Technologies de l'information et de communication

Les TICs

Chapitre 2 : Initiation à la technologie de web

Partie 1 : Présentation de l'internet

1.1. Définition
1.2. Applications
1.3. terminologies
1.4. Réalisation d'un site statique

Partie 2 : La recherche sur le web

2.1. Outils de recherche
2.2. Affinage de la recherche
2.3. Autres outils de recherche

2. La recherche sur le Web

2.1 Outils de recherche

2.1.1 Les moteurs de recherche

Un **moteur de recherche** est **une application web** permettant à un utilisateur d'effectuer une **recherche en ligne** (ou **recherche internet**), c'est-à-dire de trouver des ressources à partir **d'une requête** composée **de termes**. Les ressources peuvent notamment être des **pages web**, **des articles** de **forums**, des **images**, des **vidéos**, des **fichiers**, **des ouvrages**, **des sites pédagogiques**, des **applications**, des **logiciels open source.**

Fonctionnement

Le fonctionnement d'un moteur de recherche se décompose en trois processus principaux :

1. **L'exploration** ou *crawl* : le web est systématiquement exploré par un robot d'indexation suivant récursivement tous les hyperliens qu'il trouve et récupérant les ressources jugées intéressantes. L'exploration est lancée depuis une ressource pivot, comme une page d'annuaire web.

   Un moteur de recherche est d'abord un outil d'indexation, c'est-à-dire qu'il dispose d'une technologie de collecte de documents à distance sur les sites Web, via un outil que l'on appelle robot ou bot.

   Un robot d'indexation dispose de sa propre signature (comme chaque navigateur web). Par exemple, Googlebot est le user agent (signature) du crawler de Google ; BingBot est celui de Bing ; AppleBot celui de Apple.

2. **L'indexation** des ressources récupérées consiste à extraire les mots considérés comme significatifs du corpus à explorer. Les mots extraits sont enregistrés dans une base de données organisée comme un gigantesque dictionnaire inverse ou, plus exactement, comme l'index terminologique d'un ouvrage, qui permet de retrouver rapidement dans quel chapitre de l'ouvrage se situe un terme significatif donné.

   Les termes non significatifs dans un contexte donné s'appellent des mots vides.

   Les termes significatifs sont associés à un *poids*. Celui-ci reflète à la fois la probabilité d'apparition du mot dans un document et le « pouvoir discriminant de ce mot » dans une langue, conformément au principe de la formule TF-IDF.

   L'indexation s'effectue après un ensemble de traitement syntaxiques et sémantiques (lemmatisation, catégorisation, suppression des mots "vide" ..) et algorithmiques selon différentes techniques basées aujourd'hui sur l'intelligence artificielle .

Le **TF-IDF** (de l'anglais *term frequency-inverse document frequency*) est une méthode de pondération souvent utilisée en recherche d'information et en particulier dans la fouille de textes. Cette mesure statistique permet d'évaluer l'importance d'un terme contenu dans un document, relativement à une collection ou un corpus. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document. Il varie également en fonction de la fréquence du mot dans le corpus. Des variantes de la formule originale sont souvent utilisées dans des moteurs de recherche pour apprécier la pertinence d'un document en fonction des critères de recherche de l'utilisateur.

3. **La recherche** : correspond à la partie ***requêtes*** du moteur, qui restitue les résultats.
Un algorithme est appliqué pour identifier dans le corpus documentaire (en utilisant l'index), les documents qui correspondent le mieux aux mots contenus dans la requête, afin de présenter les résultats des recherches par ordre de pertinence supposée.

   Les moteurs de recherche les plus simples se contentent de requêtes booléennes pour comparer les mots d'une requête avec ceux des documents. Mais cette méthode atteint vite ses limites sur des corpus volumineux.

   Les moteurs plus évolués sont basés sur le paradigme du modèle vectoriel : ils utilisent la formule TF-IDF pour mettre en relation le poids des mots dans une requête avec ceux contenus dans les documents. Cette formule est utilisée pour construire des vecteurs de mots, comparés dans un espace vectoriel, par une similarité cosinus.

   Pour améliorer encore les performances d'un moteur, il existe de nombreuses techniques, la plus connue étant celle du PageRank de Google qui permet de pondérer une mesure de cosinus en utilisant un indice de notoriété (célébrité) de pages.

