Méthodologie de recherche et interopérabilité des outils

Marième Iba SARR
Jacky BAYILI
Pierre - Yves DUSSARAT

Méthodologie de recherche

Introduction

Processus de recherche

- Définition du Mot Clé
- Quelles aides au choix des mots clés ?
- Les moteurs de recherches

Les nouveaux outils d'aide à la recherche

- La catégorisation des résultats
- Les réseaux sémantiques

Tableau syntaxique des moteurs de recherche

- Analyse et bilan

Mise en application par un exemple concret : Comment cerner et distinguer l'image d'une structure sociale sur le Web ?

Introduction

Conscients de la masse informationnelle considérable disponible en permanence grâce aux technologies de l'information et de la communication, il semble nécessaire de se forger une « culture de la recherche » suffisante pour ne pas être victime de cette surabondance.

Dans la considération de l'outil Internet, comme une immense bibliothèque de savoirs et d'informations, l'utilisation de mots clés pour s'immerger dans cette masse de données semble une base de départ relativement fiable pour permettre à chacun de commencer à faire un tri et à sélectionner son information en fonction de ses besoins.

L'ampleur quantitative des ressources emmagasinées sur le réseau contribue davantage à un faible niveau qualitatif, qualifiable de « bruits informationnels », qu'à une évolution en matière de libération de l'information.
Bien que de nombreux outils servent de supports, en vue de classifier les données et de sélectionner au mieux les plus conformes à nos attentes, il va sans dire qu'un apprentissage de ces techniques et la segmentation de l'information restent des domaines à définir avant de s'immerger dans la sphère des « Ntic ».
Qu'il s'agisse d'ordre purement privé, ou professionnel, il est important de resituer la typologie de l'information, lors de la recherche, pour juger de l'exploitabilité des données recueillies.

Information	Blanche	Grise	Noire
Type	Scientifique, technologique, commercial, juridique, financier, stratégique, personnel		
Niveau	Tactique, opérationnel, stratégique		
Domaine opératoire	Documentaire, de situation, d'alerte		
Intérêt	Fatal, utile, pertinent	Pertinent, critique	Critique
Accès	Public	Restreint	Strictement limité
Classification	Non protégée	Protection restreinte	Confidentielle - Secrète
Disponibilité	80 %	15 %	5 %
Acquisition - Exploitation	Légale sous réserve de respecter les droits de propriété.	Domaine juridique non clairement défini. Risques d'ordre jurisprudentiel.	Illégale, l'acquisition relève de l'espionnage. Risques très élevés.
Forme	Formelle (texte) ou informelle (conversation, rumeur)		
Sources	Ouvertes	Autorisées - Fermées	Clandestines
Coût	Faible	Faible	Élevé
Rentabilité	Élevée	Très élevée	Faible

On considérera, dans le cadre d'une recherche conventionnelle et purement légale, l'exploitation de résultats basés sur la collecte d'informations blanches et grises déclinables sur différents types de pages et dont les sources seront plus ou moins fiables.

L'objet de cet article est de donner les moyens à tout à chacun de comprendre les données reçues et de savoir faire un tri rationnel en fonction de l'analyse de la validité des sources ainsi que de nombreux autres facteurs.

Processus de recherche

Définition du mot clé :

Mot caractérisant le contenu d'un document ou d'un fichier, servant de principal critère de recherche dans un fichier ou dans un système de gestion de base de donnée.

Initialement défini comme la donnée utilisée dans les balises d'une page html pour en définir le contenu, le mot clé a vu sa fonction élargie à la notion de recherche sur Internet.Il constitue ainsi à la fois les informations constituées dans le code primaire de la page (html) mais aussi dans la page elle-même (texte).

