
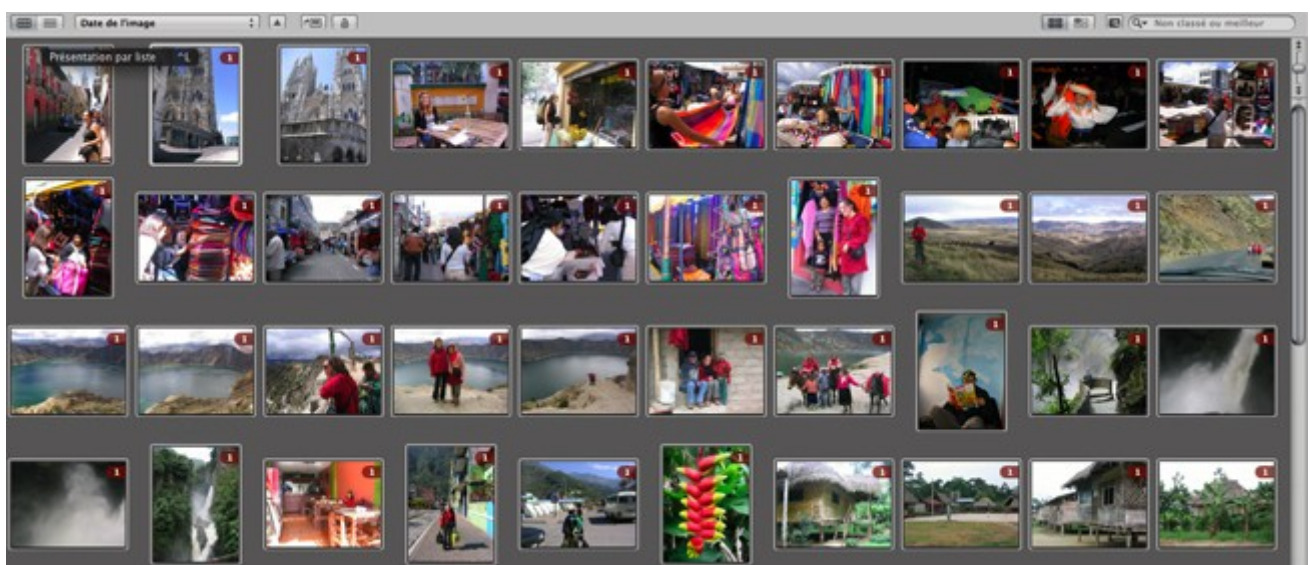


Description et indexation automatiques des documents multimédias : du fantasme à la réalité

La numérisation des images et des contenus audiovisuels permet désormais de les traiter automatiquement. De nombreux travaux ont été menés pour tenter de bâtir des systèmes de description et d'indexation automatiques de ce type de documents suscitant autant d'enthousiasme dans le monde de la recherche que d'inquiétude dans celui des professionnels de l'information-documentation.

L'[indexation](#)  automatique des documents multimédias a pour but de permettre, par le biais de techniques automatiques ou semi-automatiques, l'exploitation de collections de documents. L'apparition de ce domaine de recherche, en ce qui concerne les images et les documents audiovisuels, date de la première moitié des années quatre-vingt-dix, et est donc encore récente. Son émergence a suscité un double mouvement, d'enthousiasme chez les chercheurs qui y ont vu un domaine nouveau d'investigation et qui, selon leur habitude, ont beaucoup promis afin d'attirer des financements pour mener leurs activités, et d'inquiétude chez certains professionnels de la documentation audiovisuelle qui y ont vu une remise en cause de leur métier, voire un danger de disparition de leur emploi. Quelques années ayant passé depuis ces débuts, il est intéressant de faire le point aujourd'hui. Quel est l'objet actuel de l'indexation automatique ? Quelles sont ses possibilités, ses applications, ses perspectives d'évolution ? C'est à ces questions que nous allons tenter de répondre en nous intéressant plus particulièrement au cas des images fixes.




Que signifie indexer une telle collection d'images ?

