Les traitements documentaires automatiques et le passage du temps

Lyne DA SYLVA
École de bibliothéconomie et des sciences de l'information –
Université de Montréal

Résumé :

Dans cet article, nous examinons le sort des documents qui ne sont pas destinés à vivre longtemps et qui ne méritent ainsi aucun traitement documentaire traditionnel. Nous défendrons la thèse que les traitements automatiques, bien que produisant des résultats de moindre qualité que les traitements humains, ont leur place pour le traitement de certains documents éphémères. Ils doivent cependant répondre à des normes de qualité qu'il faut définir de nouveau dans le contexte numérique.

Traitement documentaire : contexte et définitions

Pour parler de traitement documentaire, nous retiendrons ici le contexte d'un service d'information qui dessert une collectivité d'utilisateurs dans leurs besoins d'information, spécifiquement de ressources d'informations consignées ; bien sûr, étant donné la thèse présentée ici, le support numérique est celui qui nous intéresse particulièrement. Ce service d'information peut être traditionnel (bibliothèque, centre de documentation ou d'archives, etc.) ou non. Notamment, un contexte qui sera repris au cours de ce travail est le réseau Internet, qui donne accès maintenant à un nombre important de documents divers. Dans ce cas, le service d'information est incarné par l'architecture mise en place par les outils que sont les moteurs de recherche ou les annuaires, ou encore les bases de données textuelles ou documentaires élaborées par un organisme. La collectivité d'utilisateurs pourra couvrir l'ensemble des internautes ou être limitée aux gens autorisés à consulter la base de données documentaires.

Les traitements automatiques qui nous intéressent portent essentiellement sur des versions numériques de l'analyse documentaire. Ceux qui viennent le plus facilement à l'esprit sont les applications d'indexation automatique ou de résumé automatique. Mais dès que l'on a accès à une version numérique d'un texte, d'autres deviennent possibles.

Exigences des traitements traditionnels

La notion de cycle de vie d'un document est cruciale pour plusieurs aspects de la gestion documentaire. Un service d'information y est très sensible. Notamment, la pérennité que l'on reconnaît aux documents qui entrent dans la chaîne documentaire entraîne des attentes : pour ces documents jugés précieux, les utilisateurs potentiels exigent, de la part du service, un traitement à la hauteur de leur valeur. Ainsi, le traitement documentaire traditionnel doit rencontrer un certain nombre d'exigences de qualités, garantes auprès de leurs utilisateurs de la fiabilité du service d'information.

Les exigences de qualités portent non seulement sur la sélection et la préservation des documents, mais aussi sur l'analyse documentaire, la production de représentations fidèles des documents et le mode d'accès à ceux-ci. Parmi les exigences, on relèvera celles qui suivent ; elles seront reprises dans l'examen des performances des systèmes automatiques.

La normalisation : celle-ci est la première étape vers la garantie de qualité. Les normes (Afnor, 1990, 1993, par exemple) portent à la fois sur la méthodologie appliquée et sur les produits résultants. L'existence de normes publiées encadrant l'activité de l'analyse documentaire permet d'assurer la qualité et l'interopérabilité des systèmes et l'échange des informations.

La mise au point de politiques et procédures : afin de préciser le flou laissé dans les normes nationales et internationales (flou nécessaire, pour permettre leur application dans différentes situations), un service se doit de mettre au point des politiques et des procédures contextualisées selon les genres de documents et les profils de ses utilisateurs.

L'évaluation et la validation : l'établissement de processus de validation et d'évaluation des produits a pour but de vérifier que les procédures sont appliquées adéquatement et réalisent le mandat du service d'information.

La rigueur et la cohérence : seules ces dernières peuvent assurer d'obtenir les mêmes résultats entre indexeurs et dans le temps. De cette cohérence découle la possibilité pour les utilisateurs de retrouver à leur tour les documents qui les intéressent. Bien sûr, on connaît les difficultés posées par ces exigences. Nombre d'études attestent de la relativement faible cohérence entre indexeurs (Leininger, 2000) ou entre rédacteurs de résumés (Schlesinger *et alii*, 2003). Le détail atteint par certaines politiques d'analyse documentaire a précisément pour but de réduire les possibilités d'interprétations personnelles menant aux incohérences. Les mesures d'évaluation et de validation tentent de contrôler la qualité des résultats. Bref, la rigueur et la cohérence sont des objectifs déclarés et poursuivis, qui sont néanmoins difficiles à atteindre. On verra dans la suite la différence qui oppose, à ce chapitre, les traitements humains aux traitements automatiques.

L'exactitude et la fidélité : la représentation des documents doit, bien sûr, être fidèle à l'information contenue dans ceux-ci. Cette exigence devrait être la première ; mais dans l'enseignement de l'analyse documentaire, on souligne que l'exactitude doit être atteinte dans le respect des normes, des politiques, des procédures ; elle en serait donc une conséquence naturelle. L'analyse documentaire humaine est particulièrement réussie quant à ces aspects. Cela découle notamment de la capacité qu'ont les analystes de réellement saisir le sens, le contenu des documents.

Mais encore : pour mener à bien la tâche d'analyse documentaire, une bonne connaissance des caractéristiques des documents contenus dans la collection constitue l'un des deux points de départ pour une description réussie ; le deuxième réside dans une aussi bonne connaissance des utilisateurs, leurs profils, leur niveau d'expertise, leurs habitudes de travail et de recherche d'information. En vérité, ces deux considérations sont les piliers de toute l'entreprise.

Le paradoxe des traitements automatiques

Comment se positionnent les traitements automatiques par rapport à ces exigences ? Plutôt mal, il faut le dire.

