensif » de Campbell, le système présente en grand format l'image en cours de sélection, et présente dans un graphe compressé d'images, le parcours de l'utilisateur, et donc toutes les images vues, ainsi qu'une nouvelle sélection d'images proposées à l'utilisateur pour poursuivre son exploration. Le système est dit ostensif car aucune requête ne lui est explicitement posée ; elle est déduite de son comportement.

3.4.2.3. 3D

L'utilisation de trois dimensions nécessite une représentation en deux dimensions sur la surface de l'écran. Pour permettre à l'utilisateur de se construire une représentation 2D de cette image 3D, les interfaces se basent sur une représentation *métaphorique* du monde réel. Par exemple, le système LyberWorld (Hendley *et al.*, 1995) met en place la métaphore du cône. Un cône représente un ensemble de documents groupés autour d'un thème. Ces cônes sont ensuite organisés en arbres selon la structure hiérarchique des thèmes des réponses. L'interaction se fait par manipulation directe sur les cônes, par translation et rotation. Ce système intègre également la métaphore du « paysage ».

3.4.3. *Perspectives*

Les perspectives dans la visualisation des résultats sont donc à examiner selon les trois axes suivants :

– Technologie de visualisation : l'accès à l'information est dépendant de l'artefact technique qui met en relation l'utilisateur avec l'information. L'écran, le clavier et la souris sont toujours les moyens les plus utilisés pour accéder à l'information. L'évolution technologique va dans le sens de davantage de mobilité, vers une informatique omniprésente et multimédia. Les techniques de visualisation seront dépendantes de cette évolution.

– Technique de visualisation : l'accès à de grandes quantités d'information pose le problème de leur perception par l'utilisateur. En premier lieu, le passage à l'échelle influence directement le nombre de réponses potentielles à une requête et pose donc le problème de la navigation parmi ces réponses. Par ailleurs, l'utilisateur doit être capable de s'assurer que l'information proposée par le système est bien pertinente, non pas dans l'absolu mais par rapport au *corpus*. Dans ce cas également, une perception du contenu potentiel peut l'aider à établir son propre jugement de pertinence.

– Couplage entre interaction et modèle de recherche d'information : le modèle ostensif est un exemple d'influence mutuelle entre un modèle de recherche d'information et la technique de navigation dans les documents. Si de nouveaux modèles deviennent nécessaires pour permettre un passage à l'échelle, alors une étude de leur influence sur la présentation des résultats ainsi que sur la manière dont l'utilisateur pourra communiquer son besoin d'information est également nécessaire. Ce dernier point évoque le domaine connexe à la visualisation qui est celui de l'« expressibilité » de la machine et du besoin de l'utilisateur en une communication réellement multimodale : voix, image, gestuelle.

4. Evaluation des performances de recherche : vers des protocoles adaptés à la taille des *corpus*

Depuis 1984, le DARPA (*Department XXX*) encourage la recherche en recherche d'information. C'est ainsi qu'est né, avec le partenariat du NIST (*National XXX*), la campagne d'évaluation internationale TREC qui a jeté les bases des techniques d'évaluation en mettant en place des protocoles d'évaluation et des méthodologies de construction de *corpus* de test. Ces derniers sont construits à l'issue des résultats obtenus par chaque système participant selon une méthode de *pooling* ou échantillonnage. La qualité de ces *corpus* de test conditionne l'ensemble de l'évaluation puisqu'ils en représentent le mètre-étalon. Le type d'évaluation proposé dans TREC est orienté vers une approche comparative des systèmes et reste en vigueur actuellement. Le traitement du paramètre volume a été introduit tout d'abord, dans la campagne TREC, à l'aide de la tâche *Very Large Corpus* (en 1996) puis plus récemment (en 2004) *via* la tâche *Tera-Byte*.

Dans la perspective du passage à l'échelle, un certain nombre de difficultés doivent être résolues concernant la méthodologie d'élaboration des *corpus* de test pour que certains « effets de bord » ne faussent pas les résultats (Zobel, 1998). En effet, il faut tenir compte de trois éléments :

– les tailles de collections étant très importantes, et par conséquent le nombre de documents à classer nettement élevé, il s'ensuit un dilemmme entre la taille opportune de l'échantillon, à retenir expérimentalement, et la faisabilité du jugement manuel. En pratique, une meilleure exhaustivité des résultats nécessite, dans le cas de larges collections, une augmentation conséquente de la taille de l'échantillon. Or, cela implique un investissement considérable lors du déroulement de la phase de collecte des jugements ;