Image : © INRIA

Des débuts difficiles...

L'expression même d'indexation automatique est sujette à discussion. Elle n'est, tout d'abord, pas largement partagée. À la suite des travaux d'IBM sur le système QBIC (*Query by Image Content*), le domaine s'est développé sous le vocable général d'indexation multimédia « par le contenu », pour le distinguer d'une indexation qualifiée de manuelle. Cette expression supposerait ainsi qu'il y aurait, d'une part, une indexation basée sur le contenu même et par conséquent objective, face à une indexation basée sur l'interprétation, donc subjective, et de ce fait appelée à être reléguée au rayon des pratiques désuètes. On trouve souvent un tel argumentaire dans les thèses du domaine, mais il ne fait que trahir l'ignorance des doctorants, voire de leurs encadrants, vis-à-vis du milieu professionnel de la documentation audiovisuelle et de ses méthodes.

La naissance du domaine a par ailleurs été marquée par une grande effervescence. Rares sont les universités américaines qui n'ont pas lancées leur projet d'indexation automatique d'images. Du coup, il y a eu compétition et surenchère, chacun voulant se démarquer et faire mieux ou, pour le moins, faire une meilleure publicité pour ses propres travaux. Beaucoup se sont rués dans les agences de photo et ont promis d'y remplacer les documentalistes... Puis le souffle est retombé, car une fois que tout le monde eut essayé de retrouver des images de couchers de soleil en utilisant des [histogrammes de couleur](#) , il a bien fallu se rendre à l'évidence que ce n'est pas avec ce genre de technique qu'on allait pouvoir résoudre des problèmes réels. Le chemin s'avérait plus ardu que prévu. La place était alors libre pour les quelques équipes qui avaient choisi de travailler plus à fond dans le domaine, les autres abandonnant le sujet.

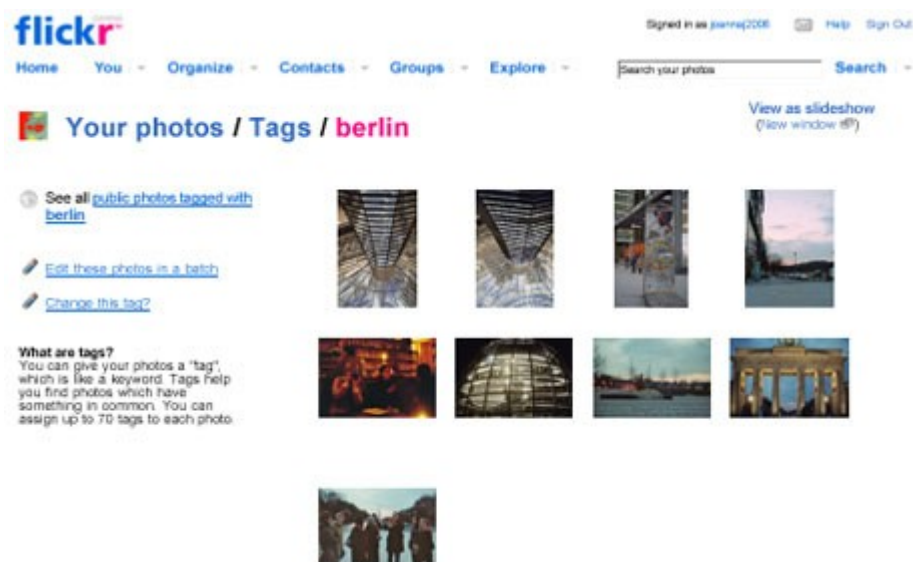
Une autre difficulté vient de l'emploi du terme indexation qui, dans la bouche ou sous la plume de beaucoup, signifie autant la structuration, la description que l'indexation proprement dite des documents, voire la recherche de documents ou la navigation dans des collections. Cette confusion, qui n'est pas sans créer des problèmes de communication avec d'autres domaines, est aussi le signe que les problèmes à résoudre n'étaient pas encore complètement identifiés.