Normes, procédures et évaluation : le développement des produits se déroule généralement de façon *ad hoc*, c'est-à-dire en respectant éventuellement des méthodologies de développement informatiques, mais pas les méthodologies d'analyse documentaire. De surcroît, les projets sont habituellement pilotés par des informaticiens, sans recours aux bibliothécaires, documentalistes ou archivistes. La notion de normes portant sur la description et la gestion documentaire est absente (dans sa presque totalité) de la conception de systèmes informatiques documentaires. Les écrits sur les normes d'indexation dans le contexte numérique proviennent plutôt de la discipline des sciences de l'information que de l'informatique (voir par exemple Anderson, 1994 ; MacDougall, 2000).

En conséquence, il est difficile de faire une évaluation utile des systèmes, puisqu'il n'est pas facile de contrôler ce que l'on évalue. Et pourtant, on observe depuis une dizaine d'années un essor dans l'évaluation des systèmes automatiques de traitement de la langue : les conférences-compétitions TREC et MUC (pertinentes pour l'indexation automatique) (Harman, 1995, 1996 ; MUC, 1995, 1998) et DUC (pour la condensation automatique) (Over, 2002), etc. Les normes d'évaluation se limitent aux résultats attendus, que l'on compare (faute d'autre moyen automatique) à des résultats produits par les humains pour les mêmes tâches.

Exactitude et fidélité : les systèmes automatiques en sont encore à perfectionner les traitements afin de saisir le sens des énoncés contenus dans les documents. La difficulté principale réside dans le fait que les systèmes ne dépassent généralement pas le traitement des chaînes de

caractères (des formes de surface) du texte ; la compréhension du sens est l'objectif poursuivi par les travaux en sémantique computationnelle (voir Lappin, 2003), mais demeure un idéal lointain.

De façon plus inquiétante, les systèmes sont souvent rigides quant aux genres de documents traités. Par exemple, bon nombre de systèmes expérimentaux de condensation automatique sont spécialisés dans le résumé d'articles scientifiques (voir notamment Saggion et Lapalme, 2000, 2002) et ne peuvent résumer adéquatement un autre genre, les articles de journaux par exemple.

Également, les besoins des utilisateurs sont mal connus et peu incorporés. En fait, si l'on considère les logiciels grand public, les utilisateurs visés sont mal définis. Les travaux de recherche portant sur les utilisateurs s'intéressent généralement à leur comportement observable d'après des journaux d'exécution ou de navigation (par exemple, Cooley *et alii*, 2000 ; El-Ramley et Stroulia, 2004), mais n'abordent pas vraiment la question de leurs motivations, leurs besoins ou leur niveau d'expertise du domaine. Leur niveau d'expertise en recherche d'information est parfois pris en compte ; il est capté par les différents niveaux de recherches : simple, avancée, experte.

On constate cependant un paradoxe : les applications informatiques connaissent un succès, une popularité incontestable, que l'on ne peut expliquer que par leur utilité évidente. Il est clair que les critères de succès ne sont pas définis selon le respect des normes et des procédures, mais d'après la réussite de tâches précises, quantifiables et faciles à mesurer. Nous le verrons, les avantages des traitements automatiques se situent donc à l'extérieur de la liste de critères que nous avons examinés jusqu'à présent.

Traitements automatiques pour documents éphémères

Le moteur de développement des solutions informatiques est donc principalement pragmatique et utilitaire. La croissance phénoménale des collections numériques rend les traitements automatiques incontournables, en particulier dans l'exploitation du Web. On observe ainsi une frénésie dans les champs de recherche associés : recherche d'informations (ou repérage d'informations), « gestion de contenu », gestion de la connaissance, condensation automatique, etc.

Mais d'un point de vue théorique, peut-on légitimer ces travaux, lors même que les utilisateurs experts (issus des sciences de l'information) déplorent la piètre qualité de leurs performances ?

Il est un cercle d'activités où les approches automatiques sont inégalées par les traitements documentaires traditionnels : le traitement de documents voués à la disparition. On parle ici de documents qui échappent entièrement à la chaîne documentaire ; ils peuvent y échapper parce que leur durée de vie est trop courte, ou encore parce que l'on estime que leur valeur documentaire est trop faible, et donc ils ne mériteraient jamais d'y entrer.

Nous nous intéressons donc ici aux documents éphémères, qui disparaissent rapidement ou qui n'ont aucune « vie documentaire ».

Documents de courte vie

Quels sont ces documents « de courte vie » auxquels on fait référence ? Nous en examinerons quelques-uns ici.

Recherches ponctuelles sur le Web

On peut considérer comme des documents momentanés les réponses obtenues pour une requête envoyée à un moteur de recherche : chacune des réponses constitue en effet un très court document, dont la durée de vie n'est souvent que de quelques minutes et dont le contenu est dicté par les pratiques d'un moteur de recherche donné. Ainsi, un moteur comme Google présentera, avec chaque réponse, une description de la page visée (son URL, son format, la taille du document) ; ce sont des données extraites pour les besoins de l'affichage des réponses, qui sont disponibles lorsque l'on consulte le document mais auxquelles un butineur n'aura pas nécessairement accès, s'il parvient au site autrement que par Google (par un hyperlien d'une page amie, par exemple). On peut assimiler ces données aux descriptions de type catalographique, portant sur les propriétés physiques ou externes du document.

Également, chaque réponse de Google est accompagnée d'un court passage, qui se veut un résumé de la page repérée. Le « résumé » est en réalité un extrait de certaines parties du document ; il est construit à partir de règles internes à Google, qui assurent la présence d'au moins un des mots-clés de la requête. Il s'agit donc d'un résumé de type sélectif, selon le point de vue de la requête formulée. Ce type de résumé sélectif s'avère très utile : dès que l'on reconnaît la diversité des utilisateurs du Web, leurs intérêts variés et leurs besoins variables, on prend conscience de l'apport d'un tel résumé ciblé. Il permet de décrire chaque document selon le point de vue de l'utilisateur ; c'est une adaptation remarquable de cette technique automatique pour tenir compte des utilisateurs. Il faut noter cependant que l'extrait produit par Google est très limité par rapport aux possibilités que l'on entrevoit pour ces résumés sélectifs.