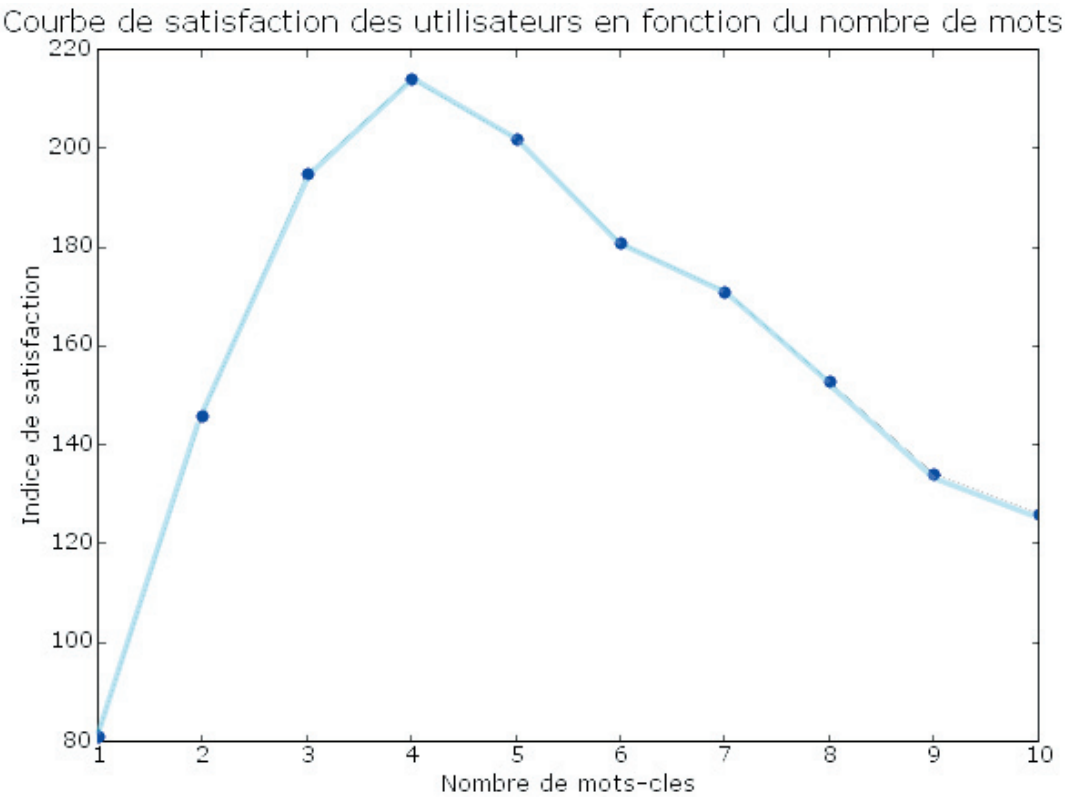
Ces mots clés sont ainsi générés par les usagers de l'outil Internet et correspondent la plupart du temps à des termes issus du langage et ne suivant pas un modèle généralisé et adopté par tous.L'internaute conçoit ses mots clés, et analyse de la pertinence de son choix en corrélation avec les réponses obtenues.On pourra ainsi parler de démarche par « tâtonnement », permettant un affinage constant au fil des mots clés entrés.

La pertinence des mots clés est étroitement liée, d'une part à leur rareté, ou caractère temporaire et éphémère qu'ils représentent.
En effet, plus un mot est commun, employé au quotidien, plus l'information qu'il générera sera massive avec une source de bruit beaucoup trop élevé pour être utilisable telle quelle.
Commence ainsi un mécanisme de réflexion pour faire émerger un terme à la fois pertinent et peu utilisé pour permettre d'obtenir un contexte d'utilisation d'information favorable.
D'autre part la notion de temporalité peut être utile dans la recherche de pages relatives à des personnalités ou événementiels, dont l'actualité médiatique est éphémère.
La valeur des mots clés, en terme de finesse des résultats, s'articule donc autour de ces axes :

- l'entrée de noms de lieux ou d'évènements
- l'entrée de noms de personnes
- l'entrée de noms relatifs à des entreprises, groupes ou marques

- l'entrée de mots spécifiques touchant une activité précise et définie
- l'entrée de mots rares

D'autre part, l'addition de mots clés dans une recherche est généralement un facteur supplémentaire permettant l'affinage des solutions et l'évolution dans l'environnement souhaité.
Plusieurs études ont permis de mettre en avant l'efficacité des recherches comparativement aux nombres de mots clés juxtaposés entrés lors de la recherche.



Quelles aides au choix des mots clés ?

Assigné à la recherche de termes qui nous sont éventuellement inconnus, dans le cadre d'un travail spécifique, il est important de faire appel à un expert en la matière qui pourra nous guider sur des pistes et thématiques en adéquation avec le travail en question.
Le partage d'informations ainsi véhiculées permettra de s'orienter vers de nouveaux termes plus pertinents pour des résultats probants.