Dans les faits, on peut distinguer plusieurs tâches : d'abord, il y a la structuration des documents, qui est le plus souvent une segmentation et qui consiste à repérer dans un document des entités d'intérêt ; ainsi cherche-t-on à retrouver les divers plans d'une vidéo, à y isoler les objets en mouvement par exemple. Ensuite, on passe à la description des documents ou des éléments issus de la structuration précédente, ce qui consiste à calculer un certain nombre de quantités à partir du contenu ; l'idée sous-jacente est bien entendu que l'on pourra par la suite réduire la manipulation des contenus à celle des descripteurs ainsi calculés ; puis vient la sélection, c'est-à-dire le choix des descripteurs utiles dans un contexte donné, en fonction de la collection considérée et des requêtes qui devront être traitées ; cette étape est suivie de près par l'indexation proprement dite, qui va consister à organiser toute cette information pour y accéder de manière efficace et efficiente ; et enfin, l'utilisation de ces descripteurs pour y retrouver l'information recherchée et répondre à une requête, ou pour naviguer dans la collection de documents. D'autres éléments doivent bien entendu être ajoutés à ceux-ci pour obtenir un système complet, avec des interfaces, de l'interactivité, etc.

Les images fixes et leurs contextes

Dans le domaine des images fixes, plusieurs contextes applicatifs sont intéressants et pourraient bénéficier de techniques automatiques.


Côté grand public, la gestion des collections de photos personnelles et familiales est un sujet d'actualité. Avec la diffusion des appareils photos numériques, les photos personnelles existent désormais en format numérique et peuvent donc être traitées au moyen de programmes informatiques. Par ailleurs, bon nombre de foyers sont équipés d'outils électroniques, que ce soit de magnétoscopes ou d'ordinateurs, qui peuvent servir à stocker et à manipuler de telles images. Dans ce cas précis, le problème vient du fait que les photos sont la plupart du temps stockées sans annotation et que toute recherche ou navigation autre que chronologique n'est pas possible ou reste très pénible.




[Flickr](#) ^W est l'un des plus célèbres sites web gratuits de diffusion et de partage de photos personnelles. Les annotations y sont collaboratives. Un *tag* est un mot-clé qui peut être attribué à une ou plusieurs photos. Il permet entre autre une identification rapide des photos.

Dans le domaine professionnel, les agences de photos ont été les premières contactées, bien qu'elles ne soient pas les seules à gérer de grandes quantités d'images. L'attention a été fortement focalisée sur le travail des documentalistes qui gèrent les photos et répondent aux requêtes des clients. Il est toutefois très vite apparu que ces requêtes étaient de très haut niveau sémantique et que les documentalistes pouvaient s'appuyer sur une bonne connaissance des besoins du client ou sur une compréhension de sa demande qui n'est pas formelle, et sur leur mémoire du contenu de la collection qu'ils gèrent. Toutes ces choses qu'un ordinateur peut difficilement opérer...

Mais il existe d'autres tâches pour lesquelles l'ordinateur peut avoir un apport intéressant. Tout d'abord dans l'aide à l'annotation, dans un fonctionnement semi-automatique :

