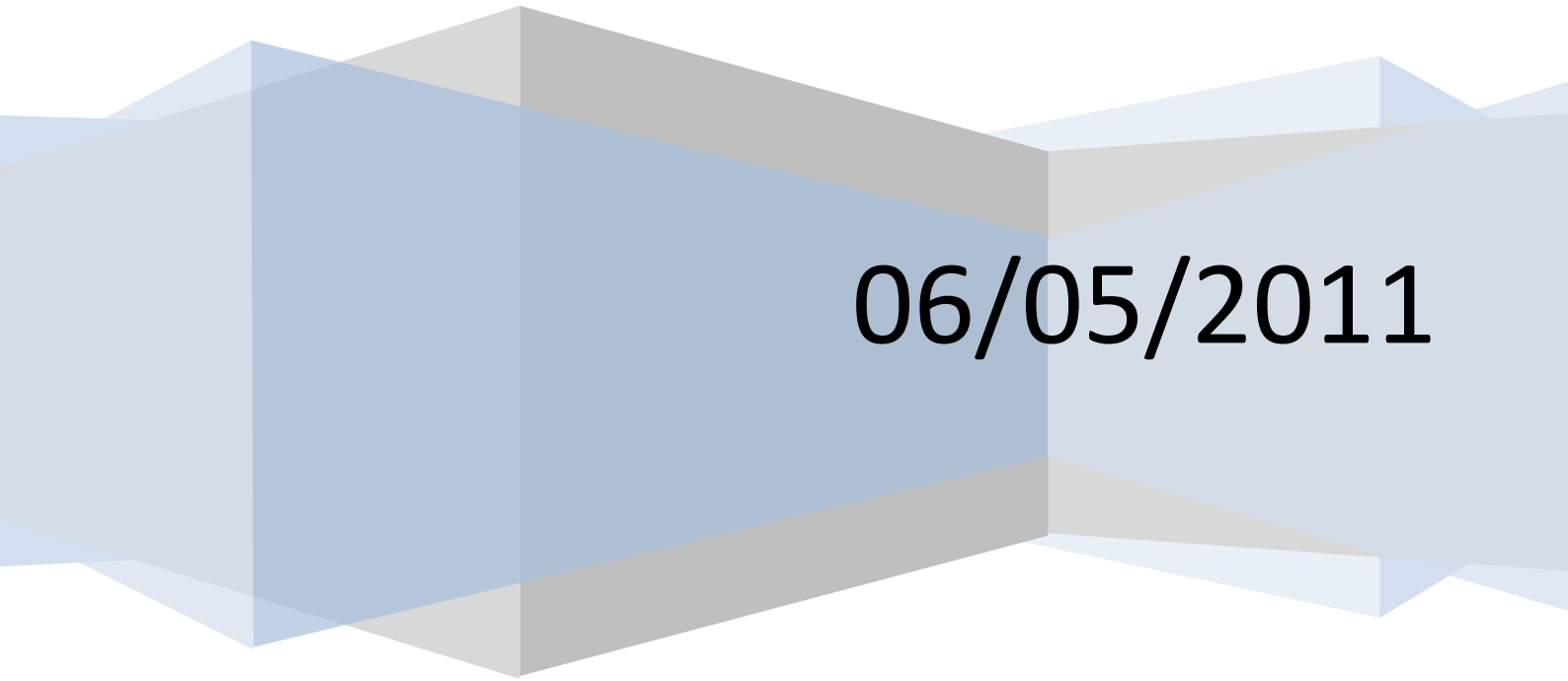
ISTIL-EPU

Livre Blanc

Présente une synthèse de nos recherches autour du thème de l'information Scientifique et Technique

Fabien COMOLET – Romain FONTAINE – Ludovic GARCIA – Florina POPESCU – Corentin SANIARD



06/05/2011

Sommaire

Introduction

Nous sommes un groupe de 5 étudiants de l'EPU Lyon 1. Nous avons effectué un projet de veille technologique dans le cadre de l'UE de Veille Technologique.

L'Unité d'Enseignement de Veille Technologique est un module de 2ème année de l'Ecole Polytechnique Universitaire de Lyon 1, filière informatique, qui s'étend sur six mois. Son objectif est double : il s'agit d'une part de former les étudiants au travail de groupe et d'autre part de leur faire découvrir et appliquer les principes de la veille technologique. Au cours de cette UE, les étudiants doivent étudier un domaine technique en détail et en réaliser le suivi sur plusieurs mois au cours desquels ils sont encadrés par deux tuteurs universitaires. Le sujet est proposé par un commanditaire, il concerne une question spécifique s'inscrivant dans un grand domaine thématique.

Notre sujet est la recherche d'API pour l'information scientifique et technique. Ce projet nous est présenté par Deuxième Labo qui recherche ainsi à favoriser l'accès aux informations Scientifiques ou Techniques. Néanmoins, il existe des problèmes pour effectuer ce type d'application. En effet, on peut citer le problème des homonymes (deux noms identiques pour deux personnes différentes) ou le problème de la signature d'un écrivain de sexe féminin : comment savoir si une femme a signé avec son nom de jeune fille ou avec celui de son mari (voire les deux) ?

