PROFILDOC

FILTRER UNE INFORMATION EXPLOITABLE

SYLVIE LAINÉ-CRUZEL

Laboratoire Recodoc
Université Claude Bernard
Lyon 1
slaine@univ-lyon1.fr

Dans un contexte où de plus en plus de textes intégraux sont ou seront accessibles en ligne, et où des moteurs de recherche indexent systématiquement le contenu des documents textuels, l'utilisateur qui lance une recherche s'expose au risque de découvrir une masse beaucoup trop importante de documents ayant un rapport avec sa question.

La question de la pertinence du sujet traité par les documents (représentation de son contenu) a fait l'objet d'une réflexion très approfondie dans toute l'histoire de la documentation. D'abord en tant que démarche intellectuelle (c'est tout l'objet des langages documentaires), puis en tant que démarche automatisée ou assistée. S'appuyant sur un simple repérage de chaînes de caractères, ou sur des méthodes plus sophistiquées incluant une analyse linguistique ou des analyses distributionnelles de formes, incluant ou non une représentation des connaissances (celle du système ou celle de l'utilisateur), de nombreux travaux s'attachent à améliorer le rappel et la précision du système, par rapport à une question fondamentale : comment ajuster au mieux la représentation du sujet traité par les documents, de manière à proposer à l'utilisateur, éventuellement en les triant, les documents qui sont le plus au cœur de la question qui l'intéresse.

Mais une autre question a été beaucoup moins approfondie : celle qui consiste à essayer d'évaluer si les documents seront exploitables par l'utilisateur. Or de nombreux facteurs peuvent contribuer à rendre un ensemble de documents inexploitables par rapport à une tâche donnée, ou un type d'attente donné, et en particulier le volume de documents récupérés.

Que faire si 500 documents sont potentiellement en rapport étroit avec le sujet de la question ? Sont-ils pour autant de même intérêt, seront-ils aussi efficaces les uns que les autres par rapport au but poursuivi par l'utilisateur ?

Il est clair que non. Il est évident que certains attendront des documents courts et synthétiques, d'autres préféreront des textes détaillés. Sur le même thème scientifique ou technique, on peut rechercher des discours polémiques ou de l'information objective, des documents pédagogiques ou professionnels. On peut rechercher des articles représentatifs d'un domaine que l'on connaît mal,

LE DOCUMENT SCIENTIFIQUE LIÉ À UN TRAVAIL DE RECHERCHE, ET PERMETTANT LA RECONNAISSANCE DE CE TRAVAIL PAR D'AUTRES CHERCHEURS, RESPECTERA UN PLAN TYPE DANS SA CONSTRUCTION

comme on peut rechercher des articles originaux et atypiques dans un domaine que l'on connaît très bien. On peut chercher des textes fondateurs, ou des articles récents.
Les critères qui permettraient donc de juger si un document est opérationnel ou non ne sont donc pas les mêmes que ceux qui permettraient de juger si le thème traité est celui qui intéresse l'utilisateur. Dans une démarche de recherche classique, les seules informations de ce type qui sont fournies par le système sont celles qui sont incorporées à la description bibliographique du document. Description qui tend d'ailleurs à se réduire, dans le cas où le texte électronique n'est pas l'image d'un document ayant par ailleurs suivi des circuits éditoriaux classiques.
Toute notre réflexion vise donc à définir, puis exploiter les informations qu'il est utile d'incorporer au système en complément du document lui-même. Cela a pour but de construire des systèmes d'accès à l'information permettant de restreindre les documents pertinents (du point de vue du sujet dont ils traitent), en limitant la réponse aux seuls documents exploitables et réellement utilisables.

