Documentation et bibliothèques

DOCUMENTATION BIBLIOTHÈQUES

La recherche d'information multilingue
Retrieving Multilingual Information
Búsqueda de información multilingüe

Élaine Ménard

Volume 52, Number 4, October–December 2006

URI: https://id.erudit.org/iderudit/1029339ar
DOI: https://doi.org/10.7202/1029339ar

See table of contents

Publisher(s)

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

ISSN

0315-2340 (print)
2291-8949 (digital)

Explore this journal

Cite this article

Ménard, É. (2006). La recherche d'information multilingue. *Documentation et bibliothèques*, *52*(4), 255–261. https://doi.org/10.7202/1029339ar

Article abstract

Amplified by the Web, the information explosion creates several problems associated with the retrieval of information: very large collections of documents that are dynamic and evolving an abundance of research data, multimedia data, interfaces between users and systems and, particularly, multilingual information. This article summarises the retrieval of multilingual information whereby a person can retrieve a document using a language that is different from the language of the document retrieved. This article also presents an analysis of the retrieval of fixed digital images in multilingual collections.

DOCUMENTATION BIBLIOTHÈQUES

La recherche d'information multilingue

ÉLAINE MÉNARD
Candidate au doctorat
École de bibliothéconomie et des sciences de l'information
Université de Montréal
elaine.menard@umontreal.ca

RÉSUMÉ | ABSTRACTS | RESUMEN

L'explosion de l'information, amplifiée par le développement du Web, génère de nombreux problèmes au moment de la recherche d'information: collections de documents gigantesques, dynamiques et changeantes, surabondance des résultats de recherche, données multimédias, interactions entre utilisateurs et systèmes, et plus particulièrement multilinguisme de l'information. Cet article trace un bilan de la recherche d'information multilingue (RIML), un type de recherche qui permet à un chercheur de repérer un document dans une langue différente de celle de sa requête. Cet article présente également une réflexion sur le repérage de l'image numérique fixe dans le contexte des collections multilingues.

Retrieving Multilingual Information

Amplified by the Web, the information explosion creates several problems associated with the retrieval of information: very large collections of documents that are dynamic and evolving, an abundance of research data, multimedia data, interfaces between users and systems and, particularly, multilingual information. This article summarises the retrieval of multilingual information whereby a person can retrieve a document using a language that is different from the language of the document retrieved. This article also presents an analysis of the retrieval of fixed digital images in multilingual collections.

Búsqueda de información multilingüe

La explosión de la información, extendida por el desarrollo de la Web, genera numerosos problemas cuando se realiza la búsqueda de información: compilaciones de documentos gigantescos, dinámicos y cambiantes, superabundancia de resultados de la búsqueda, datos de multimedia, interacción entre usuarios y sistemas y, especialmente, el multilingüismo de la información. Este artículo esboza un balance sobre la búsqueda de información multilingüe (RIML), aquella que permite a un investigador localizar un documento aún cuando el idioma de su solicitud no es el mismo que el de los documentos que busca. Este artículo presenta, asimismo, una reflexión sobre la situación de la imagen numérica fija en el contexto de las compilaciones multilingües.

Depuis quelques années, Internet est devenu un média incontournable pour la diffusion de ressources multilingues. On considère qu'il existe présentement environ 6 900 langues vivantes à travers le monde (Ethnologue, 2006). Même s'il est difficile d'estimer le nombre exact de langues écrites parmi celles-ci, en raison principalement du manque de sources d'information fiables et disponibles (Robinson et Gadelii, 2003), on peut supposer que plusieurs généreront éventuellement des documents sous forme textuelle ou autre.

Le Web constitue un vaste univers de connaissances et de cultures humaines diverses, permettant le partage d'idées et d'informations, et cela sans frontières. Toutefois, la performance des divers systèmes de recherche varie considérablement au moment de repérer des documents rédigés en une ou plusieurs langues différentes de celle de la requête (Chaudiron, 2002). Les différences linguistiques constituent souvent un obstacle majeur aux échanges de documents scientifiques, culturels, pédagogiques et commerciaux. En outre, la recherche d'informations sur Internet est confrontée au problème de la surabondance des résultats. Ce problème, loin de s'amenuiser, prend de l'ampleur avec l'accroissement du Web et l'émergence d'une grande variété de langues dans ce dernier. L'accès à ce foisonnement d'information multilingue est devenu un défi de taille.

Utilité et importance du multilinguisme

Multilinguisme et information multilingue

Avant tout, il convient de définir le concept d'information multilingue, puisque celui-ci est souvent mal interprété dans la littérature. Ainsi, un document (textuel ou autre) contenant plus d'une langue est considéré comme multilingue. On peut toutefois élargir la définition d'information multilingue en termes de collection, dans laquelle on retrouve des documents unilingues de différentes langues ou des documents comprenant eux-mêmes plusieurs langues (Oard et Dorr, 1996).

