Texte de la 257e conférence de l'Université de tous les savoirs donnée le 13 septembre 2000.

LE WEB, DU TEXTE A LA CONNAISSANCE

par Sophie CLUET

Une bibliothèque universelle et anarchique

Le *World Wide Web* est un ensemble de documents auxquels on peut accéder par le réseau Internet. Ces documents sont souvent petits (on parle plutôt de pages que de documents) et s'entre-référencent, formant ainsi une structure rappelant une toile d'araignée (Web en anglais), dont les fils seraient des références et les nœuds des documents.

Créé par des chercheurs du CERN en 1989, le Web était à l'origine essentiellement utilisé par les différentes communautés scientifiques. Il a rapidement dépassé ce cadre et représente actuellement une banque de données considérable touchant tous les sujets et accessible à tous. Ainsi, on estime sa taille actuelle à un milliard de pages et on prévoit une multiplication par 100 d'ici les deux prochaines années. Si certaines pages ne peuvent être lues que par des abonnés, un grand nombre est gracieusement mis à la disposition de tous par des organismes dont la mission est d'informer, des sociétés pour qui le Web est un support de publicité ou de vente et des individus qui veulent communiquer leur savoir au monde.

Si la taille du Web croît de façon aussi spectaculaire, c'est qu'il y règne une grande liberté. N'importe qui peut facilement rajouter son nœud à la toile (même si certains pays tentent de légiférer). Cette liberté a un prix, le Web n'est pas organisé. Imaginez une immense bibliothèque où chacun ajouterait non pas des livres mais de simples feuillets, quand il veut, n'importe où et sans remplir de fiche descriptive. La seule façon d'accéder à une information dont on ignore l'emplacement consiste alors à parcourir tous les feuillets. A l'évidence, cette tâche est au-dessus des capacités d'un humain. Mais pas d'un ordinateur …

Des bibliothécaires dépassés

De fait, les bibliothécaires du Web sont des logiciels. Ils sont appelés moteurs de recherche (par exemple Voila, Altavista ou Google) et reposent sur des programmes dits *crawlers* ou *spiders* qui parcourent le Web en continu. Le parcours s'effectue en utilisant comme point de départ les adresses Internet de certains documents (c'est à dire, quelques nœuds de la toile) et en suivant les liens de référence entre les documents (c'est à dire, les fils de la toile). La collecte systématique de tous les documents autorise la construction d'index qui permettent de retrouver toutes les pages contenant un mot donné.

Cependant, un mot peut apparaître dans des milliers, voire des millions, de documents et les regarder tous peut nécessiter plusieurs jours de travail. La qualité d'un moteur de recherche va donc dépendre de sa capacité à filtrer les pages retournées. Plusieurs techniques sont utilisées à cet effet. L'une permet à l'utilisateur d'affiner sa recherche en combinant des mots de façon plus ou moins sophistiquée (par exemple, rechercher le mot « vitesse » à coté du mot « Puma »). Une autre consiste à privilégier dans les réponses certains documents « jugés » plus intéressants. Pour cela, différentes métriques sont utilisées, l'une d'entre elles prenant en compte le nombre de références vers un document. Enfin, la nouvelle génération de moteurs de recherche (par exemple

Askjeeves ou Zapper) utilise des techniques d'analyse du langage naturel (comprenez humain) et propose de rechercher des concepts plutôt que des mots (par exemple, le concept de vitesse peut s'exprimer de différentes façons : km/h, rapidité, etc.), ou encore des mots mais dans un contexte sémantique (par exemple, rechercher puma dans le contexte des sciences naturelles). Hélas, l'analyse informatique de la langue naturelle n'est pas encore parfaitement maîtrisée. Les techniques existantes sont adaptées au traitement de quelques documents sur un thème donné plutôt qu'à celui de milliards de pages couvrant une multitude de sujets. Quiconque utilise les moteurs de recherche sait que même les meilleurs sont souvent insatisfaisants, retournant des centaines de documents sans rapport avec l'objet de la recherche.

