Documentation et bibliothèques

Vers une nouvelle génération d'outils d'analyse et de recherche d'information

A New Generation of Analysis and Retrieval Tools

Hacia una nueva generación de herramientas de análisis y de investigación de información

Dominic Forest

Volume 55, numéro 2, avril–juin 2009

URI : https://id.erudit.org/iderudit/1029091ar
DOI : https://doi.org/10.7202/1029091ar

Aller au sommaire du numéro

Éditeur(s)

Association pour l'avancement des sciences et des techniques de la documentation (ASTED)

ISSN

0315-2340 (imprimé)
2291-8949 (numérique)

Découvrir la revue

Citer cet article

Forest, D. (2009). Vers une nouvelle génération d'outils d'analyse et de recherche d'information. *Documentation et bibliothèques*, *55*(2), 77–89. https://doi.org/10.7202/1029091ar

Résumé de l'article

Les récents efforts visant à favoriser la diffusion et la circulation de l'information en format numérique ont contribué au phénomène de l'infobésité (*information overload*). Il est désormais important de concevoir des outils de recherche d'information plus adaptés aux besoins des utilisateurs afin de leur permettre de récupérer les documents pertinents répondant à leurs besoins informationnels. Dans cet article, nous ferons état, dans un premier temps, de certaines observations sur les conséquences découlant des limites des outils traditionnels en recherche d'information numérique. Dans un deuxième temps, nous exposerons les concepts et les techniques de base du domaine de la fouille de textes, en insistant sur les opérations de classification et de catégorisation automatiques. Finalement, nous montrerons comment certaines techniques de fouille de textes peuvent contribuer au développement d'une nouvelle génération d'outils de recherche d'information.

DOCUMENTATION BIBLIOTHÈQUES

Vers une nouvelle génération d'outils d'analyse et de recherche d'information

DOMINIC FOREST
Professeur adjoint
École de bibliothéconomie et des sciences de l'information
Université de Montréal
dominic.forest@umontreal.ca

RÉSUMÉ | ABSTRACTS | RESUMEN

Les récents efforts visant à favoriser la diffusion et la circulation de l'information en format numérique ont contribué au phénomène de l'infobésité (information overload). *Il est désormais important de concevoir des outils de recherche d'information plus adaptés aux besoins des utilisateurs afin de leur permettre de récupérer les documents pertinents répondant à leurs besoins informationnels. Dans cet article, nous ferons état, dans un premier temps, de certaines observations sur les conséquences découlant des limites des outils traditionnels en recherche d'information numérique. Dans un deuxième temps, nous exposerons les concepts et les techniques de base du domaine de la fouille de textes, en insistant sur les opérations de classification et de catégorisation automatiques. Finalement, nous montrerons comment certaines techniques de fouille de textes peuvent contribuer au développement d'une nouvelle génération d'outils de recherche d'information.*

A New Generation of Analysis and Retrieval Tools

The recent efforts to increase the dissemination and circulation of information in numeric format have led to a phenomenon known as information overload. It is now imperative to develop retrieval tools that are better adapted to the users' needs and that will enable them to retrieve relevant documents that meet their information needs. In this article, we will begin with a summary of observations of the consequences of the limits of the traditional tools used in numeric information retrieval. Following this, we will describe the concepts and techniques of textual searching, emphasizing automatic classification. Lastly, we will demonstrate how certain textual searching techniques contribute to the development of a new generation of information retrieval tools.

Hacia una nueva generación de herramientas de análisis y de investigación de información

Los recientes esfuerzos tendientes a favorecer la difusión y la circulación de la información en formato digital han contribuido al fenómeno de la infobesidad (sobrecarga de información). Es importante, de aquí en adelante, diseñar herramientas de búsqueda de información que se adapten mejor a las necesidades de los usuarios a fin de facilitarles la recuperación de documentos pertinentes que respondan a sus necesidades informacionales. En este artículo, realizaremos, en un primer momento, ciertas observaciones sobre las consecuencias que derivan de los límites de las herramientas tradicionales para la búsqueda de información digital. En un segundo momento, expondremos los conceptos y las técnicas de base del dominio del registro de textos, insistiendo sobre las operaciones de clasificación y de categorización automáticas. Finalmente, mostraremos cómo determinadas técnicas de registro de textos pueden contribuir al desarrollo de una nueva generación de herramientas de búsqueda de información.

Introduction

Les dix dernières années ont été caractérisées par une hausse importante du nombre d'initiatives visant à numériser et à rendre disponible, très souvent sur Internet ou à l'intérieur d'un Intranet, le patrimoine informationnel des organisations. De nos jours, la majorité des gestionnaires de l'information ne remettent plus en question les nombreux avantages que peuvent procurer les documents en format numérique (reproductibilité quasi parfaite et illimitée, dissémination rapide de l'information, etc.). Parallèlement, les nombreux avantages associés à la documentation numérique ont motivé le développement d'infrastructures performantes dédiées à la numérisation, ainsi qu'à la diffusion de l'information. À cet égard, les manifestations du numérique sont nombreuses et des plus variées. Ainsi, Bibliothèques et Archives nationales du Québec (BAnQ) est un des principaux acteurs d'un projet visant à permettre la diffusion en format numérique de plusieurs dizaines de milliers de documents généalogiques[1]. À l'échelle internationale, il importe de souligner le succès, tant en termes de la qualité de la numérisation que de la quantité de documents disponibles et consultés du projet *Européana*[2], de la bibliothèque numérique *Gallica*[3] ou encore du projet ARTFL[4]. Dans le domaine de la recherche universitaire, l'intérêt pour la diffusion numérique des résultats de la recherche scientifique est à la base de nombreux travaux en cours cherchant à promouvoir le développement de cyberinfrastructures pour la diffusion numérique des résultats de la recherche. Dans la Francophonie, les deux principaux chantiers dans ce domaine sont les projets Synergies[5] et Adonis[6].

L'évolution du numérique, indissociable du Web, est d'une telle ampleur qu'il est très difficilement quantifiable[7]. Cependant, les inconvénients du numérique,

1. http://www.banq.qc.ca/portal/dt/genealogie/genealogie.jsp
2. http://dev.europeana.eu/
3. http://gallica.bnf.fr/
4. http://humanities.uchicago.edu/orgs/ARTFL/
5. http://www.synergiescanada.org/
6. http://www.tge-adonis.fr/
7. Certaines études, dont la plus citée a été menée à l'Université de Berkeley en 2003 (*How much information ?*), ont tenté de chiffrer la croissance du numérique. Si toutes les études démontrent que la quantité de documents disponibles en format numérique augmente radialement, aucun consensus n'émerge des études lorsqu'il s'agit de chiffrer précisément cette évolution.

En d'autres termes, les technologies actuelles d'assistance à la recherche d'information numérique ne sont que très peu sensibles au contenu sémantique des documents.

dans une perspective de recherche d'information, existent. Ils ont d'ailleurs fait l'objet de nombreux travaux réalisés en milieu professionnel. À ce titre, selon des études réalisées par les firmes IDC, Ford Motor Company et Reuters et présentées par Feldman en 2004, les gestionnaires de la connaissance en entreprise affirmaient passer en moyenne entre 15 % et 35 % de leur temps de travail à chercher – très souvent infructueusement – de l'information sur le Web ou à l'intérieur d'un Intranet corporatif. Toujours selon ces études, 50 % des recherches effectuées sur le Web s'avèrent infructueuses. En outre, dans un contexte organisationnel, on estime que 40 % des usagers d'entreprises ne peuvent pas trouver sur leur Intranet l'information nécessaire à l'accomplissement de leur travail. Les conséquences de ces échecs sont radicales. Selon Feldman, 50 % des recherches effectuées en ligne en 2004 étaient abandonnées en raison de la faible pertinence des résultats générés par les moteurs de recherche.