En 2001, dans sa déclaration universelle sur la diversité culturelle, l'UNESCO s'engageait, entre autres, à *« promouvoir la diversité linguistique dans l'espace numérique et encourager l'accès universel, à travers les réseaux mondiaux, à toutes les informations qui relèvent du domaine public »* (UNESCO, 2003 : 57). Cet engagement ferme de l'UNESCO laisse déjà entrevoir l'intérêt croissant pour l'information multilingue.

Des statistiques récentes montrent que le Web est largement dominé par les langues que l'on retrouve dans les pays développés. Soulignons également que 16 langues se retrouvent dans 90,3 % des pages Web recensées en 2003, alors que 22 autres langues ne se retrouvent que dans 0,9 % des pages Web. En outre, si la langue anglaise arrive toujours en tête avec 57,4 % des pages Web répertoriées, cette proportion n'a cessé de diminuer depuis les sept dernières années ; en 1999, on rapportait 72 % des pages Web en anglais (Gey, Kando et Peters, 2005 : 427). Dans le même ordre d'idées, on observe que les utilisateurs anglophones dominent toujours le Web avec une proportion de 30,6 % (Internet World Stats, 2006), ce qui constitue toutefois un recul marqué par rapport à cette même statistique (35,9 %) publiée en septembre 2004. Deux constatations émanent de ces données. D'une part, les langues utilisées dans les pages Web sont de plus en plus diversifiées et, d'autre part, les internautes n'utilisent plus uniquement l'anglais sur le Web, comme c'était le cas il y a quelques années (Grefenstette, 2002 : 60-61). On observe donc une évolution progressive vers les environnements multilingues d'information.

Rôle et importance de l'information multilingue

Avec l'implantation massive de nouvelles technologies informatiques, la mondialisation de l'économie et la percée sans précédent du commerce électronique, les gens se voient confrontés à une augmentation fulgurante de l'information mise à leur disposition. L'accès à l'information, peu importe la langue, représente une richesse extraordinaire puisqu'elle contient *« tout le réservoir "d'idées", alimenté au fil du temps par le patrimoine, les traditions et les habitudes locales communiquées par les langues locales »* (UNESCO, 2004).

Cependant, même si l'information multilingue existe, cela ne signifie pas pour autant que le besoin d'accéder à celle-ci se fait réellement sentir. Pourtant, se priver de l'information multilingue signifie également renoncer à toutes les idées véhiculées par différentes langues et différentes cultures, ce qui représente une grande richesse pour l'individu. On peut mentionner, à titre d'exemple, la valeur extraordinaire de l'information scientifique, culturelle et pédagogique des nombreuses collections numérisées des services patrimoniaux, des musées, des bibliothèques et des archives de différentes communautés linguistiques.

Il faut rappeler que l'information multilingue est tout aussi capitale pour les entreprises qui désirent demeurer compétitives, surtout avec l'émergence du phénomène de mondialisation qui amplifie la concurrence. En effet, il est de plus en plus courant de voir une entreprise mener une partie de ses opérations dans une langue distincte de celle avec laquelle elle s'est développée, afin de traiter avec des clients potentiels, des employés, des partenaires commerciaux ou des entreprises concurrentes d'autres pays. Par exemple, les acteurs du commerce électronique reconnaissent progressivement la nécessité d'exploiter la langue et les traditions culturelles spécifiques des marchés visés (Hillier, 2003 : 12).

On constate donc un réel besoin de prévoir des mécanismes de repérage efficaces, c'est-à-dire des outils permettant l'accès facile, rapide et fiable aux collections multilingues, afin que les individus puissent en profiter. L'accès à l'information multilingue doit être considéré comme un enjeu crucial, autant pour l'individu que pour les entreprises. Toutefois, le processus de repérage de l'information dans les collections multilingues s'avère souvent fort complexe. Les systèmes de recherche d'information multilingue (RIML) tentent donc d'apporter une solution à ce problème de diversité linguistique de l'information.

Recherche d'information multilingue

Caractéristiques de la recherche d'information multilingue

L'intérêt pour l'information multilingue n'est pas récent. En 1960, Calvin N. Mooers, un des pionniers en matière de recherche d'information, envisageait déjà la création d'un système qui permettrait la traduction automatique de textes et également la production sur demande d'essais écrits touchant certains sujets précis (Salton, 1987 : 375-380). Ce chercheur avait déjà saisi l'importance de concevoir un système capable de traiter et de repérer l'information multilingue. Cependant, plus de 40 ans plus tard, ces systèmes ne sont toujours pas disponibles à grande échelle.