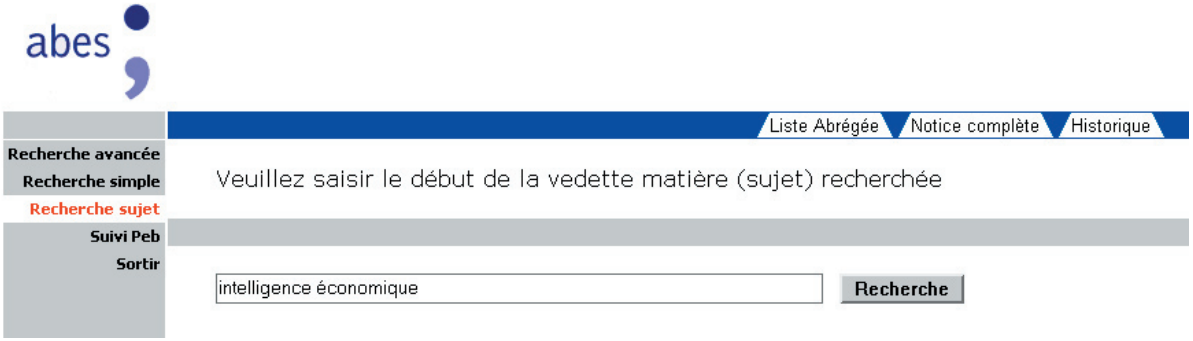
D'autre part, l'orientation vers un langage spécifique appelé RAMEAU **(**Répertoire d'Autorité Matière Encyclopédique et Alphabétique Unifié**)** permet d'étoffer la recherche en s'ouvrant à un champ plus large de termes et définitions en relation avec les mots clés déjà définis.

Exemple de procédure de recherche :

. Lors de recherches dans un catalogue (ex. **SUDOC**), choisissez une référence pertinente et repérez les autres mots-clés dans les notices bibliographiques.

2. Vous pouvez aussi utiliser le catalogue de la BNF, BN- opale plus, http://www.bnf.fr

Tapez votre mot-clé dans la rubrique « sujet » et regardez son environnement qui peut vous aider à affiner votre recherche.



Système universitaire de documentation

Système universitaire de documentation

La recherche de mots clés en langue anglaise peut être issue d'une traduction faite par le site de la BNF (http:// noticesrameau.bnf.fr/), en suivant les consignes suivantes :

- Choisir la recherche par feuillage.
- Taper un terme dans la rubrique « début de la vedette », valider.
- Cliquer sur le terme le plus approprié dans la liste proposée.
- Vers le bas de la notice apparaît l'équivalent en anglais, noté Equiv. LCSH.

Les moteurs de recherche

Définis comme étant des sites ou logiciels capables d'assister chaque internaute dans le classement de l'information sur Internet, les moteurs se chargent en effet de sélectionner et afficher des sites, grâce aux moyens de robots allant eux mêmes visiter ces pages de manière autonome. Cette méthode, communément identifiable sous l'appellation de « crawl » va donc permettre de mettre en place un index chargé de plusieurs milliards de pages.

Du point de vue de l'internaute, son utilisation est largement accessible. En effet, une fois les mots clés définis au préalable, il suffit d'entrer ces termes dans le champ de recherche pour que le moteur s'occuper d'indiquer les pages en corrélation avec la demande. Plusieurs degrés de recherche sont ainsi possibles, permettant un affinage efficace des recherches et un amoindrissement du bruit.

Plusieurs sources nous ont permis de restreindre notre recherche aux dix moteurs générant les trafics les plus importants.

TOP 10 (Septembre 2006)

Outils de recherche	% de trafic généré	Tendance
1 - Google	87.13 %	(+0.9)
2 - Yahoo!	4.32 %	(-0.19)
3 - Voilà	2.86 %	(-0.25)
4 - Live	2.14 %	(-0.57)
5 - Free	0.94 %	(+0.03)
6 - AOL	0.87 %	(-0.06)
7 - Alice	0.44 %	(+0.03)
8 - Club Internet	0.44 %	(-0.02)
9 - Altavista	0.23 %	(+0.01)
10 - Lycos	0.10 %	(-0.01)

Légende :
Stabilité du % du trafic ou changement inférieur à +/-0,1% par rapport au mois précédent.
Baisse supérieure ou égale à 0,1% par rapport au mois précédent.
Hausse supérieure ou égale à 0,1% par rapport au mois précédent.