   Les recherches les plus récentes utilisent la méthode dites d'analyse sémantique latente qui tente d'introduire l'idée de cooccurrences (présence simultanée d'une combinaison de mots) dans la recherche de résultats : (le terme « voiture » est automatiquement associé à ses mots proches tels que « garage » ou un nom de marque dans le critère de recherche).

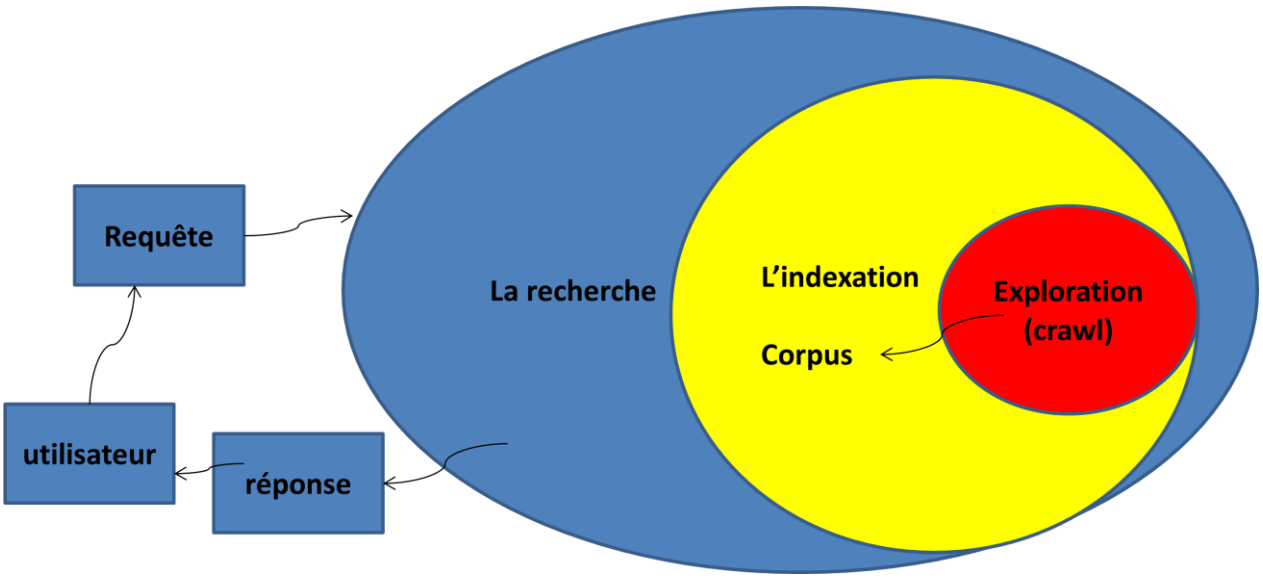

Fig 1 :Les Processus de Fonctionnement de moteurs de recherche

Des modules complémentaires sont souvent utilisés en association avec les trois briques de bases du moteur de recherche. Les plus connus sont les suivants :

2.1.2 Les répertoires

Un **annuaire web**, **répertoire web**, **annuaire Internet** ou **répertoire Internet** est un site web proposant une liste classée de sites Web.

Le classement se fait typiquement dans une arborescence de catégories, censée couvrir tout ou partie des centres d'intérêt des visiteurs. Chaque catégorie contient des sous-catégories concernant des aspects plus pointus d'un sujet donné et des hyperliens vers les sites agrémentés d'une description.

Types d'annuaires :

Un annuaire peut être généraliste, spécialisé (thématique) ou géographique :

- **les généralistes** n'excluent, *a priori*, aucun centre d'intérêt ;
- **les annuaires spécialisés et thématiques** se penchent exclusivement sur les sites ou les pages Web traitant d'un certain sujet, ou destinés à un certain public ;
- **les annuaires géographiques** enfin, peuvent à la fois se révéler généralistes ou spécialisés ; dans les deux cas, ils sont relatifs à un pays, une région, une localité.