Sur la lancée des campagnes d'évaluation du *Text REtrieval Conference* (TREC), la RIML a été abondamment étudiée. La RIML est un type de recherche permettant de repérer des documents en langue différente de celle de la requête (Oard et Diekema, 1998 : 223). Un individu peut ainsi présenter une requête dans sa propre langue et le système repère des documents dans une ou plusieurs autres langues.

La RIML est utile dans de nombreuses situations, lorsque : la collection regroupe des documents

en plusieurs langues; les documents eux-mêmes sont écrits en plusieurs langues; l'individu ne connaît pas suffisamment la langue d'un document, mais veut quand même obtenir ce document; la collection est indexée dans une langue non familière à l'individu; un chercheur veut connaître tout ce qui a été écrit sur un sujet précis, peu importe la langue; et enfin, l'individu possède les ressources nécessaires pour traduire un document dans une langue qu'il comprend (Oard et Dorr, 1996).

Deux approches sont habituellement proposées en RIML. La première consiste à traduire les documents au complet dans toutes les langues, alors que la deuxième suppose la traduction des requêtes dans la langue des documents à repérer (Oard et Ertune, 2002). Ces deux approches font généralement appel à trois types de ressources linguistiques pour la traduction: les dictionnaires bilingues ou multilingues (Hedlund *et al.*, 2004), les systèmes de traduction automatique (TA) (Chen et Gey, 2004), et les corpus parallèles ou comparables (Braschler et Schäuble, 2000). Ces différentes ressources linguistiques ont chacune démontré leurs forces et leurs faiblesses. Mais il semblerait que la démarche la plus prometteuse consiste à combiner plusieurs ressources différentes pour la traduction (Braschler, 2004: 184).

Par ailleurs, la RIML a fait l'objet de nombreux travaux de recherche depuis quelques années. Ces travaux mettent en relief de nombreux problèmes. Ainsi, plusieurs chercheurs considèrent que peu importe la ressource linguistique utilisée, le principal problème associé à la RIML demeure l'ambiguïté sémantique et syntaxique découlant de la traduction (Kishida, 2005: 439). Parmi les autres problèmes reliés à la RIML, mentionnons la manipulation des mots composés, des noms propres et du vocabulaire émergent ou le traitement de certains domaines spécifiques (par exemple, le domaine médical) (Braschler, 2004: 187-189). En outre, la fusion des résultats provenant de différentes collections multilingues demeure une question préoccupante en RIML (Nie, 2002).

Systèmes d'information et recherche d'information multilingue

Malgré les difficultés associées à la RIML, le potentiel éminent de ce type d'approche suggère que de nombreux systèmes pourraient en tirer avantage. Déjà, plusieurs moteurs de recherche disponibles sur le Web (par exemple, Google ou AltaVista) offrent la recherche monolingue dans une variété de langues. Ces moteurs permettent ensuite de traduire les documents obtenus en d'autres langues. Cependant, cette utilisation de la RIML est limitée à quelques cas d'exception et les systèmes commerciaux d'information n'offrent toujours pas de service de RIML à leurs utilisateurs (Gey, Kando et Peters, 2005: 427). Pourtant, les moteurs de recherche qui adopteraient l'approche de la RIML offriraient du même coup à leurs utilisateurs des mécanismes pour formuler, raffiner, désambiguïser les requêtes, filtrer les résultats de recherche et obtenir des documents dans la langue désirée.

Dans un autre ordre d'idées, les systèmes de question-réponse (QR) profiteraient également des avantages offerts par les algorithmes de RIML. On définit un système de QR comme un outil utilisé pour trouver une réponse courte et précise à une question posée en langue naturelle. Le fonctionnement des systèmes de QR se fait en trois étapes: analyse de la question posée en langue naturelle, traitement des documents et extraction de la réponse. L'intégration d'un algorithme de RIML aux systèmes de QR permettrait à l'individu de poser une question dans une langue qu'il connaît et d'obtenir une réponse dans cette même langue, même si la réponse est extraite d'un document rédigé dans une autre langue. Idéalement, l'implantation de la RIML peut se faire sans modifier le système de QR lui-même, par le simple ajout d'un module de traduction des questions et des réponses (Plamondon et Kosseim, 2003). L'incorporation des algorithmes de RIML aux systèmes de QR est présentement à l'étude, notamment dans le cadre d'une des pistes de recherche du *Cross Language Evaluation Forum* (CLEF). Pour le moment, les systèmes de QR bilingues sont beaucoup moins performants que les systèmes de QR monolingues, mais les chercheurs se disent encouragés par les résultats obtenus jusqu'à maintenant (Vallin *et al.*, 2005).