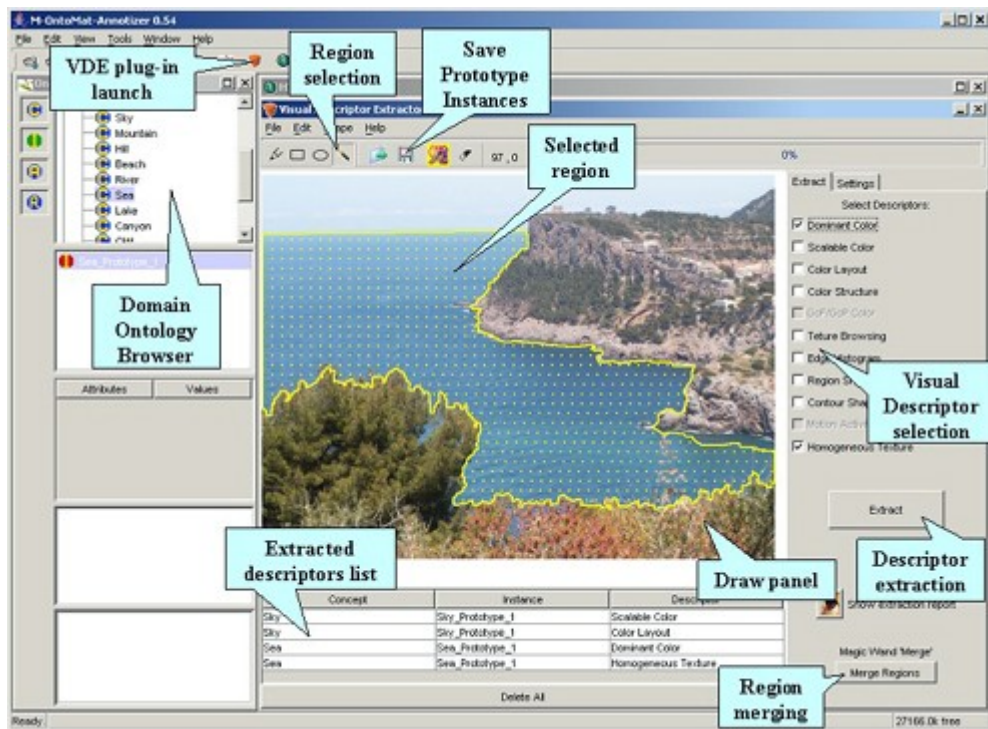
l'ordinateur peut soit proposer des annotations à un documentaliste qui les valide et complète, soit annoter automatiquement une partie de collection à partir des annotations faites manuellement sur une autre partie. Ce type de fonctionnement peut convenir à des possesseurs d'archives qui n'ont pas les moyens d'annoter manuellement l'ensemble de leur collection, par exemple les quotidiens régionaux qui récupèrent de grandes quantités d'images de leurs correspondants locaux. Dans ce domaine, le couplage entre appareils photos et systèmes de positionnement par satellites (GPS ou [Galileo](#) ) apportera une aide appréciable.

La [recherche des copies illégales](#)  devient incontournable : dès qu'un possesseur d'images veut utiliser Internet pour faire connaître son fonds, le souci du piratage devient majeur. Mais l'ampleur de la tâche est immense : comment confronter toutes les images que l'on peut trouver sur Internet avec celles d'une collection de plusieurs millions d'images ? Clairement, il ne peut y avoir de solution qu'automatisée, au moins pour un premier tri. Il y a là un problème intéressant, qui correspond à un vrai besoin et pour lequel il n'y a pas de solution manuelle possible.

D'autres applications correspondant à des collections particulières d'images, sont également envisageables, comme par exemple toutes les applications biométriques. Ainsi, le système de vidéo-surveillance londonien est-il couplé à un système de reconnaissance de visages qui a permis plusieurs arrestations. Dans le domaine policier toujours, la police judiciaire française utilise un système de reconnaissance pour comparer des images pédophiles afin de retrouver celles qui seraient prises dans une même pièce, ou retrouver des indices communs à plusieurs photos. Le domaine médical est aussi fortement demandeur tant pour la formation des médecins que pour l'exploitation des très nombreuses images produites (pour comparer des organes pathologiques, rapprocher des cas atypiques, par exemple). Par ailleurs, la météorologie, les satellites sont de très gros pourvoyeurs de données qui ne sont pas exploitées autant qu'elles pourraient l'être.


Quels instruments pour la description automatique d'images fixes ?

Face à ces besoins, de quels outils disposent les chercheurs ? Deux approches se distinguent, suivant que l'on cherche à décrire l'image dans sa totalité ou seulement des parties de celle-ci.