2.1.3 Indexations automatiques

Motivation : L'indexation de données essaye de répondre à la question suivante : *Comment organiser au mieux une collection de documents afin de pouvoir plus tard retrouver facilement celui qui m'intéresse ?*

L'in**dexation automatique de documents** est un domaine de l'informatique et des sciences de l'information et des bibliothèques qui utilise des méthodes logicielles pour organiser un ensemble de documents et faciliter ultérieurement la recherche de contenu dans cette collection. La multiplicité des types de documents (textuels, audiovisuels, Web) donne lieu à des approches très différentes, notamment en termes de représentation des données. Elles reposent néanmoins sur un socle de théories communes, telles que l'extraction de caractéristiques, le partionnement de données (ou *clustering*), la quantification, et plus généralement la recherche d'information.

En revanche, les fichiers séquentiels indexés constituent une technique d'usage très général en informatique, pour le stockage de données numériques.

En représentant les documents sous forme de vecteurs **de descripteurs**, il devient possible de les comparer, de mesurer leurs distances les uns des autres, et de répondre à des requêtes de différentes natures.

Un index est, en toute généralité, une liste de descripteurs à chacun desquels est associée une liste des documents et/ou parties de documents auxquels ce descripteur renvoie. Ce renvoi peut être pondéré. Lors de la recherche d'information d'un usager, le système rapprochera la demande de l'index pour établir une liste de réponses. En amont, les méthodes utilisées pour constituer automatiquement un index pour un ensemble de documents varient considérablement avec la nature des contenus documentaires à indexer.

2.1.4 Les navigateurs

Un **navigateur web**, **fureteur**, est un logiciel conçu pour consulter et afficher le World Wide Web. Techniquement, c'est au minimum un client HTTP.

Il existe de nombreux navigateurs web, pour toutes sortes de matériels (ordinateur personnel, tablette tactile, téléphones mobiles, etc.) et pour différents systèmes d'exploitation (GNU/Linux, Windows, Mac OS, iOS et Android). Dans les années 2010, les plus utilisés sont Google Chrome, Mozilla Firefox, Internet Explorer/Edge, Safari, Opera.

Exemples de navigateurs
Article connexe : Liste de navigateurs web.

Il existe des versions différentes des navigateurs selon le type de machines : téléphone mobile, tablette ou ordinateur. Voici pour les navigateurs les plus populaires les noms de leur version plate-forme bureautique et mobile.

- Android Browser (mobile uniquement)
- Apple Safari, Safari Mobile
- Google Chrome, Chromium, Chrome Mobile
    - Brave (Navigateur visant la protection de la vie privée)
- Microsoft Internet Explorer et Microsoft Internet Explorer Mobile, depuis remplacé par Microsoft Edge
- Mozilla Firefox, Firefox Mobile
- Opera, Opera Mini
    - Vivaldi (émanation d'Opera)
    - OperaGX (Navigateur pour les gamers)
- Lynx

2.2 Affinage de la recherche

Une requête est l'expression du besoin en informations de l'utilisateur. Elle présente l'interface entre le système de recherche de l'information **SRI** et l'utilisateur. A cet effet, divers types de langages d'interrogation sont proposés dans la littérature.

Une requête est un ensemble de mots clés, mais elle peut être exprimée en langage :

- Naturel ou quasi naturel
- Booléen
- Ou graphique à partir d'une interface graphique

2.2.1 Choix des mots clés

Un **mot clé** (orthographié aussi **mot-clé**, **mot clef** ou **mot-clef**) est un mot ou un groupe de mots utilisé pour caractériser le contenu d'un document et permettre une recherche d'informations. Une liste de mots-clés permet ainsi de préciser les thématiques du document.

Dans le cadre de la recherche d'informations, les termes de recherche sont autant que possible convertis en mots-clés au moyen d'un **thésaurus documentaire** correspondant à la manière dont sont indexés les documents.