Le moteur de recherche ne fournit pas, par contre, un ensemble de termes d'indexation qui caractériserait l'ensemble du document (rappelons que la réponse de Google se base sur le fait qu'une partie du document, et non le document au complet, semble pertinent pour la requête). L'index complet de Google n'est pas disponible pour visionnement, et ne serait vraisemblablement d'aucun intérêt, puisqu'il est constitué de l'extraction exhaustive de (presque) tous les mots de chaque document. On peut envisager ici qu'il serait souhaitable d'obtenir, à l'instar du résumé

sélectif, un petit ensemble de termes qui caractérise la portion du document qui semble d'intérêt pour la requête de l'utilisateur.

D'autres moteurs de recherche ont constaté la nécessité de présenter à l'utilisateur un condensé de la page, et ceci est fait de différentes façons (voir le site d'Abondance[1], qui décrit les principaux moteurs de recherche et annuaires selon divers paramètres, dont la façon dont ils présentent les résumés). Dans plusieurs cas, le résumé présenté a été composé par un humain et décrit la page au complet (et non le segment pertinent pour la requête). Notons enfin que les annuaires thématiques (Yahoo!, La Toile du Québec[2]) recensent des sites Web, et non, isolément, les pages des sites. Ils proposent avec les réponses un résumé de chaque site, qui a été écrit au préalable par un humain. Il s'agit donc davantage d'un résumé de type indicatif ou informatif (selon son niveau d'exhaustivité par rapport au contenu du site Web).

Pour caractériser les réponses à une requête, on voit aussi le recours à des formes de classification automatique[3]. Les moteurs de recherche qui offrent cette option examinent les résultats obtenus pour une requête et, plutôt que de les présenter simplement par ordre systématique (de pertinence), ils les regroupent d'abord en classes. Le regroupement peut s'opérer sur la base des métadonnées rattachées aux pages Web, ou bien être effectuées sur la base d'une analyse du contenu des pages. Pour donner une idée : pour une requête sur « mercure », l'algorithme de classification automatique peut déceler trois ensembles différents de mots qui co-occurrent fréquemment : un ensemble contenant « élément » et « métal », un autre contenant « planète » et « système solaire », et un autre encore avec « dieu », « messager » et « commerce ». Ainsi, les résultats peuvent être présentés en grappes. Il est à noter que la classification est dynamique, c'est-à-dire qu'elle dépend de la requête qui peut contenir plusieurs mots et qu'ainsi la classification ne peut être établie une fois pour toutes. Voir aussi Vogel (2003) pour une autre conception de la classification dynamique.

Formulaires sur le Web

Un autre type de document de courte vie est en jeu lorsqu'un utilisateur remplit un formulaire sur une page Web (pour demander un produit ou un service de l'organisme). Le formulaire rempli sert à satisfaire à la demande du client. Il peut n'exister qu'un court instant, le temps de déclencher les actions appropriées (commande d'un logiciel, réquisition d'assurance, etc.) et ne jamais être versé dans une base de données. Or, les informations fournies dans le formulaire peuvent constituer des données stratégiques pour l'organisme en question. Une indexation

1. <http://www.abondance.com> [février 2007].
2. <http://www.yahoo.com> et <http://www.toileduquebec.com> [février 2007].
3. Par exemple : WiseNut <http://www.wisenut.com> ou MetaCrawler <http://www.metacrawler.com>. Teoma <http://www.teoma.com> fait des opérations similaires dont le produit est légèrement différent. Sites visités en février 2007.

automatique, ou la création d'un résumé sous une forme narrative, peut permettre de conserver l'information désirée.

Encans virtuels sur le Web

Considérons le cas des sites Web de types « encans virtuels » (on pense ici à des sites comme *e-Bay*[4]). On y mène des transactions commerciales foncièrement éphémères. Les informations impliquées dans les transactions peuvent présenter un intérêt pour divers groupes de personnes : non seulement les consommateurs, mais aussi des sociologues ou des économistes à l'affût de nouvelles tendances de consommation, par exemple. La seule façon de les capter, c'est de le faire sur le vif, alors que les documents en question sont toujours « vivants » et que les transactions sont en cours.

Affichage d'occasions d'affaires sur le Web

Certaines collectivités se servent du Web pour afficher des occasions d'affaires : offres de partenariat, appel d'offres, etc[5]. Ces occasions sont limitées dans le temps ; elles disparaissent au fur et à mesure qu'une offre trouve preneur. La présentation des offres adopte souvent un format tabulaire, qui pourrait être exploité en tant que base de données ; mais les descriptions du détail de l'occasion d'affaires contiennent habituellement une portion narrative, écrite en langue naturelle comme le français ou l'anglais, qui requiert un traitement plus sophistiqué d'analyse documentaire automatisée. Cette fois-ci, les informations qui s'y trouvent sont utiles dans le moment présent pour les entrepreneurs, mais elles peuvent aussi s'avérer très intéressantes pour étudier l'évolution d'un marché, la force d'un secteur de l'économie, etc.

Seuls les traitements automatiques permettent d'obtenir les résultats décrits ci-dessus : les résumés sélectifs, éventuellement l'indexation partielle d'un passage et la classification des résultats, tous selon la requête formulée par l'utilisateur ; l'extraction d'analyses de relations ou tendances à partir de documents éphémères ; etc.