M-OntoMat-Annotizer, outil d'aide à l'annotation de vidéos.
Source : [aceMedia](http://aceMedia.w) 

Des descripteurs globaux

Dans la première approche, on cherche à calculer des descripteurs globaux. En l'absence de toute indication sur l'image, on en est réduit à observer le signal bidimensionnel qui la constitue. Informatiquement, une image se présente comme une table rectangulaire d'éléments élémentaires appelés pixels, chacun codant sous la forme d'un ou de plusieurs nombres (généralement trois) l'information d'intensité lumineuse  ou de couleur présente en un point. Ce sont donc ces nombres que l'on va utiliser pour décrire les images.

L'information la plus accessible est bien sûr la couleur, puisqu'elle est directement codée au niveau de chaque pixel. Il reste donc à déterminer la couleur présente dans l'image et la proportion de la surface de l'image qu'elle remplit. C'est ce qu'on appelle un histogramme de couleur. Bien évidemment, un tel procédé de description est assez simpliste mais il est suffisant pour rechercher des images de coucher de soleil dans une collection d'images de forêt tropicale... Ce procédé peut être amélioré de deux manières : tout d'abord en utilisant plus finement la distribution des couleurs. Un pixel jaune entouré de bleu ne donnera pas du tout le même effet entouré de jaune. On peut ainsi pondérer l'importance de chaque pixel dans l'histogramme en fonction de son environnement immédiat, ce qui augmente grandement la puissance descriptive de l'histogramme. Par ailleurs, de nombreuses distances existent pour comparer des histogrammes et juger ainsi de la ressemblance entre les images. Le choix de cette distance influe bien entendu sur les résultats, et est un facteur déterminant de la rapidité du système. Malheureusement, les distances les plus intéressantes sont très onéreuses à calculer.

Une deuxième approche consiste à faire abstraction de la couleur et à s'intéresser uniquement aux distributions locales des intensités lumineuses. On va ainsi chercher à déterminer dans

quelle mesure un pixel clair est plutôt entouré de pixels clairs ou foncés. L'idée est de décrire la texture de l'image, c'est-à-dire le fait qu'une image de nuage ne ressemble pas à une image d'herbes, de marbre ou de feuillage, indépendamment de sa couleur. Le problème est que la notion de texture n'a pas de définition formelle : de très nombreux descripteurs ont été proposés depuis les cooccurrences d'Haralick en 1973, mais chacun d'eux a un domaine d'usage restreint, sans que ce domaine ne soit lui-même bien défini. Aussi trouve-t-on des descripteurs de texture dans de nombreux outils pour le cas où... mais leur utilité n'est pas souvent prouvée.

Le troisième élément le plus utilisé est la forme. QBIC proposait ainsi une petite palette graphique : l'utilisateur traçait une forme et le système retrouvait les images possédant des formes semblables. Décrire une forme ne pose pas de problème majeur, la transformée de Fourier-Mellin convient tout à fait par exemple. La difficulté est de trouver la forme dans l'image, de la délimiter. En dehors des images ayant un fond uniforme (objet clair sur fond noir), il n'existe aucune méthode permettant de trouver des formes de manière générale. En effet, dans une image normale, les objets se superposent : ils sont constitués de parties de couleur, de texture et de formes très différentes. La notion même d'objet est mal définie. Une voiture par exemple est un assemblage d'objets (roues, gentes, carrosserie, etc.). L'utilisateur humain pense la forme comme étant celle d'un objet (la voiture en l'occurrence), et lui associe par conséquent une valeur sémantique. Mais comment l'ordinateur peut-il savoir qu'on veut la voiture avec ses roues et ses enjoliveurs et pas seulement ces deux objets pris séparément ? Pourquoi la voiture serait-elle la forme à extraire plutôt que l'essuie-glace, qui a au moins une couleur uniforme ? Ce problème de délimitation des formes, QBIC l'a résolu en extrayant les formes de toutes les images manuellement.