Plus de quatre ans après la parution des résultats décrits par Feldman, nous pouvons nous interroger sur la pertinence actuelle de ces conclusions. Compte tenu de l'évolution rapide des technologies de l'information, est-il possible que les problèmes et les insatisfactions identifiés en 2004 aient été en majeure partie corrigés quatre ans plus tard ? C'est précisément la question à laquelle s'est attaquée Feldman en 2008. Les conclusions de sa nouvelle étude sont tout aussi radicales que celles auxquelles elle était arrivée en 2004. En effet, cette nouvelle étude constate que les mêmes problèmes et les mêmes insatisfactions sont encore déplorés par les utilisateurs du Web en contexte de recherche d'information. Les propos de la chercheure sont d'ailleurs sans équivoque : « *Knowledge workers spend a lot of time looking for information.* [...] *Our surveys over the last six years show little change in the time that business users search online : roughly 9,5 hours each week.* » (Feldman, 2008, p. 9). Si les résultats de la seconde étude sont étroitement comparables à ceux de la première, ils divergent néanmoins sur un élément : l'importance du Web en tant que source première d'information. En 2004, l'utilisation du Web en tant que source d'information était importante, mais néanmoins diluée parmi d'autres sources d'information (base de données spécialisées, documentation en format papier, etc.). En contrepartie, en 2008, le Web constitue pour beaucoup la seule véritable source d'information consultée pour combler un besoin informationnel. Ainsi, selon la seconde étude menée par Feldman, 62 % des répondants affirment que le Web constitue leur première source d'information. Le Web est loin devant la deuxième source d'information la plus utilisée, à savoir les personnes physiques (collègues ou autres), une source exploitée par 8 % seulement des personnes interrogées.

Les problèmes identifiés en 2004 sont donc toujours observables en 2008. Ils sont d'ailleurs encore plus marqués, compte tenu de l'importance accordée au Web en tant que source principale d'informations. L'importance du Web, couplée aux limites des outils de recherche en ligne, ainsi qu'aux nombreuses insatisfactions des utilisateurs en contexte de recherche d'information, a motivé le développement de plusieurs solutions technologiques. Afin d'améliorer les performances des moteurs de recherche en ligne, des normes d'encodage des documents ont été développées et différents outils spécialisés ont été conçus afin de favoriser la gestion et la diffusion d'information bien ciblée (dépôts institutionnels, systèmes de gestion de contenu Web, blogue, etc.). On constate cependant que la majorité des solutions proposées, bien qu'elles soient de plus en plus sensibles aux exigences des moteurs de recherche en ligne, s'inscrivent dans le même paradigme de recherche d'information que celui qui est déjà en place. Il s'agit d'un paradigme fondé principalement sur l'indexation des documents (pris individuellement) et sur l'appariement entre les termes d'une requête et les termes retenus lors de l'indexation automatique des documents par les moteurs de recherche. Dans cette perspective, on constate que les applications ainsi que les stratégies d'analyse et de repérage de l'information numérique sont réalisées en ne considérant que la dimension de « surface » des documents, au détriment de leur véritable contenu informationnel. En d'autres termes, les technologies actuelles d'assistance à la recherche d'information numérique ne sont que très peu sensibles au contenu sémantique des documents. De plus, les modalités de présentation de l'information que l'on retrouve dans les moteurs de recherche d'information sont souvent minimales – les documents n'apparaissant que sous forme de listes. Pour cette raison, elles ne facilitent que très rarement l'accès à l'information.

En marge des outils de recherche d'information traditionnels, on trouve des prototypes de moteur de recherche qui tiennent compte de certaines dimensions sémantiques du contenu des documents. Ces nouveaux moteurs, souvent déployés au départ à l'intérieur d'Intranets corporatifs, tentent d'offrir des solutions originales aux problèmes auxquels se heurtent les moteurs de recherche traditionnels. En effet, ils cherchent à regrouper les documents partageant certains contenus communs, à identifier les principaux traits thématiques présents dans les documents récupérés et à présenter

graphiquement les résultats de recherche. Nous défendons l'idée selon laquelle les développements dans le domaine de la fouille de textes sont à la base de l'émergence d'une nouvelle génération d'outils d'analyse et de recherche d'information davantage sensibles au contenu informationnel et sémantique des documents.

Dans cet article, nous définirons d'abord le domaine de la fouille de textes, puis exposerons les principes fondamentaux des deux processus de fouille à la base de nouvelles stratégies de recherche d'information. Finalement, nous présenterons deux exemples d'outils de recherche sur le Web qui reposent en grande partie sur l'utilisation de techniques de classification et de catégorisation automatique pour la recherche d'information en ligne.

La fouille de textes[8]

Définition du domaine

La fouille de textes (aussi nommée forage de textes, ou encore *text mining*) est un nouveau domaine de recherche interdisciplinaire dont l'objectif général est le développement de techniques, d'algorithmes, d'applications informatiques et de méthodologies applicatives afin d'extraire et de structurer automatiquement, ou semi-automatiquement, de nouvelles informations à partir de grands corpus de documents textuels non structurés ou semi-structurés.

Ce domaine récent a fait l'objet de plusieurs définitions variant en fonction de l'angle sous lequel il est abordé. Ainsi, les définitions qui ont été proposées par la communauté informatique accordent une plus grande importance aux questions traitant des algorithmes de fouille. En contrepartie, les définitions proposées par les spécialistes du traitement automatique des langues (TAL) sont davantage focalisées sur les enjeux linguistiques et terminologiques. Malgré les divergences de points de vue, il semble possible de dégager un noyau commun consensuel. Celui-ci réside dans les opérations d'identification, d'extraction et de mise en relation de nouvelles informations présentes dans de grands corpus de documents. À cet égard, la définition proposée par Hearst (2003) permet de bien cerner les principaux traits caractéristiques du domaine de la fouille de textes. Selon Hearst, « *la fouille de textes est la découverte (à l'aide d'outils informatiques) de nouvelles informations en extrayant différentes données provenant de plusieurs documents textuels. Un élément fondamental de ce processus réside dans les relations identifiées entre les informations extraites afin d'identifier de nouveaux faits ou de nouvelles hypothèses à explorer* » (2003, notre traduction).

8. Cette section est inspirée de notre contribution à un ouvrage d'introduction aux sciences de l'information, sous la direction de Clément Arsenault et Jean-Michel Salaün (2009, à paraître).

La première étape de tout processus de fouille de données réside dans le développement ou la constitution d'un corpus de documents.

Les origines de la fouille de textes proviennent principalement d'un autre domaine, celui de la fouille de données. Le domaine de la fouille de textes peut d'ailleurs être perçu comme une variante plus complexe que celui de la fouille de données. Si ces deux domaines font appel à des algorithmes et à des processus de traitement identiques, ou à tout le moins très comparables, ils sont radicalement différents en ce qui a trait à la nature des données qu'ils traitent. En effet, les données traitées dans le domaine de la fouille de données sont normalement de nature structurée (bases de données, entrepôts de données, etc.). Par contre, les processus de fouille de textes sont appliqués sur des données textuelles (de grands corpus de documents) dont l'une des principales caractéristiques est d'être beaucoup moins structurées.

Démarche méthodologique

La démarche méthodologique caractéristique des applications de fouille de textes est également inspirée de celle que l'on retrouve au cœur de nombreux projets de fouille de données. Cette démarche est de nature itérative.

La première étape de tout processus de fouille de données réside dans le développement ou la constitution d'un corpus de documents. Il est essentiel que le corpus de documents soit constitué en tenant compte des objectifs à atteindre par le processus de fouille. À l'étape de constitution du corpus, quatre grandes familles de caractéristiques doivent être évaluées et prises en considération. La constitution d'un corpus à des fins de fouille de textes implique certains choix en ce qui concerne les caractéristiques :

- générales (provenance, taille, date de création, etc.) ;
- technologiques (support, format, etc.) ;
- informationnelles (thématiques et sujets abordés) ;
- linguistiques des documents (langue, genre, registres, etc.).

Cette opération est fondamentale, car la qualité des résultats ultérieurs dépend directement de la justesse des choix effectués lors de cette première étape.