Documents à contenu variable

Dans le monde numérique, certains documents sont en mouvance constante. Nous pensons ici, par exemple, aux pages contenant la une des journaux en ligne, aux nouveautés de sites établis ou aux annonces quotidiennes de sites commerciaux ou gouvernementaux. Souvent, la description de ces pages Web se résume à la mention « nouveautés ». Or il existe des contextes d'utilisation (le journalisme ou la veille par exemple) où l'on aimerait une description plus détaillée, et où il n'est pas toujours aisé de faire l'inventaire des sources quotidiennement à la main. Certains moteurs de

4. <http://www.ebay.com> [février 2007].
5. Exemple : Federal Business Opportunities, <http://www2.eps.gov/spg/index.html> [2004].

recherche offrent un service spécial consacré à ces pages de nouvelles qui changent souvent[6] ; le service est cependant limité à une liste de sites sélectionnés au préalable.

Il y a aussi les pages de type « journal de bord » (*weblogs* ou *blogs* ou encore blogues), phénomène récent, dans lesquelles les auteurs consignent leurs états d'âme du moment mais aussi leurs observations sur l'actualité. Et là aussi, l'indexation faite par les moteurs de recherche n'est pas assez fréquente pour suivre le rythme des parutions.

De par leur instabilité, il est impensable de penser mettre du temps à décrire ce type de documents de façon détaillée. Même leurs créateurs ne font que peu d'efforts pour leur assurer une description utile. Ils constituent donc un type de document pour lesquels les traitements automatiques peuvent devenir une solution.

Documents dont la valeur n'est pas validée

La majorité des pages Web « ordinaires » ne reçoivent pas de traitement documentaire élaboré. Si certains documents du Web sont traités à la source par leur créateur ou le distributeur, leur assurant une plus grande visibilité, nombre de documents demeurent sans traitement documentaire. Une description automatisée au moment de la requête ou du butinage peut s'avérer la seule option pour l'utilisateur en quête d'informations. Cela permet en outre à des communautés virtuelles de se créer, de se doter de documentation à partager, sans avoir à se constituer une structure formelle de gestion documentaire.

Pour sortir du cadre des moteurs de recherche sur le Web, regardons maintenant les archives de courriels présentes dans les organismes publics ou privés. Elles sont généralement considérées davantage utilitaires qu'informationnelles. Cependant, elles contiennent un trésor d'informations sur divers aspects des affaires traitées par l'organisme. On pourrait en extraire des analyses portant sur les associations entre émetteur, contenu et destinataire (en d'autres termes, *qui* envoie *quoi* à *qui*). Également, on pourrait y repérer des communautés tacites, basées sur les habitudes de communication. Ce type d'informations (et d'autres semblables) peut avoir une très grande valeur marchande ; elles sont donc d'un très grand intérêt pour les décideurs.

Pour tenter de réduire le volume des informations, il serait souhaitable aussi de procéder à un type de condensation (automatique) de leur contenu. Notons d'ailleurs que la classification ou catégorisation automatique de courriels est une fonctionnalité souvent recherchée par les sociétés dont les services aux clients ou aux utilisateurs utilisent beaucoup le courriel.

Pour leur part, les forums de discussion *Usenet* permettent à des groupes d'intérêt donnés d'afficher des messages à l'intention de leurs membres (voir par exemple les forums *Usenet* de

6. Par exemple, Google News <http://news.google.com/> [février 2007], NewsTrawler <http://www.newstrawler.com/nt/nt_home.html> [février 2007].

Google[7]) et ainsi d'échanger de l'information ou de l'expertise. Ces messages sont volatils. Ils représentent en quelque sorte les *blogs* des années 1980-1990, mais partagent plusieurs caractéristiques avec les collections de courriels. Ils contiennent notamment des masses d'informations qui pourraient être utiles, si seulement celles-ci étaient organisées, triées, classées, indexées ou résumées autrement que chronologiquement ou par fil de discussion. Ici aussi, un utilisateur aimerait pouvoir consulter ces données selon son point de vue, ou selon un angle donné ; des techniques de classification dynamique selon une requête ou un profil d'intérêt seraient tout à fait appropriées.

Collections éphémères

À la notion de document éphémère, dans le contexte d'utilisation du Web on peut ajouter celle de « collection éphémère » ; il s'agit ici de l'ensemble des réponses données pour une requête. En effet, il est possible pour un utilisateur de conserver un ensemble de résultats, auquel il peut vouloir revenir ultérieurement. La collection est doublement éphémère : elle est constituée de façon *ad hoc*, par rapport à une requête à un point temporel donné, et peut ne plus jamais être nécessaire en temps que collection ; mais aussi, elle est formée d'hyperliens vers des documents qui risquent constamment de disparaître.

Il est intéressant de noter que la collection de réponses peut devenir un objet d'étude en soi, même désincarnée des pages auxquelles elle fait référence. En effet, on peut commenter « la réponse donnée par Google à une requête donnée à une date précise », comme témoin du comportement du système ou bien du contenu du Web. C'est une notion de collection qui est tout à fait inusitée, si on la compare aux contextes traditionnels.

D'ailleurs, cette collection éphémère, stockée en tant que page HTML (page Web, donc) sur un site public, peut devenir à son tour une réponse à une requête ultérieure ; cet ensemble est impossible à reproduire, étant donné la volatilité des documents sur le Web.

D'autres applications utiles

Nous ajouterons ici d'autres utilisations possibles des traitements automatiques à celles que nous avons déjà évoquées (indexation automatique, condensation automatique, classification automatique).