Les moteurs de recherche reposent sur une utilisation exclusive de l'informatique et, notamment, de l'intelligence artificielle. Les portails, une autre famille de logiciels de recherche sur le Web, privilégient l'intelligence humaine. Pour reprendre l'analogie avec le monde des livres, un portail peut être comparé à une encyclopédie, qui ne contiendrait que des informations trouvées sur le Web, que l'on ne pourrait pas feuilleter librement mais seulement consulter en suivant la table des matières. La plupart des portails traitent d'un domaine particulier (par exemple, le commerce pour Kelkoo), mais d'autres (comme par exemple Yahoo) ont la vocation de couvrir toutes les thématiques. La construction de chaque rubrique d'un portail repose sur des spécialistes qui (1) recherchent sur le Web les pages pertinentes, (2) analysent leur contenu et la façon dont elles sont organisées, (3) conçoivent la table des matières, et enfin (4) remplissent les pages de l'encyclopédie à partir de morceaux de texte sélectionnés. L'énormité de la tâche explique le fait que les portails sont incomplets. Une équipe d'hommes, même nombreuse et bénéficiant des meilleurs outils informatiques, n'est pas à même de suivre les modifications apportées quotidiennement au Web par des milliers de personnes à travers le monde. De plus, la table des matières qu'ils proposent n'est certainement pas adaptée à toutes les questions que l'on peut se poser. Certaines questions restent tout simplement in-formulables.

XML, le sens en plus

En résumé, s'il n'existe pas d'outil réellement satisfaisant pour rechercher des informations sur le Web c'est que les hommes ne sont pas assez rapides pour suivre son évolution ou les machines assez intelligentes pour comprendre le contenu des documents qui s'y trouvent. Tout indique que le Web va continuer à croître de façon exponentielle. Le progrès ne peut donc venir que des machines. En l'occurrence, il est déjà en marche.

Pour accéder au Web, une personne doit utiliser un logiciel appelé « navigateur ». Pour reprendre l'analogie avec la toile d'araignée, un navigateur permet de suivre les fils de la toile en « cliquant » sur les références présentes dans un document nœud. A chaque « clic », le navigateur va chercher un document et l'affiche. Ce que voit la personne est le résultat d'un travail d'affichage et est assez différent du texte récupéré par le navigateur. Actuellement, ce texte est au format HTML (pour HyperText Markup Language), un mode de représentation dont le défaut majeur est qu'il ne donne aucune aide aux logiciels d'analyse de contenu des documents.

Le *World Wide Web Consortium* (W3C), un consortium regroupant plus de 400 organismes (gouvernements, sociétés, laboratoires de recherche, etc.) et dont la mission est de promouvoir des standards pour une meilleure infrastructure du Web, propose l'adoption d'un nouveau format : XML (eXtensible Markup Language). Comme HTML, XML fait partie de la famille des langages informatiques à balises (en anglais *mark*). En d'autres termes, un document XML ou HTML contient du texte annoté par des balises. Cependant, alors que les balises sont

utilisées en HTML pour décrire la façon d'afficher le document, les balises XML donnent de l'information sur son contenu. La manière de présenter le document à l'écran est décrite séparément par une « feuille de style ».

Pour mieux comprendre cette différence fondamentale dans l'utilisation des balises, considérons la même information en HTML puis en XML.

L'auto-radio ARX24 est le dernier né de la série et remplace le ARX14. Bien qu'équipé d'enceintes d'une valeur de 590F<\i>, il ne coûte que 1200F

Le texte ci-dessus comprend quatre balises HTML de deux types différents. Il en existe de nombreuses autres. La première balise « » encadre le code de l'auto-radio, indiquant ainsi qu'il doit être affiché en caractères gras (« bold » en anglais, d'où le 'b' utilisé seul en ouverture de balise et préfixé d'une barre '/' en fermeture). Les balises «   » indiquent quant à elles un affichage en italique. Un simple coup d'œil suffit à un humain pour comprendre ce texte. Par contre, il est très probable qu'un logiciel d'analyse de textes quelconques l'interprétera mal. Il comprendra sans doute qu'il s'agit de la description d'un produit mais, par exemple, associera le prix « 590F » au produit de code « ARX24 » ou « ARX14 ».

Voyons maintenant une représentation XML possible de cette information (il en existe de nombreuses autres).

```
<produit>
        <référence> ARX24 </référence>
        <désignation> auto radio </désignation>
        <prix>
                <montant>1200 </montant>
                <unité>FF</unité>
        </prix>
        <description>
                Dernier né de la série, cet auto-radio est équipé d'enceintes
                d'une valeur de 590F.
        </description>
</produit>
```