La seconde étape de la démarche générique consiste à filtrer et, le cas échéant, à normaliser le lexique (l'ensemble des mots) du corpus. L'opération de filtrage du lexique est composée traditionnellement de plusieurs

La troisième étape de la démarche consiste à convertir le corpus initial dans un format pouvant être traité par les algorithmes de fouille.

sous-opérations. La première d'entre elles consiste à supprimer certains mots non pertinents pour l'analyse. Le filtrage du lexique peut être effectué à l'aide de plusieurs techniques, certaines étant de nature linguistique, d'autres de nature statistique. Une première opération a pour but de supprimer l'ensemble des mots fonctionnels présents dans le texte. Ce processus est réalisé en retirant les termes figurant dans une liste prédéfinie de mots fonctionnels. Il est aussi souhaitable d'appliquer certains filtres statistiques au lexique du corpus afin d'en éliminer les unités qui, tout en ne figurant pas dans la liste des mots fonctionnels, ne sont pas pertinentes pour l'analyse. La pertinence des termes est très étroitement associée à leur potentiel discriminant. Ainsi, il importe de supprimer les mots dont la fréquence est supérieure ou inférieure à certains seuils, souvent déterminés empiriquement.

Dans un dernier temps, il est d'usage d'appliquer un processus de généralisation sensible aux variantes sémantiques et syntaxiques présentes dans le corpus. Il importe alors d'appliquer au lexique du corpus une opération de lemmatisation[9]. L'opération de lemmatisation est réalisée d'abord en effectuant un marquage morphosyntaxique des différents lexèmes à analyser, ensuite en comparant ceux-ci à un dictionnaire. Ce processus permet de dégager une liste de lemmes propres à une langue donnée. S'il est impossible de lemmatiser les données à traiter (par manque de ressources linguistiques, par exemple), il est souhaitable de recourir à un processus d'amputation des terminaisons (*stemming*), lequel génère une liste de *stems* (racines).

La troisième étape de la démarche consiste à convertir le corpus initial dans un format pouvant être traité par les algorithmes de fouille. Cette opération est réalisée en structurant les documents du corpus en une matrice de vecteurs dans laquelle chaque document (ou segment de document) est représenté par l'absence ou la présence, binaire ou pondérée, de chaque unité lexicale retenue à l'étape précédente.

C'est à la quatrième étape de la démarche que sont réalisées les opérations permettant plus spécifiquement d'extraire et de structurer les informations présentes dans le corpus. Dans une perspective de fouille de textes, la majorité des opérations d'extraction et de structuration des informations sont réalisées en utilisant des algorithmes développés dans les domaines de l'intelligence artificielle et de l'apprentissage machine.

Il est possible de distinguer deux principales familles d'algorithmes en fonction de la quantité d'information externe au système qu'ils requièrent. La première famille d'algorithme est celle des techniques supervisées. La principale particularité des techniques supervisées réside dans leur capacité à projeter certaines caractéristiques des documents préalablement connues et apprises par le système sur un ensemble de documents pour lesquels les mêmes caractéristiques ne sont pas encore connues. En vertu de cette particularité, les techniques supervisées impliquent donc d'abord une phase d'apprentissage (réalisée sur un corpus d'apprentissage) et, ensuite, une phase de test (ou d'application) lors de laquelle l'apprentissage effectué par le système est projeté sur de nouveaux documents (en contexte de test ou d'application concrète). Les tâches de catégorisation automatique – qui consistent à attribuer une ou plusieurs catégories à chaque document d'un corpus – sont traditionnellement accomplies en utilisant des algorithmes d'apprentissage supervisés.

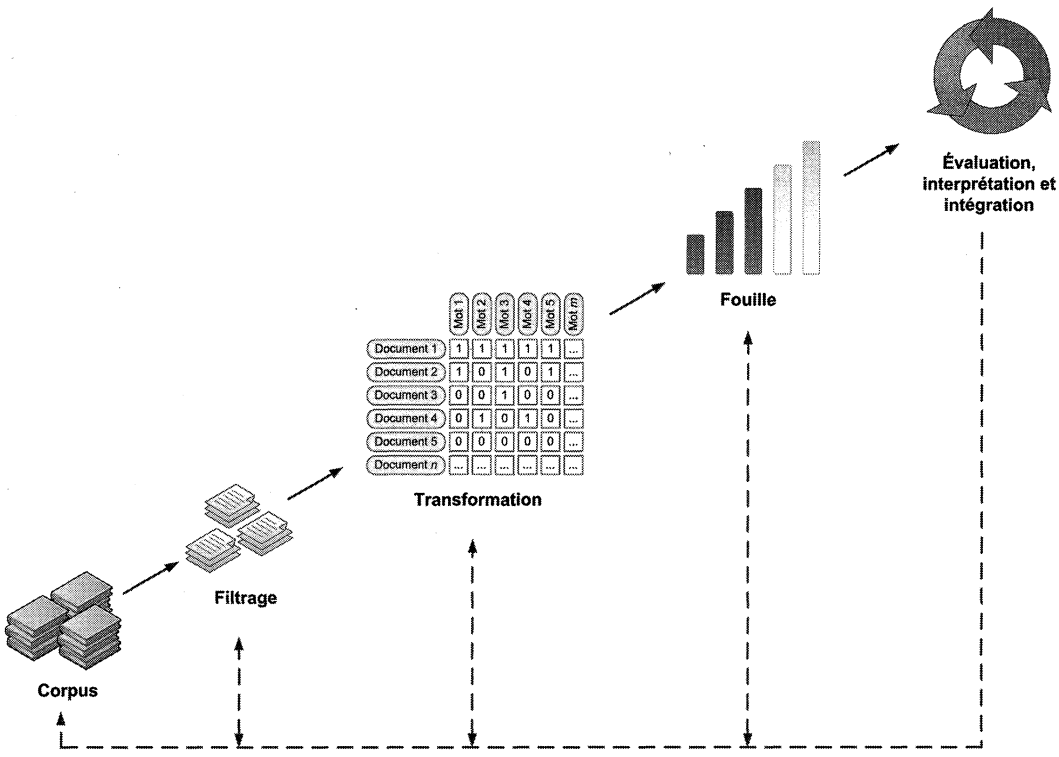
La seconde famille regroupe les algorithmes ne faisant intervenir aucune connaissance externe au système. Ces algorithmes, qualifiés de non supervisés, cherchent à extraire automatiquement certaines informations ou structures d'informations récurrentes dans un corpus. Contrairement aux algorithmes supervisés, les algorithmes non supervisés ne requièrent normalement pas d'ensembles d'apprentissage. Ils sont plutôt directement appliqués en contexte en tentant de décrire certaines régularités statistiques qui sont caractéristiques aux documents. Les tâches de classification automatique (*clustering*) – qui consistent à regrouper les documents d'un corpus dans un certain nombre de classes sur la base d'un ou de plusieurs critères de similarité – sont traditionnellement accomplies en utilisant des algorithmes non supervisés.

Outre ces deux principales opérations de catégorisation et de classification automatiques, on retrouve aussi dans le domaine de la fouille de textes des algorithmes davantage sensibles aux caractéristiques linguistiques des documents. Elles permettent d'assister l'identification d'entités nommées (dates, noms propres, lieux, etc.) et de liens sémantiques entre des termes.

La cinquième étape de la démarche réside dans l'interprétation, l'évaluation et l'intégration des résultats générés par les algorithmes de fouille de textes. Les opérations d'interprétation et d'évaluation sont des plus complexes, car elles sont dépendantes de plusieurs facteurs extrinsèques au processus de traitement des documents textuels. En effet, l'interprétation et l'évaluation des résultats de la fouille ne peuvent être réalisées adéquatement sans tenir compte de la nature des docu-

9. En linguistique, opération consistant à ramener les formes fléchies (conjuguées, plurielles) à des formes standard (infinitif ou singulier). La lemmatisation implique normalement un processus complexe visant à lever toute ambiguïté sémantique.

Figure 1
Méthodologie générique de la fouille de textes (d'après Fayyad *et al.* 1996)



ments traités, du contexte de réalisation des traitements, des objectifs poursuivis, etc.