Les recherches actuelles en condensation automatique de textes (voir Hovy, 2003) se penchent en grande partie sur la problématique de condensation multitextes : comment synthétiser les informations contenues dans un ensemble de documents. Les travaux ont porté beaucoup,

7. <http://www.google.com/grphp?hl=en&tab=wg&ie=UTF-8> [février 2007].

jusqu'à présent, sur des collections de dépêches de nouvelles, mais le besoin de créer des synthèses pour une collection quelconque existe partout. On voit ici l'exemple d'un outil (ou d'une technique) qui est spécifique à un genre. Les outils développés pour le genre journalistique ne sont pas efficaces pour traiter d'autres genres comme les courriels ou les appels d'offres. Or, la recherche et le développement de systèmes automatiques sont toujours dictés par le type de problème que l'on a tenté de résoudre. Ainsi, l'adaptation de systèmes existants à de nouveaux genres nécessite un travail non négligeable ; nous y reviendrons par la suite.

Une autre application possible de la condensation automatique serait de constituer rapidement, à partir de documents trouvés sur le Web, une biographie expresse d'une personnalité quelconque ; une telle possibilité intéresse les services de renseignements nationaux, pour leur permettre de profiter de tout document qui peut disparaître sans préavis ou pour pouvoir produire rapidement une description à jour des activités d'une personne ou d'un organisme surveillés. Bien sûr, il existe également de forts intérêts commerciaux pour ce type d'application : la possibilité d'obtenir un portrait rapide des activités d'une société privée ou encore d'un politicien peut aider à constituer un plan d'affaires. Comme les collections de documents utilisés pour les résumés multitextes peuvent varier à l'infini, seules les approches automatiques sont en mesure de rendre cet objectif commercialement viable.

L'indexation automatique peut prendre bien des formes autres que ce qui est implanté dans les index des moteurs de recherche. Par exemple, la tradition d'indexation fine des livres (en anglais, *back-of-the-book indexing*) peut inspirer l'élaboration d'outils permettant de naviguer à l'intérieur d'un document numérique assez volumineux (voir notamment Da Sylva, 2002, 2004a et 2004b). Cette indexation fine n'est pas envisageable pour toute une collection mais un utilisateur peut vouloir la produire sur demande, pour un nouveau document obtenu ; elle est donc tout indiquée pour le traitement de documents éphémères. La possibilité de faire ceci est absente des outils actuellement disponibles pour naviguer dans les documents du Web.

De plus, les analyses automatiques permettent d'extraire des informations tacites ou implicites contenues dans des documents existants (forage textuel) ; on peut donc en extraire des informations au-delà de ce que l'auteur avait inclus explicitement (on pense ici par exemple à l'analyse de textes en génétique afin de découvrir, par des analyses linguistico-statistiques, des corrélations entre enzymes, protéines, gènes, etc.). Le forage textuel est l'analogue textuel du *data mining* (forage de données) utilisé dans les grandes bases de données numériques ou factuelles. Cette application dépasse les possibilités de l'extraction d'informations (voir Appelt 1999 ; Humphreys *et alii* 2000), qui permet d'obtenir une description schématique du contenu d'un document – et qui est aussi une application qui nous intéresse ici. Ces types de traitements ne sont

pas du tout envisageables manuellement, vu la taille des bases de données constituées par ces énormes collections de documents.

Une autre application aussi, dont l'intérêt croisse avec le nombre de documents ajoutés sur le Web, est celle de la sélection automatique de documents : elle consiste à rechercher des documents répondant à un profil donné (idéalement, en tenant compte du contenu du document et non seulement de ses propriétés physiques externes) et à les ajouter alors à une collection existante (un annuaire thématique ou une bibliothèque numérique, par exemple). Même dans le cas où les résultats seraient validés par un humain, la possibilité de traitement automatique augmente la productivité potentielle d'une telle opération.

Avantages des traitements automatiques

Les traitements automatiques offrent plusieurs avantages, qu'il est utile ici d'opposer à leurs faiblesses citées ci-dessus, tout en les contextualisant.

Abordons d'abord, bien sûr, la question de la rapidité de traitement. Celle-ci permet de traiter des collections de taille monumentale. Il faut par contre comprendre que si les traitements simples sont effectivement très rapides même sur une collection énorme, certains traitements plus sophistiqués ne sont pas instantanés ; d'autres encore exigent des années de travail pour toute une équipe (par exemple, les systèmes de traduction automatique, de correction automatique, etc.) tout en n'arrivant jamais à un taux de succès de 100 %.

Ensuite, le traitement automatique, réalisé par un système unique, peut être décrit comme « centralisé » ; par conséquent, on peut s'attendre à une meilleure cohérence des résultats dans le temps. On gagne beaucoup ici. Une exception possible à cette règle serait les systèmes adaptatifs, qui peuvent « apprendre » d'après les habitudes de travail de l'utilisateur (et ainsi donner des résultats différents au fur et à mesure que le système s'adapte à celui-ci). Il ne sera pas possible alors d'atteindre une parfaite cohérence dans le temps ; ce comportement évolutif ne doit cependant pas être vu comme un désavantage, mais bien comme un atout – comme l'expérience acquise par un analyste chevronné.

Un autre facteur favorisant la cohérence est la mémoire infaillible des systèmes automatiques. Plus fidèle que la mémoire humaine, elle assure de pouvoir récupérer le résultat des traitements antérieurs, qui peuvent servir de guide ou modèle (on peut faire ici un parallèle avec les systèmes de traduction automatique qui utilisent une « mémoire de traduction » pour récupérer des traductions antérieures). Bien sûr, l'avantage offert par l'utilisation des mémoires suppose que les archives demeurent intactes et lisibles, ce qui s'avérera en fait un défi de taille.