– Les valeurs de rappel ne sont pas fiables, et par conséquent la qualité de la comparaison issue de l'évaluation est remise en cause. En effet, seuls les documents figurant à une distance inférieure ou égale à la taille fixée de l'échantillon (relativement petite par rapport à la taille de la collection en raison du point précédent) sont jugés. Au de-là, les autres documents sont jugés *par défaut* non pertinents alors qu'ils peuvent être effectivement pertinents. Cela biaise les résultats, notamment dans le cas de systèmes qui rappellent des documents pertinents dits difficiles ou de manière générale, qui ne contribuent pas à la constitution de l'échantillon ;

– Les résultats obtenus, à concurrence de la taille de l'échantillon, ne sont pas rationnellement extrapolables, notamment dans le cas de requêtes ayant de très nombreux documents pertinents ou de requêtes dont le nombre de documents communs trouvés par l'ensemble des participants est faible ;

Dans le but de remédier à ces difficultés, des améliorations de l'échantillonnage ont été proposées (Zobel, 1998). On cite particulièrement la méthode ISJ (*XXX*) qui utilise un système de recherche d'information interactif pour la sélection des documents à juger et la méthode *Move To Front* basée sur la sélection, pour chaque système participant, d'un nombre de documents dépendant de ses performances.

Par ailleurs l'objectif de *haute précision* induit par le volume nécessite sans doute la révision des métriques standard d'évaluation de la pertinence (Shah *et al.*, 2004). En l'état actuel de nos connaissances, peu de travaux ont traité ces questions qui restent donc à ce jour largement ouvertes.

5. Bilan

Ce travail a cherché à cerner la problématique du passage à l'échelle en recherche d'information, à identifier ses dimensions puis à présenter, pour chacunes d'elles, un court état de l'art, illustrant les principales difficultés, et des perspectives. Le travail de synthèse entrepris dans ce cadre a en outre permis de capitaliser les acquis de chaque partenaire, de synthétiser et d'intégrer les réflexions qui en sont issues, puis de dresser les bases de travaux futurs qui s'inscrivent dans les principales directions suivantes : organisation et annotation de collections volumineuses, personnalisation

du processus de recherche d'information, élaboration de méthodologies d'évaluation des performances de recherche dans des collections de grande échelle, définition de techniques de visualisation adaptées au volume.

5.1. *Organisation et annotation des collections*

L'objectif de cet axe est de mener des réflexions sur des modèles (physique, logique et théorique) de représentation des collections et des informations. Ces modèles ont pour objectif d'intégrer différents niveaux d'abstraction (du métadocument vers des granules plus fins) et de permettre une réduction de l'espace de recherche lors du traitement de la requête et une présentation synthétique des informations. Les recherches futures doivent mettre l'accent sur les points suivants :

– caractérisation des collections et définition de techniques de segmentation sur la base de métadonnées descriptives (connectivité des informations, contenu, fraîcheur, etc.) permettant de diriger le processus d'évaluation de requêtes ;

– augmentation de la granularité de l'information avec la définition d'un support théorique pour l'expression de la représentativité locale et globale.

5.2. *Personnalisation du processus de recherche d'information*

L'objectif de cet axe est de proposer des modèles de recherche d'information permettant une meilleure prise en compte de l'utilisateur et des usages notamment sur le type de besoin (privilégiant sans doute la précision sur le rappel). Dans ce contexte, les travaux s'orientent davantage vers :

– l'étude de la composante utilisateur au travers des caractéristiques (besoins, connaissances, préférences, contexte, etc.) permettant de situer son profil par rapport au contenu de la collection ;

– l'exploitation du profil de l'utilisateur pour réduire l'échelle de la collection considérée en réponse à un besoin en information ;

– l'analyse de la traçabilité d'un utilisateur au travers des sessions de recherche antérieures en vue d'ajuster l'espace de recherche en fonction de l'évolution de son profil ;

– la prise en compte d'autres paramètres dans le traitement de la requête : utilisateur (profil), usage, niveau d'abstraction de l'information.

5.3. *Elaboration de méthodologies d'évaluation des performances de recherche dans des collections de grande échelle*

L'évaluation des performances d'un système de recherche d'information est évidemment un point fondamental. L'objectif de cet axe est d'étudier les protocoles

d'évaluation intégrant l'efficience et l'efficacité. Dans la perspective d'évaluation des différents scénarios dans le cas de collections volumineuses, des réflexions sont à mener dans le sens de :

– la définition de nouveaux protocoles d'évaluation qui tiennent compte tant du volume d'informations traitées que de leur disparité ;

– la mise au point de nouvelles métriques adaptées à l'évaluation conjuguée de l'efficacité et de l'efficience des systèmes de recherche d'information ;

– la mise en œuvre de méthodologies de construction de collections de test volumineuses.