Deux issues sont alors possibles. D'une part, on peut renoncer à segmenter l'image et à la découper selon un quadrillage prédéfini puis essayer de regrouper les carreaux entre eux. C'est simple, mais approximatif. D'un autre côté, de nouveaux algorithmes essayent de ne découper l'image qu'en quelques régions, moins d'une dizaine. Ces régions sont plus grossières, mais semblent bien mieux correspondre aux zones d'intérêt dont on a besoin pour décrire une image. Une fois ces zones délimitées, on peut soit les décrire par leur forme, mais celle-ci reste assez peu fiable, soit par leur couleur ou leur texture. On obtient ainsi une description de l'image en quelques zones qui paraît bien plus puissante que les descriptions globales utilisées jusqu'à présent.

Des descripteurs locaux

Une image peut être recadrée, tournée, agrandie, et on veut pouvoir la reconnaître tout de même ! Seulement, les descripteurs globaux sont mal adaptés pour faire face à ces transformations. On peut alors chercher à n'utiliser que des descriptions locales. Pour cela, on va rechercher des zones d'intérêt dans l'image, puis décrire chacune de ces zones. De manière générale, on peut chercher des points, des courbes ou des régions de l'image.

En ce qui concerne les régions, le problème est étudié de longue date. Les petites régions sont très instables. Celles correspondant à un objet rencontrent le problème de délimitation évoqué à propos des formes. Comment savoir quelles sont les régions les plus intéressantes dans une image ? Reste l'option de ne rechercher que quelques régions dans une image, moins d'une dizaine en général. Cela permet de séparer le centre de l'image de ce qui est soit au premier

soit au dernier plan. Mais il n'y a aucune assurance que les régions extraites aient quelque valeur sémantique. Cela dit, on ne sait pas faire mieux ! Le problème est assez similaire pour les courbes : détecter une courbe est difficile. Entre deux images d'une même scène, il suffit d'un changement très mineur d'éclairage ou de point de vue pour que la courbe soit coupée en morceaux. D'autre part, les courbes sont reliées entre elles, et chaque intersection ou jonction pose un problème sans solution immédiate.

L'emploi de points d'intérêt s'est quant à lui révélé bien plus fructueux. En effet, il est possible de définir mathématiquement ce qu'est un point d'intérêt, par exemple comme étant un point qui ressemble le moins possible à ses voisins (un minimum d'autocorrélation) ou encore comme un point porteur d'un maximum d'informations au sens de la théorie de l'information. Ensuite, ces points résistent bien aux transformations que l'on peut faire subir aux images : on dit qu'ils sont répétables. La méthode de Harris modifiée par Cordelia Schmid est apparue comme une des meilleures sur ce plan. Autour de chacun de ces points, on calcule alors des quantités décrivant le signal, en veillant à ce que ces quantités restent égales si l'on modifie l'image, par exemple en la faisant tourner. Les deux méthodes les plus répandues et les plus performantes sont les invariants différentiels de Luc Florack, que C. Schmid est la première à avoir utilisés, dans un but de description pour la recherche d'images, et les invariants SIFT (*Scale Invariant Feature Transform*) de David Lowe.

Ces descripteurs locaux se sont révélés, dans la pratique, extrêmement performants pour des tâches de détection de copie, par exemple. Les taux de fausse détection sont très bas. Cela est dû au fait qu'ils décrivent de manière très discriminante la texture locale de chaque point. Ils permettent donc de retrouver des objets précis. C'est pourquoi ils sont si adaptés pour la détection de copies.

Avec toutes ces descriptions, qu'elles soient locales ou globales, on reste au niveau du signal. On aimerait pouvoir traiter les images avec un vocabulaire plus riche sémantiquement. Il n'y a, pour le moment, aucun moyen un tant soit peu générique de trouver des [descriptions sémantiques](#) + dans les images.