Un avantage important est la tolérance qu'ont les systèmes automatiques à l'ennui dû à la répétition : cet aspect est bien souvent à l'origine de propositions de traitement automatique. En particulier, une tâche facile mais répétitive peut avantageusement être confiée au traitement automatique (par exemple, le repérage des noms propres dans un document), libérant l'expertise humaine pour les moments où elle est requise.

Enfin, relevons la transparence du traitement possible par un système automatique. On scrute plus facilement le code d'un système informatique que le cerveau et les connaissances des analystes humains (les expériences en intelligence artificielle des années 1980 sont arrivées à de telles conclusions). En principe on peut faire en sorte que le système permette d'examiner les étapes de traitement. En réalité, cette fonctionnalité n'est pas souvent incluse, mais il serait utile qu'elle le soit. Tout de même, c'est cette propriété des traitements automatiques qui est la plus séduisante. Elle permet de décrire explicitement les étapes du traitement et peut servir, à l'occasion, de guide méthodologique pour le traitement humain.

Un dernier avantage que nous relèverons pour les traitements automatiques, ce sont les nouvelles possibilités qu'ils amènent. Nous avons indiqué ci-dessus certains résultats qui ne peuvent être obtenus que par une analyse automatique, notamment la production de résumés ou de classifications qui seraient basés sur les besoins ou les requêtes des utilisateurs. Comme ces besoins varient d'un individu à l'autre, et pour le même individu dans le temps, les systèmes automatiques permettent de faire varier les résultats selon les besoins du moment. Cela s'applique bien sûr non seulement pour le traitement des documents éphémères, mais aussi pour d'autres documents et collections. Et c'est donc au point de vue de la flexibilité que les traitements automatiques se démarquent.

Nouvelles exigences

Quelles exigences doit-on imposer aux traitements automatiques ? En effet, si l'on veut défendre cette possibilité pour des documents éphémères, on doit leur assurer un certain niveau de qualité ou de performance.

Normalisation

La prolifération d'outils de traitements automatiques peut s'avérer plus problématique qu'utile, si les efforts réalisés de tous côtés produisent des résultats qui ne peuvent être exploités ultérieurement, qui ne peuvent être échangés entre systèmes. Dit autrement, ne pas chercher une normalisation dans les résultats des systèmes fait perdurer leur réputation de joueurs non disciplinés. Des efforts de normalisation des formats de données, en harmonie avec ceux du W3C

(*World Wide Web Consortium*[8]) sur les formats de documents (illustrés par exemple par le *Text Encoding Initiative*, Sperberg-McQueen et Burnard, 1995) et de métadonnées (par exemple, les initiatives liées aux métadonnées descriptives du *Dublin Core*[9] ou de *RD*[10]), sont à encourager.

Multilinguisme

Les systèmes de traitement automatique sont souvent développés pour traiter une seule langue. Par exemple, ils peuvent exiger des ressources linguistiques comme des dictionnaires, des thésaurus, des anti-dictionnaires, etc., ou alors ils peuvent être élaborés par des méthodes statistiques, suite à un entraînement sur des corpus d'une langue donnée. Ils sont donc limités à traiter cette seule langue (ou, éventuellement, un petit ensemble de langues). La mondialisation des ressources d'information entraîne des besoins de développement de systèmes multilingues (voir une discussion détaillée dans Da Sylva, 2003). Les critères d'évaluation pour les systèmes automatiques devraient donc inclure leur capacité à traiter plus d'une langue.

Langage documentaire

Il est important de prendre conscience du fait que, du point de vue du langage documentaire utilisé, les approches les plus faciles d'un point de vue automatique ont des limites importantes. Une indexation automatique faite avec les techniques de pointe peut être très performante, mais utilisera probablement le vocabulaire libre (par extraction directe d'expressions du document). On en connaît les inconvénients en gestion documentaire : rappel plus bas et précision trop élevée, dus à l'absence d'abstraction de la terminologie utilisée dans l'ouvrage, ce qui mène à une faible cohérence d'indexation. Le défi ici est donc de poursuivre les travaux pour doter ces systèmes de ressources permettant l'indexation par assignation de termes issus d'un vocabulaire contrôlé. Les divers travaux en construction automatique de thésaurus (voir notamment Bertrand-Gastaldy et Pagola 1992 ; Grefenstette 1994 ; Hearst, 1998 ; Sundblad, 2002) représentent un point de départ, mais sont confrontés à la tension entre thésaurus spécialisés (trop étroits mais précis), et thésaurus généraux (trop ambigus mais à portée plus large) ; ces derniers dépendent alors des avancées en désambiguïsation lexicale en contexte (voir Stevenson et Wilks, 2003), pour guider le système vers la bonne interprétation à donner aux mots ambigus ou polysémiques.

Au sujet des langages documentaires, regardons un cas problème, soit la terminologie mouvante des documents d'actualité. Le traitement automatique des dépêches de nouvelles est compliqué par le fait qu'un nombre important de néologismes apparaît sur une base quotidienne : des noms de personnes, d'organismes, de concepts émergents, d'innovations technologiques, etc.

8. <http://www.w3c.org> [février 2007].
9. <http://dublincore.org> [février 2007].
10. <http://www.w3.org/RDF/> [février 2007].

Une des pistes de recherche à poursuivre sera de développer des techniques d'analyse adaptées à ce vocabulaire constamment en évolution (voir entre autres Nadeau et Foster, 2004).