Les algorithmes supervisés peuvent être évalués selon les mesures classiques de rappel et de précision. En contrepartie, l'évaluation des résultats générés par les algorithmes non supervisés ne peut être accomplie en utilisant de telles mesures objectives. Les processus non supervisés permettent principalement d'identifier et de décrire certaines caractéristiques récurrentes observables statistiquement à l'intérieur du corpus de documents. Pour cette raison, il est fréquent de comparer les résultats obtenus par un algorithme non supervisé à ceux obtenus en utilisant plusieurs autres algorithmes comparables. Ceci permet de s'assurer de la stabilité des résultats. Idéalement, les mêmes patrons (*patterns*) devraient être observés, indépendamment de l'algorithme utilisé. Lorsque cela est possible, il peut aussi être utile de comparer les résultats à ceux obtenus par une analyse manuelle.

Par ailleurs, l'interprétation des résultats des algorithmes de fouille ne peut être dictée par aucun cadre théorique qui ferait abstraction du contexte dans lequel l'opération de fouille est réalisée. Finalement, les résultats doivent normalement faire l'objet d'un processus d'intégration à l'intérieur d'une application finale plus complexe dans laquelle le processus de fouille ne constitue qu'une étape bien précise. Les applications finales intégrant des processus de fouille de textes sont de plus en plus nombreuses et variées. Parmi celles-ci, on trouve entre autres les applications de veille scientifique, de gestion électronique des documents et de recherche d'information. La figure 1, inspirée de Fayyad *et al.* (1996), présente les principales étapes de la méthodologie générique de fouille de textes.

Le développement d'outils d'analyse et de traitement de l'information numérique intégrant efficacement des fonctionnalités de fouille de textes est actuellement un important domaine de recherche dans les milieux académique et industriel. Certains domaines d'activités (dont, entre autres, ceux de la gestion des connaissances, de la bioinformatique et du génie biomédical) ont très rapidement perçu la pertinence d'intégrer des opérations de fouille de textes dans leurs pratiques. Depuis quelques années, plusieurs projets de recherche en sciences humaines et sociales explorent des modalités d'application de stratégies de fouille de textes en tenant compte des particularités d'analyse propres à chaque discipline (Forest, 2006). En sciences de l'information, les processus de fouille de textes sont de plus en plus souvent couplés aux outils de recherche d'information. Plus loin dans l'article, nous exposerons deux exemples d'outils de recherche d'information en ligne de nouvelle génération. Ces outils sont fondés sur l'application d'une opération de classification automatique des documents au sein du processus de recherche. Mais avant d'aborder plus précisément cette dimension applicative, il importe

La distinction entre les opérations de catégorisation et de classification se manifeste aussi dans le domaine de la fouille de textes.

d'expliciter les particularités de l'opération de classification automatique dans son acception dans le domaine de la fouille de textes. L'opération de classification automatique est au cœur des travaux de fouille de textes. Il s'agit d'une opération distincte de la catégorisation automatique, mais qui lui est complémentaire.

Catégorisation et classification automatiques

Les opérations de catégorisation et de classification automatiques font l'objet de travaux de recherche dans les secteurs de l'intelligence artificielle et de l'apprentissage machine. Elles s'inscrivent dans deux catégories de techniques informatiques. La catégorisation automatique figure dans la catégorie des méthodes dites supervisées ; les méthodes supervisées impliquent une intervention directe de l'utilisateur dans le processus réalisé. Cette intervention prend différentes formes : dans le cas de la catégorisation automatique, l'intervention consiste à guider le processus d'apprentissage en indiquant au système quelles catégories doivent être associées à certains documents. L'ensemble des documents catégorisés manuellement porte le nom d'ensemble d'apprentissage. Sur la base de cet ensemble, l'algorithme de catégorisation effectue un apprentissage (la nature de celui-ci est fonction de la nature de l'algorithme employé) et projette l'information apprise sur un ensemble de test, constitué de documents dont le système ne connaît pas *a priori* les catégories auxquelles ils appartiennent.

Quant à la classification automatique, elle figure traditionnellement dans la catégorie des méthodes dites non supervisées. Les différentes méthodes utilisées ici sont réalisées de manière autonome, sans aucune intervention de l'utilisateur. Ces méthodes non supervisées intègrent, tout comme les méthodes supervisées, un mécanisme d'apprentissage. Cependant, celui-ci opère uniquement sur la base de l'information soumise au système, sans aucun recours à des informations ou métadonnées spécifiées par l'utilisateur. L'algorithme peut nécessiter l'utilisation d'un ensemble d'apprentissage et d'un ensemble de test, mais ce n'est pas obligatoire. L'objectif de l'opération consiste à regrouper les données en ensembles homogènes en employant uniquement les traits caractéristiques ayant servi à décrire chaque élément.

La distinction entre les opérations de catégorisation et de classification se manifeste aussi dans le domaine de la fouille de textes. Ces deux opérations correspondent à deux types d'analyse distincts. L'opération de catégorisation automatique, réalisée en employant certaines métadonnées préalablement connues, est très étroitement reliée à des processus d'analyse de nature prédictive ou normative. Par exemple, une telle opération semble des plus adaptées à l'identification du contenu thématique des documents. L'opération de catégorisation présuppose cependant que nous disposions à priori d'informations valides ou attestées pouvant être employées pour effectuer l'apprentissage. En ce sens, sur le plan théorique, l'opération de catégorisation automatique fait nécessairement appel à des informations externes aux données à traiter. En d'autres termes, la réalisation de cette opération implique l'application d'une structure d'informations sur les données textuelles. Dans une tâche de catégorisation thématique, cette structure externe prend la forme d'une taxinomie de catégories thématiques ou d'un plan de classification.

Par contre, l'opération de classification automatique est effectuée en ne considérant que les données provenant exclusivement des documents textuels soumis au système. L'opération de classification vise à regrouper des documents (ou des segments de documents) en ne considérant que leur contenu. Il s'agit d'une opération exploratoire et descriptive.

Dans son application au traitement automatique des documents, le processus de catégorisation implique plusieurs considérations. Premièrement, le processus de catégorisation présuppose l'élaboration a priori d'une taxinomie ou d'une hiérarchie de catégories thématiques adaptées au contenu et aux spécificités de la collection à traiter. Deuxièmement, ce processus implique non seulement l'identification, à partir de l'ensemble des documents, des caractéristiques (linguistiques et statistiques) qui serviront de base à la catégorisation, mais aussi le choix d'une méthode d'attribution des catégories aux documents.

Pour Charniak (1993), le défi de la classification automatique s'organise autour de trois axes. Le premier axe concerne la description des différents éléments à soumettre à l'opération de classification. En effet, il importe à cet égard de ne retenir que les traits caractéristiques discriminants, sur la base desquels les différents segments de documents seront comparés. Très souvent, cette description implique l'application d'une ou de plusieurs fonctions de filtrage et de nettoyage. Pour que le résultat soit pertinent, il importe que le processus de classification repose sur des descripteurs significatifs. Les décisions prises à cet égard doivent tenir compte aussi bien des objectifs de la classification que de la nature même des objets à classer. Le deuxième axe concerne la fonction discriminante, à la base de toute opération de classification. Cette fonction discriminante peut faire appel à plusieurs critères : l'identité, la similarité, l'homogénéité, l'équivalence, etc. Finalement, le dernier axe concerne le choix de l'algorithme employé pour réaliser