Les prémisses : le découpage

L'une des caractéristiques évidentes liées aux documents scientifiques et techniques est leur forte structuration, implicite ou explicite. Pour être acceptés par la communauté qui va les exploiter, les différents types de documents respectent des contraintes de production précises. Le discours scientifique, contrairement au roman ou à l'essai, n'est pas linéaire. Un document lié à un travail de recherche, et permettant la reconnaissance de ce travail par d'autres chercheurs, respectera un plan type dans sa construction (plan IMRED, plan OPERA…), comportera un résumé, une bibliographie, éventuellement une brève présentation biobibliographique des auteurs… Un travail universitaire obéira à des règles de présentation (introduction, état de l'art, conclusion, annexes etc.), un rapport d'activité ou un manuel technique auront leurs propres règles. Un ouvrage pédagogique sera organisé en chapitres, comportera une introduction, parfois une préface et, éventuellement, un glossaire ou un index…
Dans un environnement électronique, la notion même de document devient mouvante et ambiguë. Si l'on tente de l'aborder par le biais du lien ou de la citation (le document étant alors « ce qui est référencé » dans un autre document), on trouvera une très grande diversité dans les pratiques. A une extrémité des possibles, on trouvera l'intégralité d'un site Web. A l'autre extrémité, on trouvera un paragraphe ou une occurrence précise d'un terme dans un texte.
Faut-il alors définir le document par la forme physique qu'il adopte dans le système ? Une page HTML pourrait être un document, mais un fichier contenant une image incluse dans la page visualisée pourrait, elle aussi, prétendre au statut de document.

Dans le projet Profildoc mené au sein de l'équipe Recodoc de l'université Lyon 1, l'exploitation ultérieure de l'information est notre fil conducteur. Le critère de découpage des documents est donc lié à l'usage (1).
Une enquête qualitative réalisée auprès d'un certain nombre de chercheurs et d'étudiants (2) a tenté de cerner les pratiques d'exploitation des articles scientifiques. Les principaux résultats qui en ont été dégagés sont les suivants :
– la lecture d'un article scientifique est rarement séquentielle, et souvent partielle ;
– les parties du texte auxquelles va s'intéresser le lecteur varient selon la tâche qu'il est lui-même en train d'effectuer ;
– la consultation de certaines parties du document permettra de décider de l'usage ultérieur de l'article (à exploiter ou non, à lire intégralement ou non…)
Enfin, l'enquête a permis de confirmer que les utilisateurs ont une très bonne connaissance des règles de production appliquées par les auteurs, et savent à l'avance qu'ils vont trouver dans un article scientifique l'exposé d'une démarche sous une forme normalisée. Le résumé, la table des matières, l'introduction, l'état de l'art, la discussion des résul-

LES UTILISATEURS SAVENT À L'AVANCE QU'ILS VONT TROUVER DANS UN ARTICLE SCIENTIFIQUE L'EXPOSÉ D'UNE DÉMARCHE SOUS UNE FORME NORMALISÉE

tats ou la bibliographie sont situés dans le document d'une manière qu'ils savent localiser. A chaque type de document correspond un ensemble de règles de production, même si cela n'a jamais été explicité.
La première partie de notre travail a consisté à définir des critères de découpage des documents, dont les principes généraux sont les suivants :
– une unité documentaire doit avoir une autonomie d'usage. Son exploitation doit pouvoir être significative, indépendamment des autres unités documentaires constitutives du document (contrainte d'autonomie) ;
– la taille d'une unité documentaire est comprise dans l'intervalle suivant : au minimum un paragraphe de texte, au maximum quelques pages écran (contrainte ergonomique) ;
– une unité doit être cohérente du point de vue du thème qu'elle traite et, pour cela, doit être identifiée en tenant compte le plus possible de la structuration choisie par l'auteur : chapitres, sous-chapitres, paragraphes… (contrainte de cohérence).
Chaque unité documentaire sera décrite par un *type* qui précise le rôle qu'elle joue au sein du document. A titre indicatif, nos types actuels dans le domaine des sciences expérimentales sont les suivants : *Résumé et mots-clés, Table des matières, Introduction, Description du contexte général, Description du thème, Description de la méthode, Environnement, Développement, Expérimentation, Résultats, Discussion, Conclusion, Bibliographie*.
Les premières analyses que nous avons commencé à mener sur d'autres secteurs scientifiques (et en particulier dans le domaine juridique[1]) font clairement apparaître que cette caractérisation est très fortement liée au domaine et devra être redéfinie dans chaque champ disciplinaire.

Types d'informations associées aux documents

Chaque unité documentaire (partie de document) est décrite par un ensemble de « propriétés » : des propriétés s'appliquant à l'ensemble du document, et des propriétés liées spécifiquement à l'unité documentaire.