À ce même chapitre des langages documentaires, nous incluons la forme des résumés produits de façon automatique. Notons d'abord que plusieurs systèmes expérimentaux sont élaborés en prenant comme modèle des résumés d'auteurs, qui sont différents des résumés documentaires. Parmi les différences qui distinguent ces deux types de résumés, le plus important est l'absence de connaissances des principes d'organisation de l'information, auxquels nous reviendrons ci-dessous. Par ailleurs, il est encore aujourd'hui très difficile (voire impossible) d'élaborer des systèmes de condensation automatique qui saisissent réellement le sens des documents. Donc, les systèmes existants procèdent par heuristiques, la plus courante étant de créer non pas un condensé du document, mais un « extrait », produit par juxtaposition de phrases clés extraites du document. Les phrases clés sont sélectionnées d'après un ensemble de critères variés ; ces méthodes sont très ingénieuses, mais elles possèdent une limite importante, celle de réutiliser uniquement des phrases existantes (avec quelques bémols dont nous faisons abstraction ici). Il faut comprendre que le résumé résultant est très près du style de l'auteur, et que par conséquent la qualité du résumé produit est très variable d'un document à l'autre. Il est difficile, avec les techniques existantes, de voir comment résoudre ce problème ; cela demandera davantage de recherche dans la formulation finale des résumés.

Cet aspect de contrôle du vocabulaire (et celui de normes partagées, ci-dessus) ne pourra être incorporé aux systèmes de traitement automatique qu'en éduquant les concepteurs de ces systèmes au sujet des impacts de l'absence de normalisation et d'uniformisation.

Ces réflexions nous amènent à la conclusion suivante : que les critères d'évaluation pour les traitements automatiques, différents de ceux applicables pour les traitements manuels, doivent être dictés par une modélisation efficace des processus d'analyse documentaire en jeu.

Modélisation et principes d'organisation de l'information : pistes pour l'évaluation

Une bonne partie du cursus en science de l'information vise à inculquer aux apprenants ce que nous appellerons les « principes d'organisation de l'information ». Il s'agit de principes qui dictent quelles sont les propriétés d'un système d'information réussi. Parmi ces principes, on trouve celui de l'exactitude (*accuracy*), de la représentation suffisante et nécessaire, etc. On en trouve une description notamment dans Svenonius (2000, p. 68 et suivantes). L'ensemble de ces principes décrit les objectifs à atteindre (qui peuvent l'être plus ou moins, bien sûr, même dans un système d'information non automatique). Ces objectifs font abstraction de la méthode utilisée pour arriver

aux résultats, et ont ainsi le potentiel d'être exploités pour décrire le niveau de succès d'un système automatique.

Il existe un lien important entre les principes d'organisation et la notion de modélisation, si fondamentale à l'élaboration de systèmes informatiques. Il devient très intéressant d'explorer jusqu'à quel point l'activité de gestion documentaire peut être modélisée avec succès ; ceci exigera de puiser dans plusieurs champs disciplinaires, dont la linguistique, la psychologie cognitive, l'épistémologie, la communication, etc. Chose importante, une bonne modélisation peut produire des critères et des schèmes d'évaluation très pertinents.

Conservation des résultats

Enfin, plus spécifiquement, pour traiter les documents éphémères, une considération importante est celle de la conservation des résultats et des documents : les traitements automatiques trouveront plus facilement leur place s'ils permettent, en plus, de contrer la volatilité des documents visés. Certains documents éphémères, normalement voués à la disparition, pourront être considérés dignes de conservation par l'utilisateur qui les aura repérés : il faudrait donc fournir au sein même du système une façon de conserver le document, ou de stocker l'information extraite. Dans certains cas, le résultat du traitement pourra remplacer le document original, et c'est ce qui pourra être conservé à long terme. Cela créera donc des documents fantômes, des représentants de documents qui seront disparus. Bien sûr, alors, la qualité du traitement automatique devra être à la mesure de cette exigence de réutilisation du simple représentant.

Conclusion

Il apparaît que les traitements automatiques représentent une solution au traitement documentaire de documents éphémères, qui seraient normalement exclus de la chaîne documentaire. Ils leur permettent de connaître une certaine pérennité, si cela est désiré par un utilisateur ou un groupe d'utilisateurs. De plus, les traitements automatiques rendent possibles pour ces documents des traitements impossibles à atteindre manuellement.

Les exigences attendues de ces systèmes doivent être adaptées à la nature du traitement ; notamment, on prévoit que, à titre d'instrument d'évaluation, on aura besoin d'une solide modélisation du traitement documentaire, déterminée par les principes qui régissent l'organisation de l'information.

Bien sûr, à mesure que la qualité des résultats connaîtra des améliorations notables, on verra un intérêt à étendre l'application des traitements documentaires automatiques à des documents durables. Il existe d'ailleurs quelques cas (isolés, mais intéressants) d'introduction de traitements

automatiques dans des services d'information, où l'expertise humaine est réorientée. Pour permettre une amélioration dans les performances des systèmes de traitement documentaire automatique, il importe que les connaissances issues de la tradition bibliothéconomique et des sciences de l'information soient mises à contribution et déterminent les exigences des systèmes élaborés.

Bibliographie

AFNOR (Association française de normalisation), *Information et documentation : principes généraux pour l'indexation des documents*, Paris, Afnor, 1993.

AFNOR (Association française de normalisation), *Principes directeurs pour l'établissement des thésaurus multilingues*, Z47-101, Paris, Afnor, 1990.

ANDERSON, J.-D., « Standards for indexing : revising the American National Standard Guidelines Z39.4 », *Journal of the American Society for Information Science*, vol. 45, n°8, 1994, p. 628-36.

APPELT, D.-E., « Introduction to information extraction », *AI Communications*, vol. 12, n° 3, 1999, p. 161-172.

BERTRAND-GASTALDY, S. et PAGOLA, G., « L'analyse du contenu textuel en vue de la construction de thésaurus et de l'indexation assistées par ordinateur ; applications possibles avec SATO », *Documentation et bibliothèques*, vol. 38, n°2, 1992, p. 75-89.