Statistiques délivrées par Xiti et 1° position

De l'étude de ces moteurs, nous avons la encore plus segmenter nos résultats, de manière à obtenir des « familles » de moteurs s'appuyant sur les mêmes concepts et technologies pour l'affichage de résultats.

- Yahoo (YST) fournit notamment les résultats de Lycos, Altavista, Alltheweb ou Hotbot
- MSN Search fournit les résultats de MSN ou Live.com

- Voila fournit les résultats du portail Orange
- Exalead fournit les résultats de AOL (en France)

Les nouveaux outils d'aide à la recherche :

De nombreuses innovations sont apparues depuis trois ans dans le domaine du traitement et de la présentation des résultats des outils de recherche : la catégorisation des résultats (clusterisation), les réseaux sémantiques (cartographie).

La catégorisation des résultats

Cette technologie collecte les documents par les biais d'autres moteurs de recherche et les organise en une hiérarchie significative de dossiers.
L'innovation de cet outil est qu'il permet de rassembler les documents par aires sémantiques.
Il permet donc d'affiner les requêtes, de suggérer de nouvelles pistes de recherche, de nouveaux liens vers d'autres thèmes.
En bref, les technologies de catégorisation des résultats amènent du sens, de la signification, de la structuration et sont appelées, d'une certaine manière, à jouer le même rôle que les thésaurus classiques.

Les réseaux sémantiques (cartographie)

Une nouvelle manière de présenter les résultats vient d'être développée par des métas moteurs (voir tableau), sous forme de cartes, de réseaux sémantiques, calculés à partir des liens sémantiques entre les pages Web.

Au lieu de classer les documents à des catégories thématiques, les pages Web sont reliées les unes aux autres, en fonction des mots-clés communs qu'elles partagent. Les résultats sont présentés graphiquement, sous forme de nœuds et de liens : les nœuds, qui correspondent aux pages Web trouvées; les liens entre les nœuds représentent les relations entre les pages Web.

Cette technologie permet d'affiner les requêtes (par choix de mots-clés), de visualiser des liens entre sites Web que l'on n'aurait pas pensé à associer, d'élargir les recherches sur les sites proches, de mettre en évidence des réseaux d'acteurs sur telle ou telle thématique, avec des indications précieuses sur l'importance de tel ou tel site (par le nombre de liens qu'il reçoit)
En définitive, ces métas moteurs graphiques développent une nouvelle cartographie de l'information permettant une meilleure représentation.

TABLEAU des moteurs

Analyse tableau

La mise en place d'un tableau comparatif nous a permis de voir les différences et spécificités syntaxiques des moteurs jugés les plus fonctionnels et pertinents.

Bien que les requêtes basiques soient construites de manières similaires pour chaque moteur, d'autres, utilisées de manière à affiner une recherche, n'ont pas la même structure. Il est donc assez délicat de parler d'interopérabilité entre ces outils, étant conscients qu'ils ne le sont jamais complètement.
D'autre part de nombreuses variances ont pu être constatées pour une même requête sur les différents moteurs mettant en exergue une complémentarité de l'information.Nous avons en effet choisi d'afficher le nombre de résultats obtenus par requête afin de montrer la pluralité des résultats obtenus selon les différents moteurs questionnés.