D'autres systèmes tireraient également avantage de l'ajout d'algorithmes de RIML. C'est notamment le cas des agents intelligents que l'on définit comme des logiciels qui permettent de chercher et de traiter l'information, et dont le fonctionnement repose sur un principe d'automatisation des tâches (Bouis, 1999). En règle générale, on distingue deux types d'agents intelligents: les agents d'interface ayant comme fonction d'assister les usagers au cours de leurs activités sur le Web et les agents d'information ayant pour objectifs de trouver, d'analyser et de repérer une grande quantité d'informations (Detlor et Arsenault, 2002: 404). L'introduction d'un algorithme de RIML, jumelé au potentiel des agents intelligents, améliorerait considérablement la recherche d'information, l'aide à l'usager, la veille et l'aide au commerce électronique.

L'intégration des algorithmes de RIML s'avère également intéressante pour des systèmes axés sur le repérage d'informations géographiques, la recherche de documents multimédias (images fixes, images en mouvement et fichiers sonores), la production de résumés, etc. Toutefois, ces applications représentent toujours de véritables défis pour les algorithmes de RIML.

Raisons du retard de l'implantation des algorithmes de RIML

Comme nous pouvons le constater, la RIML apparaît comme un complément idéal à plusieurs systèmes d'information. Toutefois, le déploiement des algorithmes de RIML à grande échelle continue d'accuser un retard important. Cinq raisons majeures expliquent le délai de mise en place des algorithmes de RIML dans les systèmes d'information.

Premièrement, la plupart des algorithmes de RIML reposent sur un mécanisme de traduction automatique qui présente toujours certaines faiblesses. D'une part, la traduction automatique n'est pas nécessairement disponible pour toutes les paires de langues (Braschler, 2004 : 187) et, d'autre part, elle n'offre pas, jusqu'à maintenant, la qualité de travail des traducteurs professionnels (UNESCO, 2003 : 36). En effet, la traduction automatique fait face à de nombreux problèmes d'ambiguïtés sémantiques et syntaxiques (Braschler, 2004 : 189). Plusieurs stratégies de désambiguïsation ont été suggérées, mais sans grand succès (Kishida, 2005 : 439-443). À cela s'ajoute le fait que les algorithmes de RIML utilisant d'autres ressources linguistiques (dictionnaires ou corpus) ne sont pas plus compétitifs (Kishida, 2005 : 437). En conséquence, nous pouvons conjecturer que lorsque les mécanismes utilisés pour la traduction, notamment les logiciels de traduction automatique, seront plus performants, les algorithmes de RIML le seront aussi.

Deuxièmement, l'intérêt mitigé pour les algorithmes de RIML s'explique par le fait que, souvent, les processus de traduction et de repérage se font séparément, c'est-à-dire que le choix des mots traduits se fait de façon indépendante (Nie, 2002). Par exemple, lorsqu'un dictionnaire est utilisé comme ressource linguistique, la traduction retenue par cette technique est celle qui apparaît en premier dans le dictionnaire, étant considérée comme la plus courante. Il en résulte donc une perte sémantique parfois importante, qui serait évitée par l'intégration des différentes étapes de la RIML (Nie, 2002).

Troisièmement, les difficultés rencontrées au moment de la fusion des résultats expliquent en partie le retard dans l'implantation des algorithmes de RIML. Généralement, on retrouve deux catégories d'architectures : l'architecture centralisée, où les collections de documents en diverses langues sont considérées comme une seule et même collection, et indexées dans un seul index, et l'architecture distribuée, où les documents en différentes langues sont indexés et repérés séparément. C'est avec cette seconde architecture que survient le problème de fusion des résultats. En effet, les documents en différentes langues sont repérés séparément, demandant alors au système de fusionner les différentes listes de résultats obtenus (Nie, 2002). Cette fusion s'effectue difficilement puisque deux collections différentes n'utilisent pas nécessairement les mêmes mots pour exprimer les mêmes concepts, ce qui donne des pondérations différentes (Braschler, 2004 :194). Plusieurs stratégies de fusion ont été proposées depuis quelques années (Chen et Gey, 2004) et continuent de faire l'objet de nombreuses recherches. Toutefois, aucune d'entre elles n'a vraiment réussi à s'imposer jusqu'à maintenant (Gey, Kando et Peters, 2005 : 420).

Quatrièmement, peu d'efforts ont été faits afin d'identifier les besoins des utilisateurs des systèmes de RIML et d'avoir une meilleure compréhension de leurs comportements de recherche (Petrelli, Beaulieu et Sanderson, 2002). Or cette méconnaissance a nui considérablement au déploiement des algorithmes de RIML dans les systèmes d'information. En effet, certaines techniques, comme le contrôle de pertinence et la structuration des requêtes, mettent en relief l'importance de tenir compte des besoins et des comportements des utilisateurs. Une meilleure connaissance de ces besoins et ces comportements semble donc nécessaire pour mener à bien la conception de meilleurs algorithmes de RIML (Gey, Kando et Peters, 2005 : 420).