Des combinaisons de descripteurs

Les images sont polysémiques, c'est-à-dire qu'elles peuvent avoir plusieurs sens. On peut de plus envisager différents types de description : les mots clés et concepts permettent d'avoir accès à une partie de leur information, un histogramme de couleur donne un autre type d'information, complémentaire du premier. Une des principales difficultés est de faire cohabiter au sein d'un même système toutes ces descriptions, qui sont de type très différent d'un point de vue informatique, puis de les faire collaborer. Il faut pour cela un formalisme capable de combiner les différentes facettes de la



Les images comme les mots peuvent être polysémiques. Source : Wikipédia
description d'une image, un langage de requête plus riche, puis un mécanisme de mise en correspondance des requêtes avec les descriptions. Les graphes conceptuels fournissent un outil intéressant pour combiner les descripteurs. Pour les requêtes, le problème est de trouver un langage qui

ne soit pas trop abscons, ou une interface qui permette une traduction efficace entre le langage naturel et le langage de requête interne au système. C'est là un sujet de recherche actif.

Avec de telles méthodes, on se rapproche d'une manipulation plus sémantique des images. Du point de vue des traiteurs d'images, on a donc formidablement progressé. En fait, on commence à obtenir uniquement une petite partie de ce que l'on trouve dans un texte : des mots qui désignent des parties de contenu. De tels mots permettent d'utiliser les images dans les [moteurs de recherche](#) ⁺ les plus courants du web : c'est effectivement une grande avancée. Mais on va aussi se confronter aux mêmes limites que celles rencontrées par les textes eux-mêmes : ambiguïtés, manque de précision... Dans le cas des images, la linguistique ne pourra malheureusement pas nous aider.

Que faire une fois la description effectuée ?

Décrire ne suffit pas. Un autre paramètre important est la taille des bases à gérer. Tant que l'on s'en tient à des tailles modestes, il est toujours possible pour un ordinateur de passer toutes les images en revue pour résoudre une tâche. Pour des collections de plusieurs millions d'images, cela n'est plus possible si l'on veut garder des temps de réponse assez courts.

Le problème de l'indexation

On fait alors face à un double problème. Premièrement, les données numériques ne peuvent être gérées de manière efficace par les [systèmes de gestion de bases de données](#) ^W (SGBD) habituels. La cause est multiple. D'une part, ces données se présentent sous forme de vecteurs numériques de grande dimension, et on doit utiliser toutes les dimensions à la fois, ce que les SGBD ne savent pas bien faire. D'autre part, on ne fait pas de requêtes exactes, mais approchées : on cherche des images ayant des descripteurs voisins d'un descripteur requête, et non pas strictement égaux. Du coup, on ne cherche pas une valeur, mais les données proches d'une valeur. Dans un espace de grande dimension, les plus proches voisins d'une requête ont toutefois tendance à être aussi éloignés que ses plus lointains voisins, et pour s'assurer qu'un descripteur est bien le plus proche, il faut bien souvent scruter une bonne partie des données, ce qui est très inefficace.

On commence à connaître quelques solutions à ce problème, mais il reste à les évaluer sur de grands ensembles de données. L'idée de base est de confiner la recherche, c'est-à-dire de réduire le plus vite possible l'ensemble des données que l'on va être forcé de passer en revue. Si l'on veut vraiment aller vite, il faut accepter de n'obtenir qu'un résultat approximatif. Les algorithmes se distinguent alors par leur manière de gérer cette approximation.

Le deuxième problème posé est celui des descriptions codées sous forme de graphes, comme ceux que l'on peut vouloir utiliser pour combiner différentes descriptions. Comparer des graphes est une opération intrinsèquement compliquée, et on ne sait pas indexer dans ce cas : on est, sauf exception, obligé de tout passer en revue. Il faut alors utiliser toutes les spécificités des graphes que l'on utilise pour arriver à accélérer les recherches.

Un composant crucial : l'interface

Autre source de problème : l'interface. Un système ne sert à rien s'il ne peut rendre service à un utilisateur. Dans le cas présent, il faut que l'utilisateur puisse exprimer son besoin d'information pour que le système retrouve les images pertinentes dans la collection. On se trouve souvent face à un fossé entre le besoin de l'utilisateur, qu'il peut exprimer en langue naturelle et qui est de très haut niveau sémantique, et les descriptions des images dont on dispose, qui sont, elles, de bas niveau sémantique. C'est le *semantic gap* : comment traduire la recherche d'une image illustrant la mondialisation économique en terme d'images bleue, verte ou rouge ?