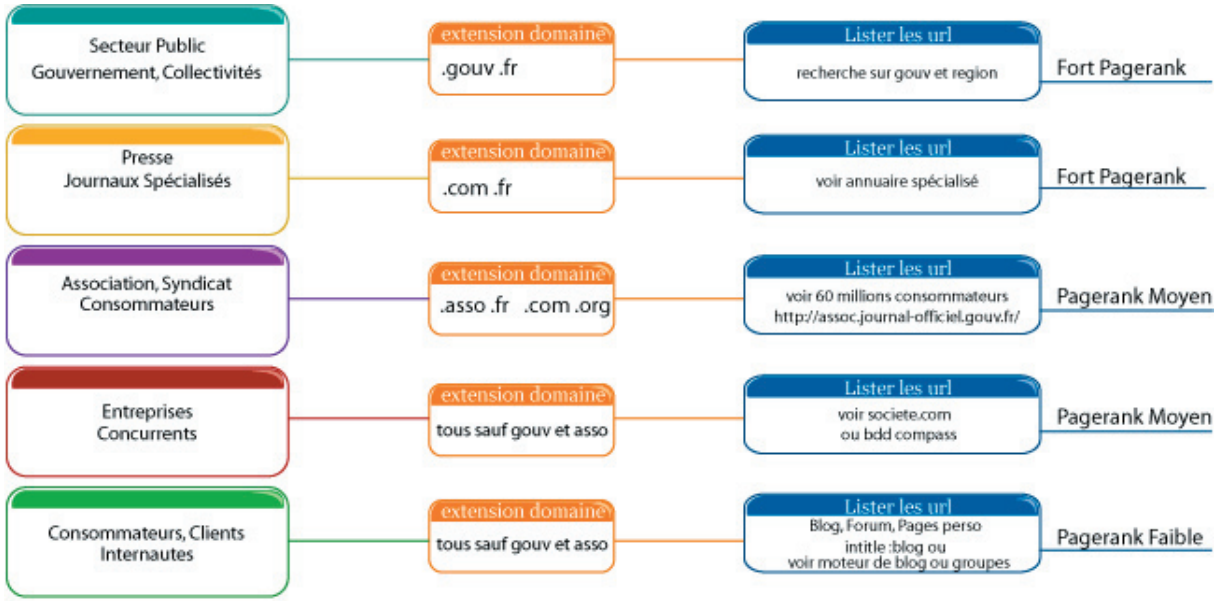
Pour une recherche fructueuse et un meilleur affinage, il semble intéressant de pouvoir mettre en corrélation les résultats obtenus sur les différents types d'outils mis à disposition de l'internaute (moteurs classiques, moteurs à clusterisation et moteurs cartographiques).

	Moteurs Standards				Moteurs à clusterisation						Moteurs cartographiques		
	Google	YAHOO!	msn	voila.fr	exalead	Vivisimo	SnakeT	iBoogie	WebClust	grokker	KartOO	MapStan
Minuscules - Majuscules (Indifferenciation de la casse)	COMPETITIVE = Competitive = competitive 251 000 000	COMPETITIVE = Competitive = competitive 197 000 000	COMPETITIVE = Competitive = competitive 37 442 000	COMPETITIVE = Competitive = competitive 1 046 705	COMPETITIVE = Competitive = competitive 29 762 389	COMPETITIVE = Competitive = competitive 46 420 000	COMPETITIVE = Competitive = competitive	COMPETITIVE = Competitive = competitive	COMPETITIVE = Competitive = competitive	COMPETITIVE = Competitive = competitive	COMPETITIVE = Competitive = competitive	COMPETITIVE = Competitive = competitive
Lettres accentuées (indifférenciation des accents)	Faible variance des resultats matière 72 600 000 matiere 72 700 000	Faible variance des resultats matière 23 500 000 matiere 23 700 000	Faible variance des resultats matière 6 384 771 matiere 6 384 771	Faible variance des resultats matière 2 131 729 matiere 2 134 729	aucune différence matière 7 814 881 matiere 7 814 881	aucune différence matière 2 041 899 matiere 2 041 899	aucune différence matière 181 matiere 181	Faible variance des resultats matière 4 980 000 matiere 4 970 000	aucune différence matière 200 matiere 200	aucune différence matière 141 matiere 141	aucune différence matière matiere	aucune différence matière matiere
ET (Permet de lier plusieurs mots)	+ intelligence+economique 3 430 000	+ intelligence+economique 1 340 000	+ intelligence+economique 334 000	+ intelligence+economique 70 320	+ intelligence+economique 360 395	+ intelligence+economique 145 021	+ intelligence+economique 180	+ intelligence+economique 294 811	+ intelligence+economique 200	+ intelligence+economique 123	+ intelligence+economique	+ intelligence+economique
OU (Permet le remplacement de mot (en majuscule))	OR competitive OR competition 399 000 000	OR competitive OR competition 438 000 000	OR competitive OR competition 78 005 666		OR ou OPT competitive OR competition 69 349 000	OR competitive OR competition 103 820 000		OR competitive OR competition 482 000 000	OR competitive OR competition 199	OR competitive OR competition 141	OR competitive OR competition	OR competitive OR competition
SAUF (Permet l'exclusion de mots ou expressions)	- flux -rss 13 300 000	- flux -rss 17 900 000	- flux -rss 5 196 000	- flux -rss 555 000	- flux -rss 5 076 431	- ou AND NOT ou NOT flux -rss 6 197 000	- flux -rss 180	- flux -rss 11 500 000	- flux -rss 198	- flux -rss 140	- flux -rss	- flux -rss
Expressions (recherche d'une expression complète)	“” "l'intelligence économique" 295 000	“” "l'intelligence économique" 278 000	“” "l'intelligence économique" 50 736	“” "l'intelligence économique" 16 357	“” "l'intelligence économique" 64 802	“” "l'intelligence économique" 51 917	“” "l'intelligence économique" 51 917	“” "l'intelligence économique" 50 941	“” "l'intelligence économique" 200	“” "l'intelligence économique" 123	“” "l'intelligence économique"	“” "l'intelligence économique"
Troncature (Permet l'affichage de la racine d'un terme)					* tel* (telephone, telecom..) 348 000 000							
Synonymes (Recherche élargie en rapport avec le mot)	~ ~finances (Budget, money...) 2 000 000 000											
Recherche sur le nom de domaine	site: site:www.economy.com 56 800	site: site:www.economy.com 86 800	site: site:economy 11 315	domain: domain:economy 1 800 000	site: site:economy 6 193	site: site:economy 37					domain: domain:economy	site: site:www.economy.com
Recherche sur les adresses des liens (Affiche les documents contenant l'url du site)	link: link:http://www.univ-tln.fr 313	linkdomain: linkdomain:www.univ-tln.fr 9 000	link: link:http://www.univ-tln.fr 2 691	link: link:http://www.univ-tln.fr 1 463		link: link:http://www.univ-tln.fr 33	link: link:http://www.univ-tln.fr 180	link: link:http://www.univ-tln.fr 240	link: link:http://www.univ-tln.fr 153	link: link:http://www.univ-tln.fr 124	link: linkdomain: link:http://www.univ-tln.fr	link: link:http://www.univ-tln.fr
Recherche de sites similaires (Affiche des sites aux fonctions semblables)	related: 31						like: like:http://www.univ-tln.fr 180	like: like:http://www.univ-tln.fr 951	like: like:http://www.univ-tln.fr 200	related: related:http://www.univ-tln.fr 122	like: related: like:http://www.univ-tln.fr	related:

	Moteurs Standards				Moteurs à clusterisation							Moteurs cartographiques	
	Google	YAHOO!	msn	voila.fr	exalead	Vivísimo	SnakeT	iBoogie	WebClust	grokker	MapStan	KartOO
Recherche dans le cache (Permet la lecture du site tel qu'apparu à la dernière visite du moteur)	cache: cache:http://www.univ-tln.fr Site tel qu'il était 5 jours avant											
Informations sur le site (repertorie diverses infos sur le site cherché)	info: info:http://www.univ-tln.fr affiche des info par google											
Recherche sur les mots (Affichage de définitions relatives aux mots tapés)	define: define:RSS affiche définitions pertinentes	define: define:RSS affiche définitions					define: define:RSS affiche url de définitions	define: define:RSS affiche url de définitions	define: define:RSS affiche url de définitions	define: define:RSS affiche url de définitions		
Recherche sur la date (Classement par date)	daterange: "flux rss" daterange:2452439-2452439				date: ou sort: flux rss sort:new flux rss sort:old							
Recherche sur le type de fichiers (affiche des documents en fonction de leurs extensions)	filetype: rss filetype:pdf 1 110 000				filetype: rss filetype:pdf 71 248							filetype: rss filetype:pdf
Recherche dans le titre (Recherche basée sur la présence d'un mot ou expression dans le titre)	allintitle: allintitle:"recherche et developpement" 10 400										allintitle: allintitle:"recherche et developpement"	allintitle: allintitle:"recherche et developpement"
Recherche dans le titre (Recherche basée sur la présence d'un mot ou expression dans le titre)	intitle: intitle:recherche et developpement 3 290 000	intitle: intitle:recherche et developpement 810,000		title: title:recherche et developpement 47 573	title: title:recherche et developpement 240	title: title:recherche et developpement 1 690	title: title:recherche et developpement 180	title: title:recherche et developpement 86 000	title: title:recherche et developpement 200	title: title:recherche et developpement	title: title:recherche et developpement	title: title:recherche et developpement
Recherche dans l'URL (Recherche basée sur la présence d'un mot ou expression dans l'url)	allinurl: allinurl:"veille-intelligence" 403										allinurl: allinurl:"veille-intelligence"	allinurl: allinurl:"veille-intelligence"
Recherche dans l'URL (Recherche basée sur la présence d'un mot ou expression dans l'url)	inurl: inurl:veille+intelligence 45 300	inurl: inurl:veille+intelligence 164		url: url:veille intelligence 2 273		url: url:veille intelligence 9	url: url:veille intelligence 180		inurl: inurl:veille intelligence 170		inurl: inurl:veille+intelligence	url: inurl: url:veille intelligence
Recherche dans le texte (Recherche basée sur la présence d'un mot ou expression dans le texte)	intext: intext:"université du sud" 68 600						intext: intext:"université du sud" 180	intext: intext:"université du sud" 4	intext: intext:"université du sud" 200		intext: ntext:"université du sud"	text: text:"université du sud"
Reherche sur le nom de domaine (se base sur le nom de domaine pour obtenir une réponse)		hostname: hostname:free.fr 39				host: host:free.fr 2 695 000	hostname: hostname:free.fr 8		hostname: hostname:free.fr 8	hostname: hostname:free.fr 8		host: host:free.fr