Enfin, on peut se demander s'il est possible d'offrir un service de RIML au même coût que la recherche dans un environnement monolingue. La RIML exige des opérations entraînant des frais importants. Par exemple, chaque traduction demande l'examen de plusieurs équivalences possibles et le coût résultant doit être calculé au moment même où le système doit optimiser le temps de recherche en sélectionnant les meilleurs résultats et cela, le plus rapidement possible (Oard, 2002). L'indexation devient donc un facteur essentiel pour améliorer l'efficacité et diminuer les frais d'exploitation des algorithmes de RIML.

En résumé, la piètre performance des ressources utilisées pour la traduction, le manque d'intégration des différentes étapes, l'incapacité d'offrir un mécanisme adéquat de fusion des résultats, la méconnaissance des besoins et des comportements des utilisateurs, et les coûts prohibitifs expliquent le retard dans l'adoption des algorithmes de RIML par les systèmes d'information.

Le cas du repérage d'images dans les collections multilingues

En plus de la diversité linguistique observée sur le Web, on constate le développement croissant de bases de données et de collections composées de différents types de documents textuels ou multimédias, ce qui complexifie également le processus de repérage. Depuis le XIX^e^ siècle, le catalogage, la classification et l'indexation se préoccupent surtout de documents textuels. Aujourd'hui, toutefois, l'accès au

matériel multimédia soulève autant d'intérêt pour les chercheurs, sinon plus.

Considérons, par exemple, le repérage de l'image numérique fixe. Ce type de repérage offre des caractéristiques différentes de celles de la recherche de documents textuels. Par exemple, le type de requêtes, la manière dont les requêtes sont formulées, la méthode utilisée pour le repérage, la manière dont la pertinence des résultats est évaluée, la participation de l'individu au processus de recherche et les différences cognitives fondamentales d'interprétation du matériel visuel plutôt que textuel se distinguent par rapport au repérage de documents textuels (Clough et Sanderson, 2003a). En outre, le repérage d'images dans une collection multilingue est compliqué par le fait que l'individu doit faire ses requêtes dans une langue différente de celle de la collection (Clough et Sanderson, 2003a). Par conséquent, les systèmes de repérage d'images ont tout à gagner à inclure un algorithme de RIML.

État actuel

En règle générale, deux approches sont utilisées pour l'indexation et le repérage de l'image numérique : les systèmes basés sur le contexte ou sur le contenu. L'approche basée sur le contexte suppose que le repérage s'effectue à l'aide des métadonnées textuelles associées à l'image, alors qu'avec les systèmes par contenu, la phase d'indexation des images n'implique à aucun moment l'utilisation du langage. Les images sont identifiées par des valeurs associées à certains paramètres (couleur, texture et forme, notamment), plutôt que par des éléments textuels (Boudry et Agostini, 2004 : 97). De plus, les systèmes par contenu n'obligent pas l'individu à conceptualiser sa requête avec des mots, en donnant la possibilité d'effectuer une requête à l'aide d'un croquis ou d'une image similaire. Ces systèmes permettent ainsi de s'affranchir de toute contrainte linguistique (Boudry et Agostini, 2004 : 97).

Cependant, malgré le potentiel incontestable des systèmes par contenu, on constate que les informations numériques extraites de l'image n'ont pas nécessairement de sens pour l'individu et que ces systèmes restent complexes à mettre au point et à utiliser. Pour le moment, les méthodes d'indexation et de repérage par contenu de l'image demeurent au stade expérimental (Jörgensen, 2003). Même si plusieurs chercheurs considèrent que l'efficacité du repérage d'images passe par une combinaison des deux approches (par contenu et par contexte), les métadonnées textuelles continuent à jouer un rôle majeur dans le repérage de l'image numérique fixe.

En 2003, le *Cross Language Evaluation Forum* (CLEF) développait un nouveau champ de recherche (ImageCLEF) portant spécifiquement sur l'utilisation des algorithmes de RIML dans les systèmes de repérage d'images. Quatre expériences ont alors été réalisées. Cette initiative, qui s'est poursuivie en 2004 (18 expériences) et en 2005 (27 expériences), est également inscrite à l'ordre du jour du CLEF 2006. Comme on le constate, l'intérêt pour l'intégration des algorithmes de RIML aux systèmes de repérage d'images, même s'il est plutôt récent, s'accroît d'année en année.