Propriétés liées à l'ensemble du document

Certaines de ces propriétés sont des indications bibliographiques classiques (le titre du document, de la revue, les noms des auteurs, le pays et l'année de publication…).
Mais les propriétés les plus intéressantes du point de vue d'un filtrage des documents exploitables ne sont pas des titres de revues ou des noms de personnes. Ceux-ci ne sont significatifs que pour les utilisateurs qui ont une connaissance approfondie du domaine, des supports éditoriaux qui les diffusent et des communautés qui les produisent.
Or les nouvelles technologies diversifient l'accès à l'information, et permettent de sortir du cloisonnement disciplinaire qui a longtemps été caractéristique du travail scientifique. Pendant longtemps, les scientifiques n'ont pu découvrir le travail de leurs pairs que par l'accès à des index, à des bases de données thématiques, à des revues de sommaires ciblées. Toutes ces sources permettaient d'identifier la littérature produite par les chercheurs du même domaine.
Les nouveaux outils d'accès à l'information (serveurs hébergeant des bases très diverses, agents intelligents, moteurs de recherche) balaient un champ beaucoup plus vaste : à la fois interdisciplinaire, et aussi comprenant des productions dont l'étiquetage « scientifique » est plus ou moins reconnu et officiel. D'où l'augmentation du bruit dans les réponses, mais aussi une richesse extraordinaire : la possibilité de découvrir des travaux qui se situent hors du champ d'investigation auquel le chercheur avait jusque-là accès.
Pour tous ces nouveaux documents qui deviennent accessibles, le chercheur n'a plus les moyens de juger de la validité du texte, de son statut, du courant dans lequel il se situe, ni du mode de discours qu'il adopte. Plus précisément, sa grille d'analyse qui fonctionne efficacement et rapidement dans son propre champ disciplinaire n'est plus applicable. Il ne peut se construire une opinion qu'en lisant attentivement l'intégralité du document, opération longue et coûteuse au sens cognitif.

LES NOUVELLES TECHNOLOGIES DIVERSIFIENT L'ACCÈS À L'INFORMATION, ET PERMETTENT DE SORTIR DU CLOISONNEMENT DISCIPLINAIRE CARACTÉRISTIQUE DU TRAVAIL SCIENTIFIQUE

Il est donc de plus en plus indispensable que la caractérisation d'un certain nombre d'informations soit prise en charge en amont, lors de l'incorporation du document au système. Cette caractérisation peut se faire au travers d'une grille simple.
Dans le projet Profildoc, nous codons actuellement au niveau du document les indications suivantes (3) :
– type d'environnement éditorial (thèses ou mémoires, publications professionnelles, publications orientées grand public ou public averti, publications de recherche fondamentale, autres) ;
– profession de l'auteur (étudiant, spécialiste, journaliste ou médiateur, divers) ;

1. Mohammed GABSI, *Analyse et caractérisation des publications scientifiques spécialisées dans le domaine juridique : propositions pour une description dans le cadre du projet Profildoc*, Mémoire de DEA Sciences de l'information et de la communication, ENSSIB, 1998.

– champ disciplinaire de l'auteur (grands domaines) ;
– communauté de l'auteur (étudiant, universitaire, grand groupe industriel, PMI-PME, secteur public ou parapublic, individu ne se réclamant pas d'une communauté).

Propriétés liées à l'unité documentaire

Des indications propres à l'unité documentaire (partie de document) sont codées également :
– type d'unité (liste ci-dessus) ;
– forme discursive (descriptive, narrative, argumentative, discours rapporté) ;
– style : littéraire, textuel avec données numériques ou formalisées, à dominante factuelle ou numérique, formel numérique (formules de calcul ou équations), iconique (schémas, figures…), formel non numérique (symbolique, programmes, etc.).
Le repérage automatique d'un certain nombre de ces propriétés au moyen d'outils d'analyse de données textuelles fait l'objet d'une thèse en cours à Recodoc[2].