Figure 2
Représentation d'un regroupement plat

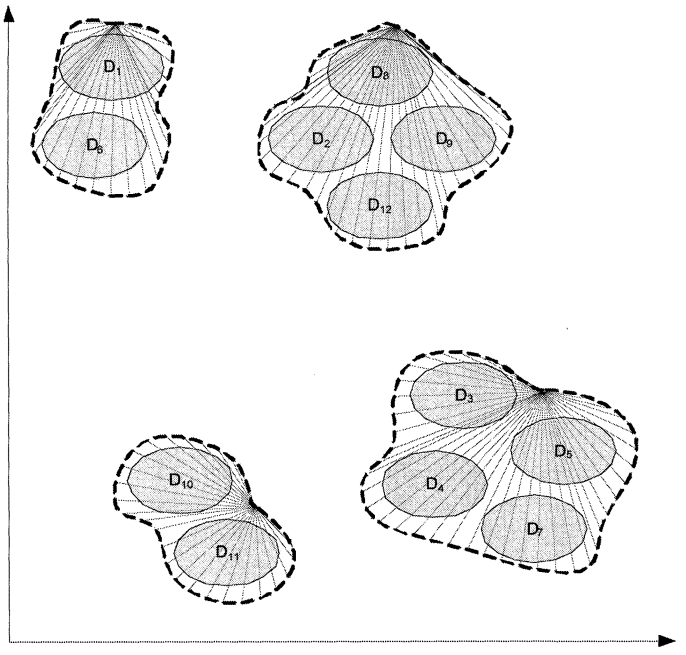
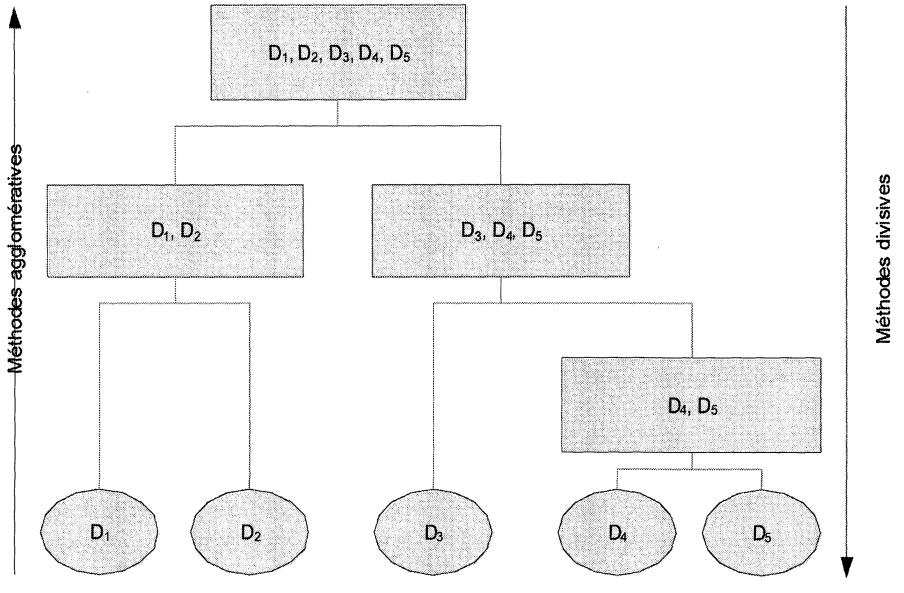


Figure 3
Représentation d'un regroupement hiérarchique



Actuellement, les processus de classification et de catégorisation automatiques ont été les plus exploités dans le contexte de la recherche d'information.

l'opération de classification. Ce choix implique des enjeux à la fois théoriques et pratiques qui relèvent tant de la nature des données à regrouper que des caractéristiques de la classification souhaitée.

Plusieurs techniques informatiques ont été explorées afin d'accomplir des tâches de classification des documents textuels (Jain, Murty and Flynn, 1999). On distingue les différentes méthodes de classification en fonction de la rigidité des regroupements effectués (*hard clustering* vs *soft clustering*) (Manning et Schütze, 1999). Les méthodes de *hard clustering* permettent de positionner un document dans un seul groupe. Par contre, certaines techniques de *soft clustering* permettent de situer le document dans plus d'un ensemble en précisant, pour le document, son degré d'appartenance à chacun de ceux-ci.

Par ailleurs, les techniques de *hard clustering* peuvent aussi faire l'objet de distinctions supplémentaires. Il est possible de distinguer, d'une part, les techniques effectuant des regroupements ou des partitionnements plats (*flat*) et, d'autre part, les techniques effectuant des regroupements hiérarchiques.

Les regroupements plats sont caractérisés par le fait qu'aucune relation précise n'est déterminée entre les différents regroupements générés (figure 2). Les méthodes de cette catégorie sont très souvent de nature itérative. En effet, elles procèdent d'abord en déterminant un nombre fini de regroupements, ensuite en raffinant chacun des regroupements effectués.

Quant aux techniques de classification permettant d'effectuer des regroupements hiérarchiques, elles se caractérisent par la production de nœuds qui représentent chacun une sous-classe d'un nœud de niveau supérieur (figure 3). Les méthodes les plus fréquemment employées pour la classification hiérarchique peuvent être regroupées en deux sous-catégories. La première englobe les méthodes dites « agglomératives » (*bottom-up*) ; celles-ci procèdent d'abord en identifiant tous les éléments à classer, ensuite en effectuant successivement plusieurs regroupements. La seconde sous-catégorie englobe les méthodes dites « divisives » (*top-down*) ; celles-ci procèdent d'abord en identifiant un seul regroupement, ensuite en divisant ou en fractionnant le groupe initial.

Application à la recherche d'information

Dans le domaine de la fouille de textes, les récents développements techniques (concernant, entre autres, le développement d'algorithmes d'extraction et d'organisation automatiques d'information) et méthodologiques (desquels relève principalement la modélisation des démarches à suivre afin d'appliquer efficacement les techniques) ont récemment été explorés au sein d'applications de recherche d'information en ligne. D'ailleurs, depuis quelques années, plusieurs prototypes d'applications industrielles d'analyse et de recherche d'information ont été proposés afin d'intégrer certaines techniques de fouille de textes davantage sensibles à la dimension sémantique des documents.

Actuellement, les processus de classification et de catégorisation automatiques ont été les plus exploités dans le contexte de la recherche d'information. Ces deux processus, bien qu'ils puissent être exploités afin d'assister différentes tâches liées à l'analyse de l'information numérique, sont principalement utilisés dans un contexte de recherche d'information afin de lever l'ambigüité sémantique des termes de requête. Cette perspective d'application repose en partie sur l'hypothèse suivante : les mots avec lesquels les termes de la requête cooccurrent sont de bons indices pour identifier la signification précise des termes de la requête en question. Cette hypothèse est d'ailleurs à la base même de l'utilisation du modèle vectoriel pour la recherche d'information (Salton et McGill, 1983). Lors de l'utilisation de stratégies d'extraction et d'organisation automatiques d'information, l'objectif est d'exploiter au maximum cette idée en procédant, en plus, au regroupement automatique des documents. Comme nous l'avons souligné, le regroupement automatique des documents peut être accompli en utilisant soit un algorithme de classification, soit un algorithme de catégorisation. Dans les deux cas, au terme du processus, les documents sont regroupés sur la base de thématiques communes, permettant ainsi à l'utilisateur de porter son attention sur les documents inscrits dans des regroupements thématiques qui comblent plus adéquatement son besoin d'information.

Les principales applications de recherche d'information en ligne fondées sur l'utilisation de la classification ou de la catégorisation automatiques des documents sont *Grokker*[10] et *Clusty*[11]. Ces deux applications possèdent plusieurs caractéristiques communes, mais se distinguent l'une de l'autre à plusieurs égards.

La différence importante entre ces deux outils est qu'ils ne sont pas actuellement applicables sur des corpus de même type. *Clusty* interroge la partie visible du Web en entier, alors que *Grokker* n'offre actuellement que la

10. http://www.grokker.com
11. http://www.clusty.com

possibilité d'interroger le contenu des index de *Yahoo !*, de *Wikipedia* et de *Amazon*.

Abordons maintenant les similarités entre *Grokker* et *Clusty*. La principale fonctionnalité de ces deux outils de recherche d'information intégrant des opérations de fouille de textes est la classification automatique des résultats de recherche. Grâce à cette fonctionnalité, il est possible de regrouper les nombreux résultats d'une requête en fonction du lexique commun à certains documents. Ainsi, les documents partageant plusieurs mots en commun sont regroupés afin de former des classes de documents formant des univers lexicaux homogènes. Dans cette perspective, le regroupement des résultats de recherche permet, entre autres, d'assister l'identification d'ambiguïtés sémantiques qui sont très souvent une source importante de bruit dans le processus de recherche d'information.