Par exemple, deux avenues de recherche ont été explorées par ImageCLEF en 2003 : l'étude des interactions des utilisateurs avec une interface de repérage d'images, et le repérage d'images *ad hoc*, où une série de requêtes sont effectuées dans une langue pour repérer des images dont les légendes associées sont en anglais (ImageCLEF, 2003). Les résultats de ces premières expériences font ressortir principalement que les requêtes pour des images sont plutôt courtes et concernent des objets ou des lieux spécifiques. De plus, il semble que la ressource linguistique choisie pour ces expériences (un système de traduction automatique) convienne à plusieurs paires de langues, mais démontre aussi certaines limites. Enfin, ces premières expériences, basées strictement sur le repérage textuel, indiquent que le repérage serait amélioré par l'introduction de méthodes basées sur le contenu de l'image et de techniques comme le contrôle de pertinence, notamment (Clough et Sanderson, 2003b).

Poursuivant les recherches entreprises en 2003, lmageCLEF 2004 abordait trois autres avenues de recherche : le repérage traditionnel d'images (images accompagnées de légendes rédigées en anglais), le repérage à partir d'une base d'images médicales où la majorité des images étaient associées à des notes de cas (diagnostic, description en texte libre, présentation clinique, mots-clés, titre, etc.) et une piste de recherche visant à étudier les interactions des individus avec le système de repérage d'images (ImageCLEF, 2004). Plusieurs observations émanent des travaux effectués en 2004. D'abord, les individus réussissaient assez bien à repérer des images avec un système intégrant un algorithme de RIML. Ensuite, le développement des requêtes, mais aussi le contrôle de pertinence, amélioraient le repérage d'images. En ce qui concerne les images médicales, les études indiquent que l'utilisation des caractéristiques physiques de l'image aurait un impact significatif sur leur repérage. Enfin, les usagers étaient disposés à soumettre de nouveau leur requête ou à la reformuler et à examiner un bon nombre d'images, afin de trouver ce qu'ils cherchaient (Clough, Sanderson et Müller, 2004).

En 2005, les travaux du CLEF ont repris les pistes proposées en 2004 et ajouté une nouvelle avenue de recherche concernant les annotations attribuées automatiquement aux images médicales (ImageCLEF 2005). Les travaux de 2005 mettent en relief l'intérêt grandissant pour l'intégration des caractéristiques physiques de l'image pour le repérage. De plus,

ces travaux démontrent l'importance d'utiliser un mécanisme de traduction optimal dans les systèmes de repérage d'images et d'afficher les résultats de manière hiérarchique (de l'image la plus précise à la moins précise). Enfin, les résultats de 2005 soulignent le besoin d'offrir une interface donnant plus de contrôle à l'individu pour la formulation et la reformulation des requêtes (Clough *et al.*, 2005).

Parallèlement à la recherche sur l'utilisation des algorithmes de RIML pour le repérage d'images dans les collections multilingues, des travaux sur le repérage d'images en général se poursuivent depuis plusieurs années. À l'instar de la recherche en repérage d'images faisant appel à la RIML, ces études font ressortir deux éléments majeurs : l'importance d'intégrer l'approche par contenu de l'image à l'approche basée sur le contexte, ainsi que la nécessité de poursuivre l'étude des besoins et des comportements des utilisateurs des systèmes de repérage d'images (Gey, Kando et Peters, 2005 : 424).

Soulignons, en terminant, que même si le repérage d'images à l'aide de la RIML constitue un domaine de recherche relativement nouveau, les travaux effectués jusqu'à maintenant démontrent déjà tout le potentiel qu'apporterait l'intégration des algorithmes de RIML pour améliorer le repérage d'images.

Potentiel de la RIML

Le potentiel des algorithmes de RIML pour le repérage d'images est considérable. Le principal intérêt de l'intégration des fonctionnalités de la RIML aux systèmes de repérage d'images est d'en augmenter le rappel, puisque ces algorithmes donneront accès aux collections multilingues, ce qui n'est pas le cas avec les systèmes de repérage monolingues.

De plus, l'intégration des algorithmes de RIML aura pour effet de rendre les systèmes d'information plus polyvalents. Par exemple, les divers moteurs de recherche disponibles sur Internet tentent de diversifier au maximum les fonctionnalités offertes aux utilisateurs, allant de la comparaison de prix au gestionnaire de photographies, en passant par les services de messagerie. La possibilité de repérer des images dans les collections multilingues constitue une fonctionnalité intéressante que les moteurs de recherche pourraient ajouter à la gamme de produits offerts à leurs utilisateurs.

Il semble donc de plus en plus nécessaire de s'intéresser à différents paramètres dans des expériences en repérage d'images utilisant les algorithmes de RIML. Jusqu'ici, par exemple, on s'est peu préoccupé des critères touchant l'utilisabilité des systèmes. Pourtant, tout laisse croire que ces éléments apporteraient un éclairage nouveau sur les utilisateurs réels, de même que sur leur façon d'utiliser les systèmes de repérage d'images. Bref, ces critères sont d'un grand intérêt dans la conception de meilleurs algorithmes destinés au repérage d'images dans les collections multilingues (Petrelli, Beaulieu et Sanderson, 2002).