Recherche d'information du côté utilisateur

Dans le projet Profildoc, l'utilisateur doit fournir au système deux types d'informations complémentaires à propos de sa recherche.
Comme dans tout système classique, il doit préciser aussi clairement que possible le sujet qui l'intéresse. La fonction « recherche sur contenu » est réalisée dans notre prototype par le logiciel Spirit, qui trie les documents en langue naturelle selon leur proximité décroissante par rapport à la question de l'utilisateur.
Mais, par ailleurs, l'utilisateur doit donner au système des indications qui permettront de décider quels sont les documents, ou parties de documents, qu'il pourra exploiter dans le cadre de son travail en cours. Pour cela, deux formes de dialogue sont envisageables.
La première a pour principe de poser à l'utilisateur des questions auxquelles il peut répondre facilement, même s'il n'a pas réfléchi précisément à ce qu'il cherche. Qui est-il ? (culture générale, niveau de connaissances dans le domaine). Que veut-il ? (volume d'informations attendu, caractéristiques générales de cette information). Qu'en fera-t-il ? (nature du besoin ou de la tâche qui le conduit à rechercher de l'information).
À partir de ces données qui constituent un *profil* d'utilisateur (et qui peuvent varier pour un même utilisateur d'une recherche à une autre), le système va prendre des décisions sur les propriétés que doivent vérifier les unités documentaires, et sur la sélectivité du filtrage, qui laissera passer plus ou moins d'unités documentaires.
Mais un autre type de dialogue beaucoup plus direct est envisageable, car les propriétés associées aux documents sont directement exploitables par l'utilisateur s'il le souhaite, et s'il est capable de décrire précisément son besoin en termes de propriétés d'unités documentaires (forme ou style de discours attendu, domaine de compétence de l'auteur, communauté d'appartenance de l'auteur…). Dans les deux cas, l'auteur accède à un ensemble d'unités documentaires restreint et directement exploitable. À partir de ces fragments de documents bien ciblés, il va pouvoir entamer une seconde phase plus interactive de *navigation*. Cette navigation doit lui permettre de consulter facilement les parties de documents sélectionnées par des liens qui matérialisent ce qu'elles ont en commun, mais aussi d'avoir accès à l'intégralité du document lorsqu'il le souhaite[3].

DANS LE PROJET PROFILDOC LES PROPRIÉTÉS ASSOCIÉES AUX DOCUMENTS SONT DIRECTEMENT EXPLOITABLES PAR L'UTILISATEUR S'IL EST CAPABLE DE DÉCRIRE PRÉCISÉMENT SON BESOIN EN TERMES DE PROPRIÉTÉS D'UNITÉS DOCUMENTAIRES

Le rôle du spécialiste de l'information

Cette approche aurait, si elle devait se généraliser, un certain nombre de conséquences sur le rôle des spécialistes de l'information, et pourrait amener à un glissement dans leurs missions.
Depuis l'arrivée des premiers systèmes d'accès à l'information, une grande partie de l'activité des bibliothécaires et documentalistes s'est trouvée organisée autour de deux pôles :
– un rôle de médiateur entre les utilisateurs et le système (orientation, conseil, identification de sources ou formulation de requêtes) ;
– un rôle d'indexeur, analysant le contenu des documents et l'exprimant au moyen d'un langage documentaire.
Ces deux rôles traditionnels sont remis en question par l'évolution des nouveaux systèmes d'accès à l'infor-

2. Il s'agit de la thèse d'Éric GUINET, qu'il soutiendra en août 2000.

3. Ceci est l'objet de la thèse de Mohamed BEN ROMDHANE, qu'il soutiendra en août 2000.

mation (logiciels de GED, moteurs de recherche, agents intelligents). Les nouveaux systèmes sont de plus en plus conçus pour être utilisés directement par l'utilisateur final. De plus en plus, ils prennent en charge la fonction d'orientation et d'aide à la formulation de requêtes.

Le processus de représentation du contenu à partir de l'analyse du texte intégral, s'il a encore d'importants progrès à faire, est lui aussi de mieux en mieux appréhendé par les systèmes. Et surtout, on peut espérer que les différentes approches explorées par la recherche (analyse linguistique, méthodes quantitatives, approche cognitive ou pragmatique) permettront d'améliorer rapidement la pertinence[4] des systèmes d'accès.