La manière la plus courante de contourner ce problème est l'emploi du paradigme de la recherche par l'exemple : au lieu de formuler la requête textuellement, on fournit au système une image, charge à lui de trouver les images les plus ressemblantes. De nombreuses interfaces permettent de pondérer la prise en compte des différentes facettes de cette image requête, mais il faut bien reconnaître que cette fonctionnalité n'a aucun sens pour la plupart des utilisateurs.

L'utilité de la recherche par l'exemple dépend, bien entendu, des applications : c'est un mode de fonctionnement idéal pour la détection de copies, où l'image requête est l'image suspecte, et où l'utilisateur ne souhaite pas avoir à décrire l'image pour savoir si elle vient de son stock ou non. Par contre, pour une recherche dans une grande collection d'une agence de photos, cela risque de n'être ni efficace, ni pratique. Divers stratagèmes ont été proposés : le système peut proposer des images issues de la collection pour débiter la recherche, puis demander à l'utilisateur de choisir une deuxième requête dans les résultats de la première recherche, et ainsi de suite. Une variante (habituellement appelée *relevance feedback*) consiste à désigner les images les plus intéressantes du résultat fourni plutôt que d'en désigner une seule.

Autre aspect de l'interface, la manière dont elle va présenter les résultats. En ce domaine, l'imagination n'est guère au pouvoir. La plupart des systèmes présentent à l'utilisateur, quelle que soit la requête, un nombre fixe d'images, entre 12 et 20. Ces images sont ordonnées par ordre décroissant de pertinence. Ainsi, ces systèmes retrouvent toujours des images, même si elles n'ont aucun rapport avec la requête. À l'inverse, si de très nombreuses images sont pertinentes, le système n'en présentera qu'une douzaine. Quand à l'ordre selon la pertinence, force est de constater qu'il ne correspond souvent à rien pour l'utilisateur. Quelques propositions plus innovantes ont été faites, mais il faut reconnaître que la communauté de recherche sur les interfaces ne s'est pas encore intéressée au problème.

Quelles perspectives pour cette technique ?

Comme toute nouvelle technologie, l'indexation automatique des documents multimédias se présente sous deux aspects : possibilité de nouvelles applications, mais mise en cause de manières de faire antérieures. En matière d'emploi, si la première piste ouvre des opportunités, la deuxième suscite légitimement des craintes. Qu'en est-il dans le cas présent ?

L'indexation automatique offre des possibilités tout à fait prometteuses en terme de manipulation de grands volumes de données, car elle permet de réaliser des tâches simples comme la détection de copie. Pour de nombreuses autres tâches, les algorithmes sont loin d'offrir des performances, en qualité des résultats, qui permettent d'envisager une documentation complètement automatique. Il y a là des raisons profondes et sérieuses : les

langues naturelles ont été un des premiers objets d'étude de l'informatique, et leur traitement reste très loin d'être parfaitement maîtrisé. La perception d'un document fait appel à tout un univers culturel, social et personnel qui a un fort impact sur cette perception. Et qui est très difficile à appréhender et à manipuler par ordinateur.

Un autre facteur intervient aussi : le volume de documents multimédias générés et stockés augmente à une folle allure. Il est à craindre que la recherche n'avance pas aussi vite en ce moment. Que ce soit sur le web ou à la télévision, sans parler de numérisation de fonds plus anciens, c'est plutôt la noyade qui nous guette que la pénurie de matière. C'est cela, plus que l'émergence de l'indexation automatique, qui risque de changer l'environnement d'utilisation de ces documents, que ce soit par des documentalistes, des spécialistes de médias (journalistes, analystes, créateurs, diffuseurs, etc.) ou par le grand public. Puissent-ils tous trouver dans la boîte à outils de l'indexation automatique les moyens dont ils ont besoin pour faire face à cet afflux et réellement en profiter.