Conclusion

Selon l'UNESCO, la langue est le fondement de la communication entre les personnes et fait également partie de leur patrimoine : « *L'Internet étant une source d'information à vocation mondiale, il semble indispensable de mettre l'information à la disposition d'un plus grand nombre de personnes dans leur propre langue* » (UNESCO, 2003 : 35).

Parmi les défis à relever, la RIML doit prévoir des mécanismes permettant, entre autres, l'identification automatique du caractère multilingue des sources d'information (sites et documents) ; une forme d'aide à l'individu dans la formulation de ses requêtes s'adressant à des sites multilingues ; l'interrogation dans sa propre langue des sites multilingues ; l'identification et l'extraction automatique du contenu des documents multilingues ; le filtrage et le transfert de l'information multilingue en fonction des profils des individus ; la traduction dans la langue de l'individu des documents ou parties de documents rédigés dans une autre langue et, surtout, la possibilité de repérer toute forme de documents.

Face à ces nombreux défis, on note un intérêt croissant des sciences de l'information pour cette perspective multilingue et l'apparition de nouvelles avenues de recherche. Deux raisons majeures expliquent cette orientation. D'une part, à cause des considérations économiques liées au commerce électronique. Par exemple, le développement du multilinguisme sur le Web relève d'une nécessité stratégique, dans la mesure où les entreprises qui limiteraient leur champ d'action à une communauté linguistique donnée devraient aussi s'attendre à un regain de concurrence provenant d'autres régions du globe. D'autre part, pour prévenir l'exclusion sociale. Par exemple, les personnes âgées risquent de devenir le groupe socioculturel des laissés-pour-compte de la société de l'information, en raison principalement de leur manque de familiarisation avec les ordinateurs. Dans leur cas, la mise en place d'outils multilingues sur le Web (recherche d'information, traduction, etc.) constitue un moyen efficace de renverser cette tendance.

Les enjeux engendrés par la diversité linguistique du Web sont énormes. Beaucoup de travail reste à faire pour parvenir à briser l'isolement des individus et leur donner le plein accès à l'information, peu importe leur langue. ◉

SOURCES CONSULTÉES

Boudry, Christophe et Clemence Agostini. 2004. Étude comparative des fonctionnalités des moteurs de recherche d'images sur Internet. *Documentaliste*, vol. 41, n° 2 : 96-105.

Bouis, Sonia. 1999. Les agents intelligents. <*http://www.enssib.fr/autres-sites/dessid/dessid99/gedbouis.pdf*> (consulté sur Internet le 15 avril 2006).

Braschler, Martin. 2004. Combination approaches for multilingual text retrieval. *Information Retrieval*, vol. 7, n^{os} 1-2 : 183-204.

Braschler, Martin et Peter Schäuble. 2000. Using corpus-based approaches in a system for multilingual information retrieval. *Information Retrieval*, vol. 3, n° 3 : 272-284.

Chaudiron, Stéphane. 2002. La question du multilinguisme en contexte de veille sur Internet. In *Multilinguisme et traitement de l'information*, sous la direction de Frédérique Segond. Paris : Hermes Science, 63-85.

Chen, Aitao et Frederic C. Gey. 2004. Multilingual information retrieval using machine translation, relevance feedback and decompounding. *Information Retrieval*, vol. 7, n^{os} 1-2 : 149-182.

Clough, Paul et Mark Sanderson. 2003a. The CLEF 2003 cross language image retrieval task. <*http://clef.isti.cnr.it/2003/WN_Web/45.pdf*> (consulté sur Internet le 16 avril 2006).

Clough, Paul et Mark Sanderson. 2003b. Sheffield at ImageCLEF 2003. <*http://clef.isti.cnr.it/ 2003/WN_Web/46.pdf* (consulté sur Internet le 16 avril 2006).

Clough, Paul, Mark Sanderson et Henning Müller. 2004. The CLEF cross language image retrieval track (Image CLEF) 2004. <*http://clef.isti.cnr.it/2004/working_notes/WorkingNotes2004/55.pdf*> (consulté sur Internet le 16 avril 2006).

Clough, Paul *et al.* 2005. The CLEF 2005 cross-language image retrieval track. <*http://www.clef-campaign.org/2005/working_notes/workingnotes2005/hershos.pdf*> (consulté sur Internet le 16 avril 2006).

Detlor, Brian et Clément Arsenault. 2002. Web Information seeking and retrieval in digital library contexts : towards an intelligent agent solution. *Online Information Review*, vol. 26, n° 6 : 404-412.

Ethnologue. 2006. Statistical summaries. <*http://www.ethnologue.com/ethno_docsintroduction.asp*> (consulté sur Internet le 22 mai 2006).