Afin de regrouper les documents partageant de fortes similarités thématiques, les outils de recherche d'information peuvent utiliser soit des algorithmes de classification, soit des algorithmes de catégorisation. Le choix entre ces deux familles d'algorithmes dépendra de la qualité et de la quantité de métadonnées connues à propos de certains documents. Ainsi, plus il y aura de métadonnées fiables disponibles au sein du corpus ou de la base interrogée, plus il sera possible de recourir à un algorithme de catégorisation. En contrepartie, lorsque peu de métadonnées sont associées aux documents présents dans le corpus, il est recommandé de plutôt utiliser un algorithme de classification non supervisée. Un tel algorithme tentera de faire émerger automatiquement des structures thématiques fondées sur la présence d'univers lexicaux, lesquelles permettront à l'utilisateur d'orienter son parcours de découverte des documents récupérés. Pour cette raison, il est plus fréquent de recourir à un algorithme de classification automatique, notamment lorsqu'il s'agit d'interroger le Web. Par contre, il est tout à fait possible d'appliquer un algorithme de catégorisation à l'intérieur d'univers documentaires clos bien documentés par des métadonnées et de moins grande ampleur (à l'intérieur d'un Intranet corporatif, par exemple).

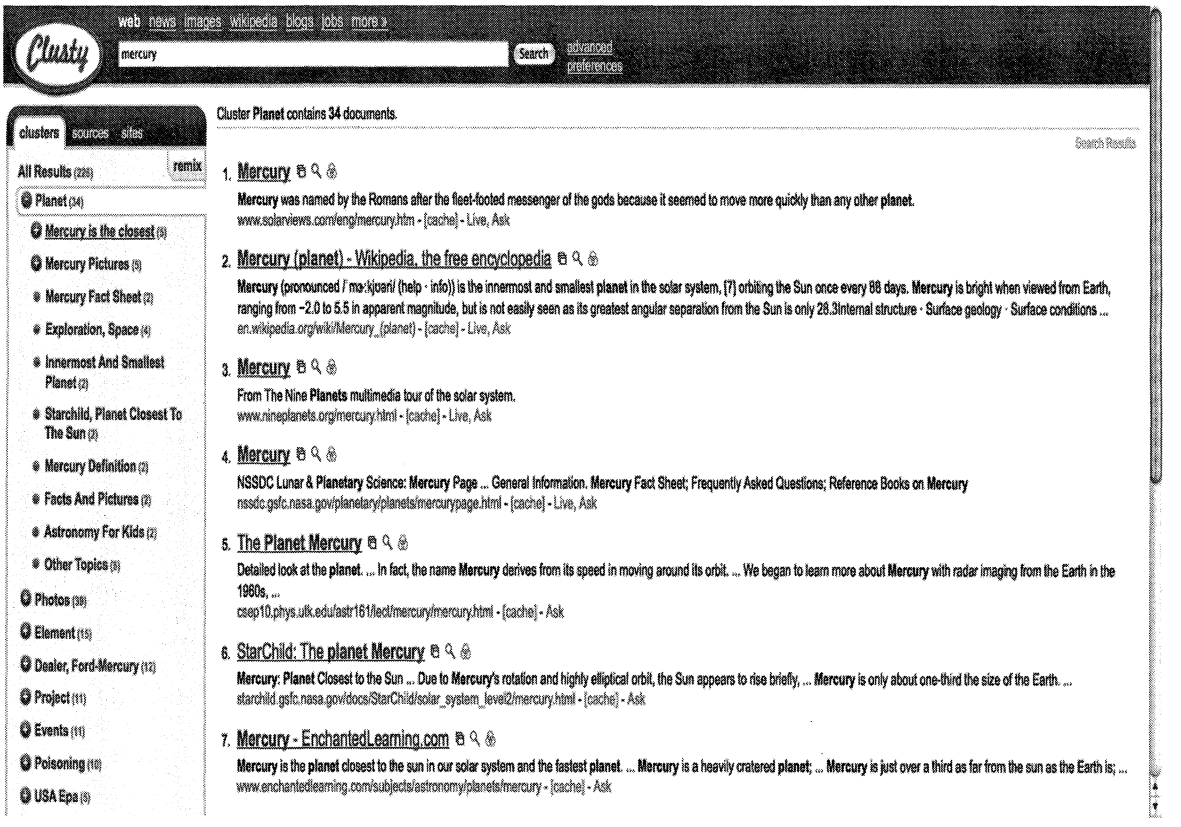
Les nouveaux outils d'analyse et de recherche d'information en ligne qui reposent sur des opérations de classification offrent un ensemble de fonctionnalités et de paramètres de base avec lesquels l'utilisateur peut interagir. Les figures 4 et 5 représentent les interfaces de recherche des deux moteurs mentionnés précédemment. Ces deux moteurs offrent des interfaces de recherche possédant de nombreuses caractéristiques communes. Ainsi, on constate que les deux moteurs fondés sur la classification automatique des documents numériques présentent, à l'instar des moteurs de recherche classiques, les résultats des requêtes sous forme de liste (à la droite de l'interface). Par défaut, les résultats sont présentés selon un tri de pertinence. Ce qui distingue les moteurs de recherche traditionnels des outils de recherche de

Ce qui distingue les moteurs de recherche traditionnels des outils de recherche de nouvelle génération repose en grande partie dans la liste des regroupements de documents.

nouvelle génération repose en grande partie dans la liste des regroupements de documents présentée dans la partie gauche de l'interface. On y retrouve l'inventaire des regroupements hiérarchiques générés par l'algorithme de classification. Ainsi, tous les documents récupérés par le moteur de recherche sont d'abord regroupés selon des mots discriminants qu'ils partagent avant d'être affichés dans la composante de droite de l'interface. Comme nous l'avons expliqué précédemment, l'opération de classification automatique consiste à regrouper des documents partageant certains critères de similarité, lesquels permettent d'identifier des thématiques ou des univers lexicaux caractéristiques des documents. Ces informations extraites par l'algorithme de classification s'avèrent utiles pour assister l'utilisateur dans l'identification des documents lui permettant de mieux répondre à son besoin d'information. Cependant, *stricto sensu*, l'opération de classification n'implique pas une composante permettant de caractériser le contenu des classes ou des regroupements ainsi générés. C'est à cet égard qu'il est préférable de réaliser une opération de catégorisation lorsque les documents nous permettent de le faire, c'est-à-dire lorsque les documents de la base interrogée sont accompagnés de métadonnées permettant d'effectuer adéquatement l'opération de catégorisation. Une solution de rechange s'impose lorsqu'aucune métadonnée n'est disponible, ou lorsque les métadonnées sont disponibles en quantité insuffisante ou encore lorsque leur qualité ne peut être attestée – ce qui est précisément le cas des documents sur le Web, du moins dans son état actuel. Il est alors possible d'identifier la thématique du contenu des regroupements issus de l'étape de classification en procédant à l'extraction automatique des termes les plus discriminants de chaque regroupement. Dans nos travaux antérieurs (Forest, 2006), nous avons démontré la pertinence de cette démarche en utilisant la mesure de pondération TF-IDF[12] comme facteur de discrimination des termes de chaque regroupement. Il est cependant possible d'employer et de combiner plusieurs mesures statistiques afin d'extraire

12. La mesure de pondération tf · idf (« *term frequency · inverse document frequency* ») (Salton, 1989) permet de modérer ou d'accentuer l'importance de la fréquence de chaque mot à l'intérieur d'un corpus. Le principe de cette mesure peut être formulé de la manière suivante : un mot sera d'autant plus efficace pour représenter le contenu d'un document (ou d'une classe de documents) s'il est à la fois fréquent dans ce document (ou dans cette classe de documents) et rare dans l'ensemble des autres documents (ou des autres classes de documents) à analyser.

Figure 4
Interface de recherche du moteur *Clusty*



les principales catégories thématiques présentes dans les documents préalablement regroupés.