Tout le texte est maintenant contenu dans une balise dont le nom est « produit », laquelle comprend d'autres balises, chacune englobant l'information relative à son nom. Clairement, il faut maintenant un peu plus de temps à un humain pour comprendre le texte. Mais rappelons que les formats HTML et XML ne sont pas destinés aux personnes mais aux logiciels. De fait, le sens du texte étant maintenant « balisé », ceux-ci ne vont avoir aucune difficulté à l'extraire et être ainsi à même de l'exploiter. Bien sûr, l'exemple choisi est volontairement simple. Le langage XML ne repose pas sur un ensemble pré-défini de balises mais donne au contraire aux programmeurs la possibilité de créer leur propre balisage. La même information pourrait ainsi être codée de façon un peu (ou beaucoup) moins lisible. Par exemple :

```
<prod ref=ARX24>
        <désiProd> auto radio </désiProd>
        <P > 1200F</P>
        <desc> Dernier né de la série, cet auto-radio est équipé
                d'enceintes d'une valeur de 590F.
        </desc>
</prod>
```

Il est important de comprendre que, bien que certainement porteur d'ordre, XML n'a pas pour vocation de brider la liberté qui règne sur le Web. Les utilisateurs qui lisent ou publient de l'information ne devraient pas voir de grande différence entre un Web HTML ou XML. Ce sont les logiciels qui s'adapteront, pas les hommes.

Xyleme ou le bibliothécaire idéal

Bien que soutenu par le W3C depuis déjà près de trois années, XML n'est hélas pas encore très présent sur le Web. A notre connaissance, il représente moins de cinq pour mille de l'information librement accessible. L'absence ou le peu de diffusion de logiciels adaptés à la navigation et à l'édition XML est sans aucun doute responsable de cet état de fait. Mais, ces logiciels sont en cours de réalisation et devraient faire leur apparition dans le courant 2001. Nous croyons qu'il se passera alors un véritable renversement en faveur d'XML. Nous, l'équipe VERSO de l'INRIA aidée de deux équipes amies des universités de Mannheim et de Paris-XI, misons sur cette révolution. Depuis un an, nous travaillons à l'élaboration d'un système nommé Xyleme qui utilisera les bonnes propriétés d'XML pour analyser finement les documents du Web, permettant ainsi une interrogation pertinente et précise de la masse de connaissances qu'ils contiennent.

Examinons les spécificités de Xyleme. Bien sûr, comme un moteur de recherche, Xyleme fait tourner des « crawlers » en continu pour collecter l'information. Également, il indexe les documents. Cette indexation est en fait assez différente de celle actuellement pratiquée par les autres systèmes dans la mesure où, comme nous allons le voir, Xyleme répond à des questions beaucoup plus élaborées que de simples recherches par mots-clé. Mais, l'expliquer nous forcerait à rentrer dans des détails assez techniques. Voyons plutôt trois autres particularités de Xyleme par rapport aux autres outils de recherche d'information sur le Web.

Stockage

Quelle que soit la question, les moteurs de recherche et portails donnent actuellement pour toute réponse une liste d'adresses de documents. L'utilisateur doit ensuite parcourir les pages correspondantes, à la recherche de l'information proprement dite. Celle-ci se trouvant rarement en un seul endroit, ce travail peut être long et, de fait, l'utilisateur abandonne souvent en cours de route. Xyleme se propose de retourner des réponses précises et synthétiques. À la question « quels sont les noms et adresses des grossistes parisiens spécialisés dans les jeux de société ? », Xyleme répondra par une liste contenant les noms et adresses des sociétés en question qui sont présentes sur le Web. Pour cela, il devra donc extraire de l'information provenant de différents documents. Ceci ne peut se faire en un temps raisonnable que si les documents sont présents dans le système. Xyleme stockera donc tous les documents XML du Web. Ce stockage permettra également de conserver l'historique de certains documents. Ainsi, le système sera à même de répondre à des questions telles que « combien de personnes ont été embauchées par la société X ces 12 derniers mois ? » ou encore « quelles sont les nouveautés de la semaine en littérature espagnole? ».

Analyse

Nous avons vu qu'un texte au format XML contenait des balises dont les noms étaient porteurs de sens. L'idée est d'utiliser ces balises pour poser des questions plus fines que de simples recherches par mot-clé et obtenir ainsi des réponses plus pertinentes. Imaginez un formulaire construit à partir des balises du document XML introduit précédemment. On voit bien comment des balises tels que « produit », « référence » et « prix » pourraient aider une personne à formuler une étude de prix sur, par exemple, un auto-radio et comment des documents balisés de manière similaire se prêteraient à la construction d'une réponse pertinente. Pour automatiser un tel mode d'interrogation, il reste cependant quelques difficultés à résoudre.