5.4. *Techniques de visualisation adaptées au volume*

Le support de présentation des résultats sur de grands volumes influence la satisfaction de la recherche des informations. Cette activité de recherche est également dépendante de l'évolution des usages et des techniques. Dans ce contexte, les travaux doivent étudier :

– l'adaptation des techniques de visualisation de grandes quantités d'informations ;

– l'influence du passage à l'échelle sur l'interface de sortie en relation avec une prise en compte de l'utilisateur et de son besoin dans un modèle de recherche d'information donné, comme le modèle ostensif ;

– les apports dans l'évolution des technologies de visualisation, la variété des possibilités d'affichage des résultats demandant la conception d'interfaces plus « plastiques » (Calvary *et al.*, 2001) et capables de mieux gérer la multimodalité, en entrée (une requête à partir d'une photo) comme en sortie ;

– le développement de l'informatique ubiquitaire (ordinateurs ultra-portables, *palm-top*, *wi-fi*, etc.) qui engendre de nouvelles réflexions et de nouvelles pistes de recherche, comme la recherche d'information *située*, c'est-à-dire fonction du lieu et du moment où s'effectue une recherche.

Les grands axes d'étude identifiés dans ce rapport méritent d'être investis en collaboration avec d'autres communautés. La préparation des collections peut bénéficier des études sur l'extraction, la classification et les résumés de données. L'évaluation de requêtes peut bénéficier des travaux sur les systèmes de gestion de données, notamment avec l'émergence des grappes de calcul et de stockage ainsi que sur les travaux en personnalisation de l'information.

Remerciements

Ce travail a été rendu possible grâce à l'action spécifique 91 « Passage à l'échelle dans la taille des *corpus* » du réseau thématique prioritaire « Bases de données et recherche d'information » du CNRS.

6. Bibliographie

Bayley P., Hawking D., A Parallel architecture for query processing over a Terabyte of text, Technical report, The Australian National University, 1996.

Beigbeder M., Mercier A., « Étude des distributions de tf et de idf sur une collection de 5 millions de pages html », *Atelier de recherche d'information sur le passage à l'échelle, congrès INFORSID*, Nancy, France, June, 2003.

Berrani S.-A., Amsaleg L., Gros P., « Recherche par similarités dans les bases de données multidimensionnelles : panorama des techniques d'indexation », *Ingénierie des systèmes d'information*, vol. 7, n° 5-6, p. 9-44, 2002.

Berry M. W., « Large-Scale Sparse Singular Value Computations », *The International Journal of Supercomputer Applications*, vol. 6, n° 1, p. 13-49, Spring, 1992.

Berry M. W., Darmac Z., Jessup E. R., « Matrices, vector spaces and information retrieval », *SIAM*, vol. 41, n° 2, p. 335-362, 1999.

Calvary G., Coutaz J., Thevenin D., « A Unifying Reference Framework for the Development of Plastic User Interfaces », *Proceedings of EHCI'01,IFIP WG2.7 (13.2) Conference*, Springer-Verlag, Toronto, Canada, May, 2001.

Chalmers M., Chitson P., « Bead : explorations in information visualization », *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, Copenhagen, Denmark, p. 330-337, 1992.

Chevallet J.-P., Nigay L., « Les interfaces pour la Recherche d'Information », *in* C. Paganelli (ed.), *Interaction homme-machine et recherche d'information*, Hermès Science, Lavoisier, chapter 2, p. 65-102, 2002.

Dumais S., « LSI meets TREC : A status report », *Proceedings of the 1st Text REtrieval Conference (TREC-1)*, NIST Special Publication, p. 137-152, March, 1993.

Elias P., « Universal codeword sets and representations of the integers », *IEEE transactions on information theory*, vol. 21, n° 2, p. 194-203, 1977.

Franz M., McCarley J. S., « How Many Bits are Needed to Store Term Frequencies ? », *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, Tampere, Finland, p. 377-378, 2002.

Frieder O., Grossman D. A., Chowdhury A., Frieder G., « Efficiency considerations for scalable information retrieval servers », *Journal of digital information*, 2000.

Hawking D., Voorhees E., Craswell N., Bailey P., « Overview of the TREC-8 Web Track », *Proceedings of the of the 8th Text REtrieval Conference (TREC-8)*, NIST Special Publication, p. 131-150, 1999.