COOLEY, R., PANG-NING, T. et SRIVASTAVA, J., « Discovery of interesting usage patterns from Web data », in *Proceedings of the WEBKDD Workshop*, 1999 (Lecture Notes in Computer Science, vol. 1836), Springer, Berlin, 2000, p. 163-182.

Da SYLVA, L., « Relations sémantiques pour l'indexation automatique. Définition d'objectifs pour la détection automatique », *Document numérique*, Numéro spécial « Fouille de textes et organisation de documents », vol. 8, n°3, 2004b, p. 135-155.

Da SYLVA, L., « Indexation automatique de documents par combinaison d'analyses statistiques et terminologiques structurées », *Actes du colloque RIAO'04*, Avignon, 26 au 28 avril, 2004a, p. 895-904.

Da SYLVA, L., « Technologies facilitatrices pour la dissémination de la communication scientifique multilingue », Communication dans le cadre du colloque *La communication scientifique en quatre dimensions*, 5 juin 2003, Montréal.

Da SYLVA, L., « Nouveaux horizons en indexation automatique de monographies », *Documentation et bibliothèques*, vol. 48, n°4, oct.-déc. 2002, p. 155-167.

GREFENSTETTE, G., *Explorations in automatic thesaurus discovery*, Dordrecht, Kluwer Academic Publishers, 1994.

HARMAN, D., *Overview of the 3rd Text Retrieval Conference (TREC-3)*, Washington, National Institute of Standards and Technology (NIST) Special Publication 500-225, US Government Printing Office, 1995.

HARMAN, D., *Overview of the 4th Text Retrieval Conference (TREC-4)*, Washington, National Institute of Standards and Technology (NIST) Special Publication 500-236, US Government Printing Office, 1996.

HEARST, M., « Automatic Acquisition of Hyponyms from Large Text Corpora », *Proceedings of the Fifteenth International Conference on Computational Linguistics*, Nantes, France, 1992, p. 539-545.

HOVY, E., « Text Summarization », in Mitkov, R., (éd.), The Oxford Handbook of Computational Sematics, Oxford ; New York, Oxford University Press, 2003, p. 583-598.

HUMPHREYS, K., DEMETRIOU, G. et GAIZAUSKAS, R., « Bioinformatics applications of information extraction from scientific journal articles », *Journal of Information Science*, vol. 26, n°2, 2000, p. 75-85.

LAPPIN, S., « Semantics », in Mitkov, R., (éd.), *The Oxford Handbook of Computational Sematics*, Oxford, New York, Oxford University Press, 2003, p. 91-111.

LEININGER, K., « Interindexer consistency in PsycINFO », *Journal of Librarianship and Information Science*, vol. 32, n° 1, 2000, p. 4-8.

MacDOUGALL, S., « Signposts on the Information Superhighway: Indexes and Access », *Journal of Internet Cataloguing*, vol. 2, n° 3/4, 2000, p. 61-79.

MUC, *Proceedings of the 6th Message Understanding Conference (MUC-6)*, San Mateo, California, Morgan Kaufmann, 1995.

MUC, *Proceedings of the 7th Message Understanding Conference (MUC-7)*, San Mateo, California, Morgan Kaufmann, 1998.

NADEAU, D. et FOSTER, G., « Real-Time Identification of Parallel Texts from Bilingual News Feed », *Proceedings of CLiNE'04*, Université Concordia, 30 août 2004, p. 21-28.

OVER, P., « Overview of DUC 2002 », *DUC 2002 Conference Proceedings*, National Institute of Standards and Technology, Gaithersburg, Md., 2002 ; <http://www-nlpir.nist.gov/projects/duc/pubs.html> [février 2007].

RAMLY, M.-L. et STROULIA, E., « Analysis of Web-usage behavior for focused Web sites: a case study », *Journal Of Software Maintenance And Evolution : Research And Practice,* vol. 16, n° 1-2, 2004, p. 129-150.

SAGGION, H. et LAPALME, G., « Selective Analysis for the Automatic Generation of Summaries », *Proceedings of the 6th ISKO Conference*, 10-13 July 2000, Toronto, Canada.

SAGGION, H. et LAPALME, G., « Generating Indicative-Informative Summaries with SumUM », *Computational Linguistics*, vol. 28, n°4, 2002, p. 497-526.

SCHLESINGER, J.-D., CONROY, J.-M., OKUROWSKI, M.E et O'LEARY, D.P. « Machine and human performance for single and multidocument summarization », *IEEE Intelligent Systems*, vol. 18, n°1, 2003, p. 46-54.

SPERBERG-MCQUEEN, C.M. et BURNARD, L. « The Design of the TEI Encoding Scheme », *Computers and the Humanities*, vol. 29, n°1, 1995, p. 17-39. Reproduit dans : Ide, N., Veronis, J., *The Text Encoding Initiative : Background and Contexts*, Boston, Dordrecht, Kluwer Academic Publishers, 1995.

STEVENSON, M. et WILKS, Y., « Word-Sense Disambiguation », in Mitkov, R. (ed.), *The Oxford Handbook of Computational Sematics*, Oxford ; New York, Oxford University Press, 2003, p. 249-265.

SUNDBLAD, H., « Automatic Acquisition of Hyponyms and Meronyms from Question Corpora », *Proceedings of the Workshop on Natural Language Processing and Machine Learning for Ontology Engineering (OLT'2002*), Lyon, France, 2002.

SVENONIUS, E., *The Intellectual foundation of information organization*, Cambridge, Mass., MIT Press, 2000.

VOGEL, C., « Using dynamic classification to expand individual/intuitive search processes », *Proceedings of International Conference on Integration of Knowledge Intensive Multi-Agent Systems (KIMAS)*, Boston, Sept. 30-Oct. 3, 2003, p. 523-528.