Mais ces progrès prévisibles sont loin de résoudre toutes les difficultés. Lorsque les logiciels seront capables de ne fournir que les documents pertinents sur un sujet (et de les fournir *tous*), d'autres difficultés apparaîtront. Car de plus en plus de documents sont disponibles dans leur intégralité : à la fois les documents qui sont passés par des circuits éditoriaux, et des documents directement mis en ligne par leurs auteurs. Corollairement, la redondance et la répétition vont aller en augmentant. La précision, la fiabilité, la valeur officielle des textes mis en ligne sera de plus en plus hétérogène. L'usage pour lequel ils auront été initialement conçus et le public auquel ils s'adressent *doivent* apparaître de manière lisible, et exploitable par les outils d'accès à l'information. La forme en est toute trouvée : il ne peut s'agir que de métadonnées.

La forme que prendront ces métadonnées est loin d'être stabilisée, même lorsqu'un certain nombre de principes ont déjà été retenus.

Toutes sortes de critères peuvent être retenus lorsqu'il s'agit de donner une « information sur l'information ». Une réflexion orientée sur l'usage et l'exploitation de l'information peut guider la définition de ces métadonnées. Quels sont les critères que les utilisateurs voudraient exploiter ? Qualité de la source, valeur institutionnelle, précision de l'information, vraisemblance, degré de synthèse… et bien d'autres encore sans doute. La question est déjà au cœur d'un certain nombre de travaux (4, 5).

Faut-il abandonner ces jugements de valeur à des indicateurs quantitatifs rudimentaires ? L'*impact factor* ne concerne qu'une partie de la production scientifique. L'analyse de la fréquence de consultation ou du nombre de demandes est un indicateur contestable. N'est-ce pas précisément lorsqu'il s'agit d'analyser le public visé, le mode de discours adopté, son formalisme, le champ scientifique ou l'école de pensée auquel il se réfère, que nous avons le plus besoin de l'intelligence humaine ?

LES NOUVEAUX SYSTÈMES SONT DE PLUS EN PLUS CONÇUS POUR ÊTRE UTILISÉS DIRECTEMENT PAR L'UTILISATEUR FINAL

De nouveaux modes de collaboration sont à inventer, entre les communautés de chercheurs et les spécialistes de l'information, pour définir les critères qui permettront d'enrichir la description du document et d'améliorer la recherche en la restreignant à des documents exploitables dans le cadre d'une activité particulière. Pour les spécialistes de l'information, quel passionnant défi à relever !

Juin 1999

4. Pertinence dans le sens où les Anglo-Saxons utilisent le terme *relevance*, c'est-à-dire où on est capable d'optimiser le calcul d'une distance entre le thème exprimé dans une question et le thème traité dans les documents.

Bibliographie

1. Michel, Christine ; Lafouge, Thierry. – « Profil-doc : un système personnalisé de requête à des bases de données en texte intégral ». – *Actes du Congrès SFBA « Les systèmes d'information élaborée »*. – Ile Rousse, 12-16 mai 1997. – Publié sous forme de cédérom. – 1997.

2. Ben Abdallah, Nabil. – *Analyse et structuration de documents scientifiques pour un accès personnalisé à l'information utile : vers un système d'information évolué*. – Thèse soutenue à l'université Claude-Bernard-Lyon 1 le 7 juillet 1997.

3. Lainé-Cruzel, Sylvie ; Lafouge, Thierry , Lardy, Jean-Pierre ; Ben Abdallah, Nabil. – « Improving information retrieval by combining user profile and document segmentation ». – *Information Processing and Management*, 1994. – vol. 32, n° 3, p. 305-315.

4. Barry, L. Carol. – « A preliminary examination of clues to relevance criteria within document representations ». – *Proceedings of the 56th annual meeting of the American Society for Information Science*, 1993. – vol. 30, p. 81-86.

5. Wang, Peiling ; Soergel, Dagobert. – « Beyond topical relevance : document selecting behavior of real users of IR systems ». – *Proceedings of the 56th annual meeting of the American Society for Information Science*, 1993. – vol. 30, p. 87-92.