L'exemple suivant illustre bien la pertinence de recourir à des procédés de classification et de catégorisation automatiques des documents pour assister la recherche d'information. En utilisant le terme « *mercury* » comme requête, les deux outils de recherche offrent à l'utilisateur la possibilité de parcourir les documents récupérés, tout en lui indiquant qu'il s'agit d'un terme très polysémique. En effet, le terme anglais « *mercury* » peut être utilisé pour désigner tantôt un dieu romain, tantôt une planète, tantôt un élément chimique, ou encore une marque de voiture, et même d'autres types de moyen de transport, plusieurs lieux géographiques, un chanteur, etc[13]. Compte tenu de la polysémie du terme « *mercury* », il n'est donc pas surprenant qu'une majorité de documents récupérés par les moteurs de recherche traditionnels sur le Web présentent des informations non pertinentes pour les utilisateurs. Ce problème pourrait être corrigé, en partie, en spécifiant la requête. Mais cette solution, très imparfaite dans les faits, implique que l'utilisateur sache *a priori* quelle information il recherche et quels sont les meilleurs termes pour y accéder. Or, en pratique, les utilisateurs parcourent (*browse*) le Web sans y chercher des documents précis. De plus, ils sont souvent peu compétents dans la formulation des requêtes. Pour ces raisons, la classification et la catégorisation automatiques des documents s'avèrent des plus fécondes pour assister la découverte d'information, car elles permettent de mettre en lumière certaines dimensions thématiques potentiellement inconnues de l'utilisateur, sans toutefois lui imposer de faire des choix inutiles.

Dans notre exemple, il n'est pas étonnant de constater le nombre élevé de regroupements proposés tant par *Clusty* que par *Grokker*. Ainsi, pour le moteur de recherche *Grokker*, on note les principaux regroupements suivants (la valeur entre parenthèses correspond au nombre de documents dans chaque regroupement) : *Planet Mercury* (39), *General* (28), *Mercury Information* (17), *Solar System* (15), *Project Mercury* (12), *Mercury in Fish* (9), *Mercury Element* (8), *Mercury Levels* (8), *Sun* (8), etc.

Voici les résultats du moteur de recherche *Clusty* (la valeur entre parenthèses correspond ici aussi au nombre de documents dans chaque regroupement) : *Planet* (34), *Photos* (39), *Element* (15), *Dealer Ford-Mercury* (12), *Project* (11), *Events* (11), *Poisoning* (10), etc.

Dans les deux applications, il importe de mentionner que les regroupements sont effectués par un algorithme

13. On retrouve dans la version anglaise de *Wikipédia* plus d'une trentaine d'entrées associés au terme « *mercury* ».

Figure 5
Interface de recherche du moteur *Grokker* (sans visualisation graphique)

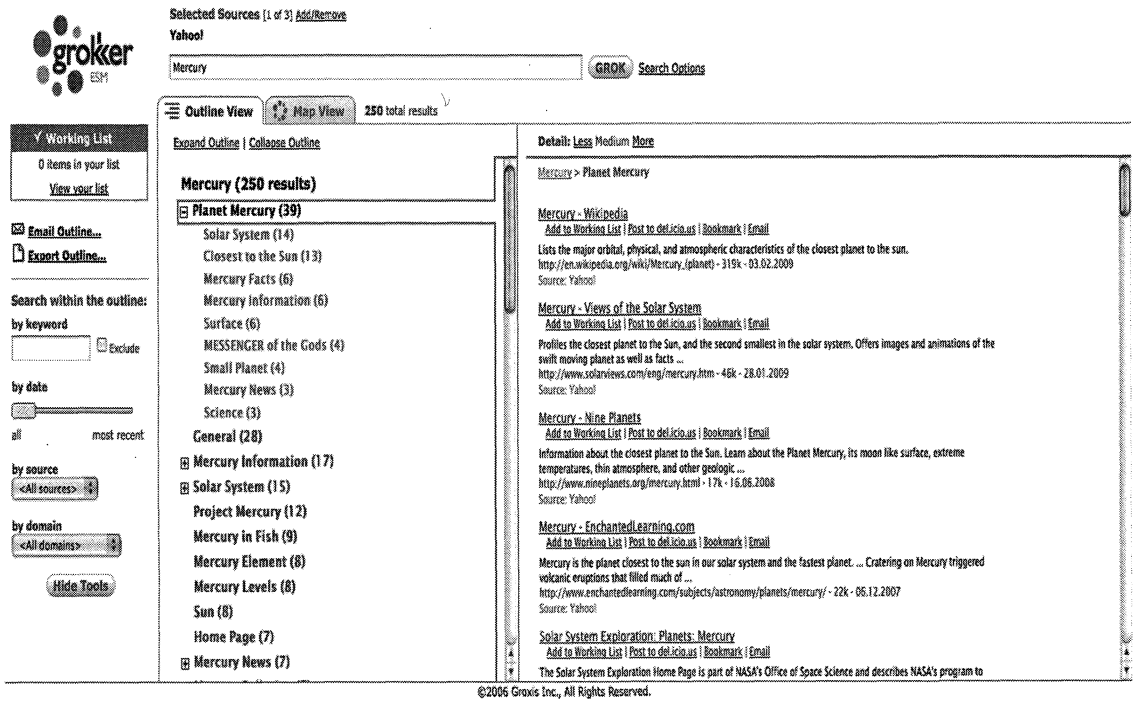
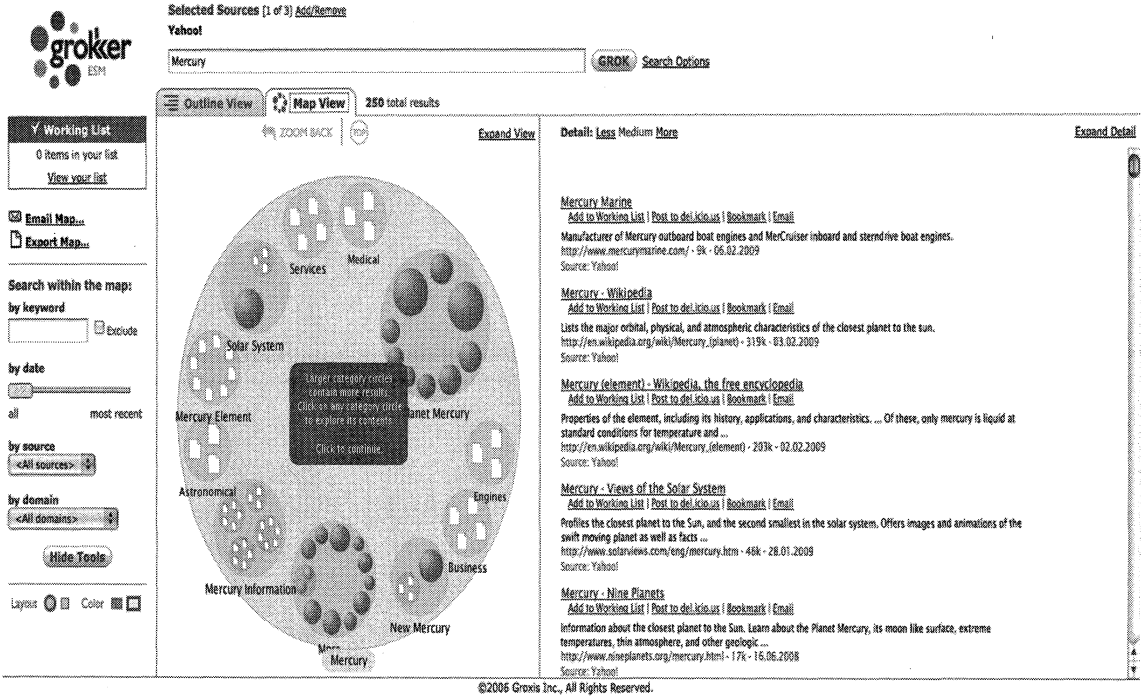


Figure 6
Interface de recherche du moteur *Grokker* (avec représentation graphique de la même structure présente dans la figure 5)



Une partie importante de l'inefficacité des moteurs de recherche en ligne actuellement disponibles découle du paradigme de recherche duquel nous semblons être prisonniers.

de classification hiérarchique. Il est donc possible d'identifier des contextes d'utilisation ou des significations plus précises au sein de regroupements plus généraux (ce qui permet, par exemple, de distinguer les images de la planète Mercure des différentes informations scientifiques à son sujet).