La plus importante est sans doute due au fait que deux documents traitant d'un même sujet peuvent contenir des balises très différentes. Ceci a été illustré précédemment. Il n'est évidemment pas concevable de donner à l'utilisateur autant de formulaires qu'il existe de balisages distincts. Xyleme va donc analyser les balises, comprendre leurs similitudes et proposer sa ou ses visions d'une thématique donnée. A l'échelle du Web, ceci peut paraître irréalisable. Plusieurs éléments font qu'il n'en est rien, même si la tâche reste extrêmement difficile, nécessitant parfois une intervention humaine restreinte. Tout d'abord, l'analyse ne porte pas sur des phrases d'une langue complexe mais sur des ensembles relativement petits de mots dont un logiciel peut, dans la majorité des cas, facilement comprendre le sens. Ceci est vrai même lorsque les mots sont abrégés comme « desc » ou agrégés comme « désiProd ». De plus, le balisage indique une certaine relation entre les mots. Par exemple, « prix » est contenu dans « produit » ce qui peut laisser entendre qu'il s'agit là d'un composant du produit. Enfin, XML nous apporte une aide considérable : les DTDs (pour « Document Type Déclaration »). Ainsi, si un document XML peut porter son propre balisage (on parle alors de documents bien-formés), il arrive plus souvent qu'il se déclare conforme à un balisage particulier décrit par une DTD (on parle alors de documents valides). Xyleme n'a donc pas à analyser tous les documents du Web, mais seulement toutes les DTDs, ce qui réduit considérablement la tâche. Ceci est vrai aujourd'hui et devrait l'être encore plus demain. En effet, des DTDs standards sont créées dans de nombreux domaines ce qui laissent présager une diminution de leur nombre et une stabilisation de leur contenu.

Une autre difficulté, plus légère, est liée à la multitude de thématiques présentes sur le Web. Ceci soulève essentiellement deux questions, comment reconnaître un document comme appartenant à un thème donné (et donc interrogeable par un formulaire donné) et comment ne pas noyer l'utilisateur dans une multitude de formulaires. Pour les raisons précédemment évoquées, associer thèmes et documents ne pose pas de problème insurmontable. Quant à l'ergonomie du système face à de trop nombreuses informations, une multitude de solutions existent. Par exemple, on peut personnaliser le système en fonction des intérêts d'un utilisateur particulier ou encore lui demander de formuler sa requête en langage naturel et utiliser certains des mots donnés pour sélectionner un formulaire et le remplir partiellement. Libre à l'utilisateur de préciser alors d'avantage sa question ou de la soumettre telle quelle au système.

Rappelons qu'à l'heure actuelle, les portails qui proposent un mode d'interrogation similaire ont été construits manuellement. Ils sont très spécialisés et ne couvrent qu'une faible portion du Web. En d'autres termes, il y a de fortes chances que vous ne trouviez pas le formulaire que vous recherchez quand vous en avez besoin et vous n'obtiendrez qu'une partie des réponses. De plus, ces systèmes reposent sur des formulaires fixes. Xyleme, parce qu'il repose sur des techniques d'analyse automatique des documents, permet la personnalisation des formulaires d'interrogation.

Dynamicité

Comme tous les moteurs de recherche, Xyleme est à l'écoute permanente du Web pour se maintenir à jour. Cependant, contrairement à ceux-ci, Xyleme analyse chaque nouveau document pour comprendre en quoi il peut intéresser certains de ses utilisateurs. Ainsi, une personne peut demander à être prévenue chaque fois qu'un nouveau site de vente par correspondance apparaît ou qu'un film de science fiction se joue dans le cinéma de son quartier. Ce type de service existe déjà sur le Web. Mais, encore une fois, les systèmes qui le proposent ont été construits manuellement et sont donc incomplets. Parce qu'il est capable d'analyser automatiquement et finement toute l'information du Web, Xyleme couvre tous les sujets de façon exhaustive.

Demain Xyleme ?

Notre année de recherche a abouti à la réalisation d'un prototype qui confirme la faisabilité technique du projet. Le succès de celui-ci dépend maintenant d'un certain nombre de facteurs, un des plus importants, et hélas, des moins maîtrisés, étant le succès que rencontrera XML sur le Web. Mais, rêvons un peu, imaginons notre pari réussi dans un futur riche de documents XML. Il y a 2250 ans, Ptolemée I^{er} chargeait Démétrios de rassembler toute la connaissance de l'humanité dans la bibliothèque d'Alexandrie. Xyleme est la version moderne de ce rêve. À la fois bibliothèque et bibliothécaire, Xyleme est universel. Constamment mis à jour, il réorganise ses rayons pour mieux vous satisfaire, répond pertinemment à vos questions les plus précises, vous informe dès qu'un événement que vous attendiez se produit. Il ne lui manque que la parole…