Un **thésaurus**, **thésaurus de descripteurs** ou **thésaurus documentaire**, est une liste organisée de termes contrôlés et normalisés (descripteurs et non descripteurs) représentant les concepts d'un domaine de la connaissance
C'est un langage contrôlé utilisé pour l'indexation de documents et la recherche de ressources documentaires dans des applications informatiques spécialisées. Les thésaurus sont donc une catégorie de langages documentaires parmi d'autres. Les termes (dans l'exemple ci-contre : *véhicule*, *navire*, etc.) sont reliés entre eux par des relations de synonymie (terme équivalent), de hiérarchie (terme générique et terme spécifique) et d'association (terme associé) ; chaque terme appartient à une catégorie ou domaine.
Par exemple un thésaurus reliant *récolte à culture*, *blé* à *céréale*, et *France* à *Europe*, permettra pour une question portant sur la *récolte* du *blé* en *France* de trouver des ressources indexées avec *culture céréale Europe*.

Pour qu'un mot-clé soit pertinent et fonctionne, il faut qu'il soit reconnu par un outil de recherche dans un index déjà constitué.

2.2.2 Opérateurs booléens

Un système de recherche d'information intègre un ensemble de modèles pour la représentation des unités d'information (documents et requêtes) ; parmi ces modèles on distingue **l'approche ensembliste ou booléenne** : c'est l'un **des premiers modèles introduit en 1983 utilisés en recherche d'information**, qui offre une représentation mathématique simple du contenu d'un document selon l'approche ensembliste

Une requête q est représentée par une expression booléenne dont les termes sont reliés par des opérateurs logiques selon le formalisme de l'algèbre de Boole : (le produit logique AND 'Λ', la somme logique OR :'V' et la différence logique NOT : '¬') permettant d'effectuer des opérations d'intersection, d'union, et de différence entre les ensembles de résultats associés à chaque terme.

Un exemple de représentation d'une requête est comme suit : $q =$
$(t1 \wedge t2) \vee (t3 \wedge \neg t4)$ $ou\ ti\ sont\ les\ termes\ de\ la\ requete$ .Pour qu'un document corresponde à une requête, il faut que l'implication suivante soit valide $d_i \Longrightarrow q$
Un document du corpus est ainsi considéré comme pertinent uniquement quand son contenu est vrai pour l'expression de la requête ; sinon il est considéré non pertinent

Si les mots-clés ne sont pas reliés par des opérateurs, la plupart des outils de recherche considèrent que, par défaut, les termes sont reliés par un « ET ».

2.2.3 Les opérateurs de proximité, L'adjacence, la troncature

- **La recherche de proximité exacte** se fait en utilisant des guillemets. Ces derniers encadrent une expression ou une suite de mots clés que l'on souhaite trouver ainsi, écrits dans cet ordre.

Exemple:
La requête q= " ***recherche d'informations sur Internet*** " donnera des documents qui comportent tous ces mots clés, dans cet ordre, et sans aucun autre mot entre ces derniers

- **L'adjacence** permet d'indiquer la position réciproque des termes de la recherche. Elle s'écrit, selon les outils : ***NEAR*** ou ***ADJ***

Exemple:
La requête : q=***effet NEAR serre*** donnera des documents sur ***l'effet de serre*** ou des documents sur ***la serre à effet***

- **La troncature** Marquée par le signe * placé à la fin d'un mot, elle permet d'élargir la recherche à tous les mots qui ont une racine commune.

Exemple:
La requête ***capit**** donnera comme résultat de recherche : capitale, capitaine, capitainerie, etc.

2.3 Autres outils de recherche

Métamoteurs de recherche

Les métamoteurs sont des outils qui permettent d'effectuer une recherche à l'aide de plusieurs moteurs simultanément. On connaît entre autres Copernic et Meta-Crawler. S'ils font gagner du temps, leurs résultats ne seront pas toujours utiles, surtout dans le cas de recherches très précises.En effet, les métamoteurs n'utilisent pas les moteurs au mieux de leurs capacités.

Bien souvent, les interrogations ne se font qu'en mode simple et ne s'adaptent pas aux particularités de chaque moteur. Une requête complexe soumise à un métamoteur risque de ne pas toujours être comprise par les moteurs qu'ils mettent à contribution.

Ces outils sont toutefois utiles pour effectuer un tour d'horizon très général d'un sujet et observer comment chaque moteur réagit à la requête en vertu de ses propres caractéristiques.

Référence

[1] https://fr.wikipedia.org/

[2] « Technologies et architectures internet » Pierre-Yves Cloux ; David Doussot ; Aurélien Géron 2ème édition mai 2002