Malgré les nombreux éléments qu'ont en commun ces deux moteurs de recherche, leurs différences mettent en lumière deux composantes essentielles de ces approches qui transcendent ce que nous avons mentionné. Le premier élément à observer est la différence importante entre ces deux applications, en ce qui a trait aux étiquettes employées pour représenter les regroupements effectués. Il s'agit d'une facette fondamentale de l'approche que nous évoquons. Les étiquettes utilisées doivent représenter adéquatement le contenu des regroupements. Or, dans un contexte de classification automatique, ces étiquettes sont extraites automatiquement en pondérant certains mots du lexique issus des regroupements effectués. L'extraction automatique des étiquettes catégorielles représente un défi majeur et un territoire de recherche des plus actifs. Comme nous sommes en mesure de le constater, surtout dans l'application *Grokker*, il est fréquent pour ces outils d'extraire des étiquettes redondantes (reflétant souvent une classification imparfaite) ou d'utiliser des étiquettes non significatives provenant de suites de mots fréquents dans les documents (et donc, par conséquent, non pertinente pour l'utilisateur). Par exemple, dans *Clusty*, l'étiquette « *Mercury is the closet* » a été, à tort, automatiquement extraite pour représenter un regroupement de deuxième niveau. Pour cette raison, nous défendons l'hypothèse selon laquelle le succès ou l'échec de ces nouvelles approches résident en partie dans la qualité de la classification effectuée, mais aussi dans l'application d'algorithmes performants permettant d'extraire les termes les plus significatifs pour représenter adéquatement le contenu des regroupements effectués (en l'absence de métadonnées pouvant être exploitées par des algorithmes de catégorisation).

Le second élément à observer, comme le montre la figure 6, est la possibilité de coupler les approches classificatoires en contexte de recherche d'information à des stratégies de visualisation de l'information. En effet, les résultats des opérations de fouille de textes sont souvent facilement représentables graphiquement. Certaines techniques de visualisation permettent de représenter de manière synthétique d'importantes quantités d'informations. Les recherches sur le Web permettent souvent de récupérer plusieurs centaines de milliers de documents. Bien que peu de ces documents soient pertinents pour l'utilisateur, la manière dont ils sont actuellement présentés, sous forme de listes non hierarchiques triées par pertinence, ne permet pas de dégager les principales dimensions de l'information présente dans les documents. Les processus de fouille de textes cherchent à extraire des structures d'information présentes dans des données textuelles. Cela est mis en évidence dans l'application des algorithmes de classification hiérarchique. L'accès aux informations présentes dans la structure classificatoire peut être facilité par l'utilisation de modalités de visualisation qui les synthétisent et les mettent en exergue. Il n'est donc pas surprenant que certaines applications de recherche d'information ayant recours à des processus de fouille de textes mettent à profit les avantages de certains processus de visualisation de l'information. Dans cette perspective, en considérant l'augmentation croissante de l'information disponible en format numérique, nous défendons aussi l'hypothèse selon laquelle l'adaption de certaines fonctionnalités de fouille de textes au sein des plates-formes de recherche d'information reposera, dans les prochaines années, sur le dévelopement de modalités de visualisation synthétiques qui permettront de mettre en relief certaines dimensions que la présentation sous forme de liste n'est pas en mesure de représenter.

Conclusion

Depuis l'avènement du Web, la quantité de documents disponibles en format numérique n'a cessé de croître. Les retombées de ce phénomène ont été des plus positives pour la diffusion et la circulation de l'information. L'omniprésence du document numérique a, en outre, justifié le développement de nombreux outils technologiques cherchant à assister l'analyse et la gestion de l'information. Cependant, malgré les développements technologiques, un problème demeure, voire s'accentue. Le volume d'information disponible fait en sorte qu'il est de plus en plus difficile de retrouver celle dont nous avons besoin. Il s'agit d'un problème très complexe dont les causes sont multiples et ramifiées. Plusieurs facteurs peuvent donc expliquer cet état de fait. À cet égard, nous sommes d'avis qu'une partie importante de l'inefficacité des moteurs de recherche en ligne actuellement disponibles découle du paradigme de recherche duquel nous semblons être prisonniers.

Nous avons exposé dans cet article comment il est possible d'exploiter des opérations de fouille de textes afin de permettre aux moteurs de recherche d'accéder à certaines dimensions sémantiques des documents. Nous avons d'ailleurs illustré notre propos à l'aide d'un

exemple impliquant les moteurs de nouvelle génération, *Clusty* et *Grokker*. Il est désormais possible d'appliquer certains processus de fouille sur les très nombreux documents disponibles en format numérique afin d'en extraire des patrons qui en sont caractéristiques. Les retombées de ces approches sont multiples. Dans le domaine de la recherche d'information, il est possible, grâce à ces stratégies, d'assister l'identification d'univers thématiques ou sémantiques permettant d'orienter plus précisément le parcours de découverte des utilisateurs du Web. Dans cet article, nous avons cependant consacré l'essentiel de nos efforts à démontrer la pertinence de l'opération de classification automatique, et, dans une moindre mesure, de l'opération de catégorisation automatique des documents. Le domaine de la fouille de textes, dont l'objectif est d'assister l'extraction et l'organisation automatique d'information, est composé de plusieurs autres processus dont pourrait grandement bénéficier la recherche d'information. Comme le souligne clairement Feldman, les développements technologiques en lien avec le domaine de la documentation ne peuvent ralentir le phénomène de l'infobésité, phénomène dont la croissance ne donne aucun signe d'essoufflement : « *Web searchers are literally buried in information today.* [...] *they are crying out for tools that will tell them what they need to pay attention to in the pile. How we will manage to apply automatic assessment and weighting factors to information to find the good stuff is certainly a research topic that will keep doctoral students going for a long time.* » (2008, p. 10)

Les prochaines années seront caractérisées par une augmentation des contenus générés dynamiquement (pensons, à titre d'exemple, aux informations disponibles en continu dans les fils RSS). Ce phénomène sera aussi accompagné d'une augmentation des contenus générés grâce à une participation plus active des utilisateurs à ce qu'il convient désormais de nommer le Web 2.0. Il s'agit là de deux facteurs parmi d'autres qui nous portent à croire que la quantité d'information numérique ne fera qu'augmenter et qu'il est désormais essentiel d'explorer de nouvelles avenues pour en faciliter la recherche, l'analyse et la gestion. ⊙

Les développements technologiques en lien avec le domaine de la documentation ne peuvent ralentir le phénomène de l'infobésité, phénomène dont la croissance ne donne aucun signe d'essoufflement.

Sources consultées

Charniak, E. 1993. *Statistical language learning.* Cambridge, Mass. : MIT Press.

Fayyad, U.M., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (Eds.). 1996. *Advances in knowledge discovery and data mining.* Cambridge, Mass. : MIT Press.

Feldman, S. 2004. The high cost of not finding information. *KMworld*, 13, 3, 8-10.

Feldman, S. 2008. What are people searching for and where are they looking ? *KMworld*, 17, 3, 8-9, 30.

Forest, D. 2006. *Application de techniques de forage de textes de nature prédictive et exploratoire à des fins de gestion et d'analyse thématique de documents textuels non structurés.* Thèse de doctorat, Montréal, Université du Québec à Montréal.

Hearst, M. 2003. *What is text mining* ? Document non-publié disponible à l'adresse www.ischool.berkeley.edu/hearst/text-mining.html.

Jain, A.K., M.N. Murty et P. J. Flynn. 1999. Data clustering : a review. *ACM computing surveys*, 31, 3, 264-323.

Manning, C.D., and H. Schütze. 1999. *Foundations of statistical natural language processing.* Cambridge, Mass. : MIT Press.

Salton, G. 1989. *Automatic Text Processing.* Reading, Mass. : Addison-Wesley.

Salton, G. et McGill, M. 1983. *Introduction to Modern Information Retrieval.* New-York : McGraw-Hill.