Hearst M. A., Karadi C., « Cat-a-Cone : an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy », *Proceedings of the 20th annual*

international ACM SIGIR conference on Research and development in information retrieval, Philadelphia, Pennsylvania, United States, p. 246-255, 1997.

Heinz S., Zobel J., « Efficient single pass index construction for text databases », *Journal of American Science on Information and Technology (JASIST)*, vol. 54, n° 8, p. 713-729, 2003.

Hendley R. J., Drew N. S., Wood A. M., Beale R., « Narcissus : Visualizing Information », *in* N. Gershon, S.G.Eick (eds), *Proceedings of Information Visualization Symposium*, IEEE CS Press, Los Alamitos, California, p. 90-96, 1995.

Huang L., A survey on web information retrieval technologies, Technical report, ECSL, 2000.

Jain A., Goharian N., « On Parallel Implementation of Sparse Matrix Information Retrieval Engine », *Proceedings of the International Multi-conferences in Computer Science : on Information and Knowledge Engineering (IKE)*, 2002.

Kobayashi M., Takeda K., « Information retrieval on the web », *ACM Computing Surveys*, vol. 32, n° 2, p. 144-173, 2000.

Kokiapoulou E., Saad Y., « Polynomial filtering in latent semantic indexing for information retrieval », *Proceedings of the 27th annual international ACM SIGIR Conference on research and development in information retrieval (SIGIR)*, Sheffield, U.K., p. 104-111, July 25-29, 2004.

Kowalski G., *Information retrieval systems : theory and implementation*, Kluwer Academic Publishers, Boston, 1997.

Lee D. L., Ren L., « Document Ranking on Weight-Partitioned Signature Files », *ACM Transactions*, vol. 14, n° 2, p. 109-137, 1996.

Lehoucq R., Sorensen D., « Deflation techniques for an implicitly restarted Arnoldi iteration », *SIAM Review*, vol. 17, p. 789-821, 1996.

Moffat A., Zobel J., « Self-Indexing Inverted Files for Fast Text Retrieval », *ACM Transactions on Information Systems*, vol. 14, n° 4, p. 349-379, 1996.

Newby G. B., « The science of large scale information retrieval », *Internet archives*, 2000.

Nigay L., Vernier F., « Design method of interaction techniques for large information spaces », *proceedings of the international conference on Advanced Visual Interfaces (AVI'98)*, L'Aquila, Italy, p. 37-46, May, 1998.

Sameh A. H., Wisniewski J. A., « A trace minimisation algorithm for the generalized Eigenvalue problem », *SIAM Review*, vol. 19, n° 3, p. 1243-1259, 1982.

Scholer F., Williams H. E., Yiannis J., Zobel J., « Compression of inverted indexes for fast query evaluation », *Proceedings of the 25th ACM SIGIR Conference on research and development in information Retrieval*, p. 11-15, August, 2002.

Shah C., Croft W. B., « Evaluating High Accuracy Retrieval Techniques », *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, Sheffield, United Kingdom, p. 2-9, 2004.

Shieh W.-Y., Chen T.-F., Shann J. J.-J., Chung C.-P., « Inverted file compression through document identifier reassignment », *Inf. Process. Manage.*, vol. 39, n° 1, p. 117-131, 2003.

Skiena S. S., *The algorithm design manual*, Springer-Verlag New York, Inc., New York, NY, USA, 1998.

Spink A., Wolfarm D., Jansen M. B., Saracevic T., « Searching the Web : the public and their queries », *Journal of American Science on Information and Technology (JASIST)*, vol. 52, n° 3, p. 226-234, 2001.

Tang C., Dwarkadas S., Xu Z., « On scaling latent semantic indexing for large peer to peer systems », *Proceedings of the 27th annual international ACM SIGIR Conference on research and development in information retrieval (SIGIR)*, Sheffield, U.K., p. 112-121, July 25-29, 2004.

Voorhees E., Harman D., « Overview of the Eighth Text REtrieval Conference », *Proceedings of the 8th Text REtrieval Conference (TREC-8)*, NIST Special Publication, p. 1-24, 1999.

Williams H. E., Zobel J., « Searchable words on the Web », *International journal of digital libraries*, vol. 5, n° 2, p. 99-105, 2005.

Zobel J., « How reliable are the results of large scale information retrieval experiments », *Proceedings of the 21th ACM SIGIR Conference on research and development in information Retrieval*, p. 307-314, August, 1998.

Zobel J., Moffat A., Sacks-Davis R., « An efficient indexing technique for full-text database systems », *Proceedings of the 8th International conference on very large databases (VLDB)*, p. 352-362, 1992.