	Moteurs Standards				Moteurs à clusterisation								
	Google	YAHOO!	msn	voila.fr	exalead	Vivísimo	SnaketT	iBoogie	WebClust	grokker	MapStan	KartOO
Recherche sur un document précis (Basée sur l'adresse du document)		url: url:http://www.univ-tln.fr affiche la page + descriptif				url: url:http://www.univ-tln.fr recherche dans le texte						
Recherche sur le texte des liens (Affichage des sites comprenant le mot cherché dans un lien hypertexte)				anchor: anchor:université du var 8 087		linktext: linktext:université						
Recherche sur les balises Meta (Repertorie les pages comprenant le mot cherché dans les balises Keywords)				keyword: keyword:université var 21 884								
Recherche sur les balises Meta (Repertorie les pages comprenant le mot cherché dans les balises desscription)				desc: desc:ingemedia 188								
Recherche sur les balises Meta (Repertorie les pages comprenant le mot cherché dans les balises d'images)				alt: alt:image 314		image: image:image 314						image: image:image
Recherche basée sur la langue de la page					language: economy language:fr 289 700							

Mise en application par un exemple concret : Comment cerner et distinguer l'image d'une structure sociale sur le Web ?



Les secteurs d'influence sur l'image d'une entité sociale sur le Web

Un modèle type de requête a pu être formulé de manière à obtenir des résultats relativement cohérents et interprétables par de nombreux moteurs :

Nom Produit et/ou Nom entité sociale intext : 1 + ou -inurl: 2 daterange: 3

1. Insérer un ou plusieurs mots pour caractériser l'image à discerner ex: avis, critique, grève, défaillance, problème

2. Insérer une ou plusieurs url à exclure ou à inclure dans la recherche ex:univ, sncf...

3. Insérer la période de recherche de publication de xxxxxx à xxxxxx

La concaténation de ces 3 opérateurs ciblent la recherche effectué sur l'entité ou son produit à une date donnée des sites indexés (+ ou - pertinent et visible) qui influent sur l'image de l'entité

Références :

Webographie

Abondance – **Informations sur les moteurs de recherche** | http://outils.abondance.com/
DSI – **Les automates de recherche** | http://www.dsi-info.ca/moteurs-de-recherche.html
Savoirs CDI – **Recherches sur Internet, où en sommes nous, où allons nous ? |** http://savoirscdi.cndp.fr/CulturePro/actualisation/Serres/Serres.htm
Journal du Net – **Kartoo, un méta moteur de recherche très visuel |** http://solutions.journaldunet.com/0105/010514_kartoo.shtml
Biologeek – **Analyse des données utilisateurs** | http://www.biologeek.com/

Bibliographie

Jan Pedersen – **Internet Search Engines : Past & Future**
Jean Paul Pinte – **Les outils de la veille pédagogique** (2005, Revue internationale des technologies en pédagogie universitaire)
Guy Forzy – **Recherche sur le web : syntaxe de base** (DRT - CRDP de Lyon)