Gey, Frederic C., Noriko Kando et Carol Peters. 2005. Cross-language Information retrieval : the way ahead. *Information Processing and Management*, vol. 41, n° 3 : 415-431.

Grefenstette, Gregory. 2002. Présence des langues sur le WWW et construction des ressources linguistiques. In *Multilinguisme et traitement de l'information*, sous la direction de Frédérique Segond. Paris : Hermes Science, 47-61.

Hedlund, Turid *et al.* 2004. Dictionary-based cross-language information retrieval : learning experiences from CLEF 2000-2002. *Information Retrieval*, vol. 7, n^{os} 1-2 : 99-119.

Hillier, Mathew. 2003. The role of cultural context in multilingual Website usability. *Electronic Commerce Research Applications*, vol. 2, n° 1 : 2-14.

ImageCLEF. 2003. Introduction. <*http://ir.shef.ac.uk/imageclef/2003/*> (consulté sur Internet le 15 avril 2006).

ImageCLEF. 2004. Introduction. <*http://ir.shef.ac.uk/imageclef/2004/*> (consulté sur Internet le 15 avril 2006).

ImageCLEF. 2005. Introduction. <*http://ir.shef.ac.uk/imageclef/2005/*> (consulté sur Internet le 15 avril 2006).

Internet World Stats. 2006. Internet users by language. <*http://www.internetworldstats.com/stats7.htm*> (consulté sur Internet le 12 avril 2006).

Jörgensen, Corinne. 2003. *Image retrieval theory and research*. Lanham MD : Scarecrow Press.

Kishida, Kazuaki. 2005. Technical issues of cross-language information retrieval : a review. *Information Processing and Management*, vol. 41, n° 3 : 433-455.

Nie, Jian-Yun. 2002. Towards a unified approach to CLIR and multilingual IR. <*http://ucdata.berkeley.edu/sigir-2002/sigir2002CLIR-04-nie.pdf*> (consulté sur Internet le 12 avril 2006).

Oard, Douglas W. 2002. When you come to a fork in the road, take it : multiple futures for CLIR research. <*http://www.glue.umd.edu/~dlrg/filter/papers/sigirpositionpaper2.pdf*> (consulté sur Internet le 15 avril 2006).

Oard, Douglas W. et Anne R. Diekema. 1998. Cross-language information retrieval. In *Annual Review of Information Science and Technology*, sous la direction de Martha E. Williams, vol. 33. Medford : Information Today Inc., 223-256.

Oard, Douglas W. et Bonnie J. Dorr. 1996. A survey of multilingual text retrieval. In *Technical Report UMIACS-TR-96-19*, University of Maryland, Institute for Advanced Computer Studies. <*http://www.clis.umd.edu/dlrg/filter/papers/mlirps*> (consulté sur Internet le 15 janvier 2004).

Oard, Douglas W. et Funda Ertune. 2002. Translation-based indexing for cross-language retrieval. <*http://www.glue.umd.edu/~dlrg/filter/papers/eciro2.pdf*> (consulté sur Internet le 12 avril 2006).

Petrelli, Daniella, Micheline Beaulieu et Mark Sanderson. 2002. User participation in CLIR research. <*http://ucdata.berkeley.edu/sigir-2002/sigir2002CLIR-12-petrelli.pdf*> (consulté sur Internet le 13 avril 2006).

Plamondon, Luc et Leila Kosseim. 2003. Le Web et la question-réponse : transformer une question en réponse. <*http://www.cs.concordia.ca/~kosseim/Publications/jft2003.pdf*> (consulté sur Internet le 13 avril 2006).

Robinson, Clinton et Karl Gadelii. 2003. Writing unwritten languages. <*http://tinyurl.com/dbvaa*> (consulté sur Internet le 13 avril 2006).

Salton, Gerard. 1987. A historical note : the past 30 years in information retrieval. *Journal of the American Society for Information Science*, vol. 38, n° 5 : 375-380.

UNESCO. 2003. Diversité culturelle et linguistique dans la société de l'information. <*http://portal.unesco.org/ci/fr/file_download.php/fo138f3685432a579c5cfc5849314368culture_fr.pdf*> (consulté sur Internet le 15 avril 2006).

UNESCO. 2004. La diversité linguistique, culturelle et biologique de la terre. <*http://portal.unesco.org/education/fr/ev.php-URL_ID=18391&URL_DO =DO_PRINTPAGE&URL_SECTION=201.html*> (consulté sur Internet le 15 avril 2006).

Vallin, Alessandro *et al.* 2005. Overview of the CLEF 2005 multilingual question answering track. <*http://www.clef-campaign.org/2005/working_notes/workingnotes2005/vallino5.pdf*> (consulté sur Internet le 15 avril 2006).