Histoire du domaine

Les premiers moteurs de recherche sont apparus en 1990 par le biais de l'université McGill qui a créé le moteur Archie. Le second, apparu en 1991 fut Gopher, qui a perduré jusqu'en 2006. Ces premiers moteurs furent créés dans le but de trouver des informations parmi les pages web. Le premier succès commercial vient du moteur Lycos en 1993, qui est à l'origine d'un projet universitaire. Depuis 1996, les algorithmes de classification n'ont pas cessés de s'améliorer, le nombre de résultats renvoyés étant devenus de plus en plus importants (voire critiques : en 1999, Google référençait plus de 60 millions de pages).

Contexte du sujet

Les concepts du sujet

Une API (Application Programming Interface) a pour objet de faciliter le travail d'un programmeur en lui fournissant les outils de base nécessaires à tout travail à l'aide d'un langage donné. Elle constitue une interface servant de fondement à un travail de programmation plus poussé. Elle permet l'interaction des programmes les uns avec les autres, de manière analogue à une interface homme-machine, qui rend possible l'interaction entre un homme et une machine.

Dans notre cas, nous recherchons des API qui permettent d'accéder à des publications sur différents domaines. Ces API pourraient permettre le développement d'un outil qui interrogerait automatiquement différentes bases de données via ces différentes API. Notre principal outil de travail est internet. Les API se retrouvent parmi des moteurs de recherche, ou des bases de données d'API (par exemple programmableweb).

Nous ne parlons pas des moteurs de recherche classiques, tels que Google ou Bing, mais des moteurs de recherche que l'on retrouve dans des institutions pour la recherche de document. Une institution possède une base donnée de ses documents, généralement accessibles au public. Pour rechercher des documents, on remplit un formulaire et on clique sur « rechercher ».

Dans ce projet, nous cherchons les moteurs de recherche qui ont une API, c'est-à-dire ceux qui proposent un outil structuré permettant d'interroger ces bases de données. Le but de ces recherches, est d'évaluer la crédibilité d'un publieur en trouvant ses publications.

Les Acteurs du projet

Acteurs du projet

Dans le cadre du projet, il y a différents acteurs qui entrent en jeu.

- **Les commanditaires** : Ce sont eux qui proposent le sujet. Notre principal commanditaire est Antoine Blanchard, installé à Edimbourg. Il s'engage aussi à répondre aux interrogations des étudiants et à participer aux réunions de lancements, de pilotages et à la soutenance finale. Étant donné qu'il est basé au Royaume-Uni, M. Blanchard a été suppléé par Pierre Maumont lors de la réunion de lancement.

- **Les tuteurs :** Les tuteurs, au nombre de deux, représentent l'institution vis-à-vis du commanditaire. Le tuteur technique assure l'accompagnement du groupe sur les aspects techniques liés à l'accomplissement du projet (site web collaboratif) et sur les aspects technologiques du sujet de veille. Cette tâche est assurée par Nadia Kabachi. Le tuteur de communication assure l'accompagnement du groupe sur les aspects organisationnels (fonctionnement du groupe), sur la production de documents (rapports et sites Web), sur les présentations orales, ainsi que sur les aspects sciences humaines du sujet de veille. Michel Lalliard sera notre tuteur communication tout au long de notre projet.

- **Les étudiants :** Au nombre de 5, chaque étudiant travaille à l'aide d'un site web collaboratif (dans notre projet, nous avons un wiki à notre disposition où chaque utilisateur peut modifier, partager des informations). Les membres de l'équipe sont :
  - Florina Popescu : chef de projet
  - Romain Fontaine : responsable de la gestion des connaissances et la gestion de la documentation
  - Fabien Comolet
  - Ludovic Garcia
  - Corentin Saniard

Acteurs finaux

- **Les publieurs :** On les retrouve dans le monde entier. Ils produisent des publications, que l'on retrouve sur internet ou dans d'autres sources. Ce projet a pour but final de pouvoir juger la crédibilité d'un publieur.

- **Les chercheurs :** Ce sont les personnes que l'on retrouve au sein des universités. Ces personnes sont considérées comme étant des spécialistes travaillant sur la conception ou la création de connaissance, de produits ou de procédés. Les chercheurs sont des acteurs ayant besoin d'un accès vers les bases de données scientifiques, ou autres API, permettant de consulter des ouvrages ou des thèses.

En dehors des acteurs présents dans le cadre du projet, il faut aussi tenir compte des aspects qui peuvent surgir vis-à-vis du sujet.

Aspects sociaux-économiques

La diffusion d'informations grâce à internet s'est accrue au cours des années. Il est maintenant facile d'avoir accès à beaucoup d'informations en même temps. Internet joue un rôle social et économique. En effet, il permet aux personnes de s'informer de manière rapide, et donc peut influencer le mode de vie (ex : Facebook). Il régit encore plus les tendances que celles que l'on peut retrouver dans une cours de récréation.

De plus, il peut permettre à une entreprise de se faire connaître facilement et rapidement. Donc, les acteurs sociaux-économiques, tel que nous, surfons beaucoup sur le net et en tirons des avantages, et aussi contribuons à le faire fonctionner.

Aspects juridiques

Le moteur final a pour but d'indexer toutes les informations, de citer les sources ou leur provenance (site web, bibliothèque etc.). Dans le cadre de notre projet, nous devons respecter le droit à l'information, à son accès.

Les œuvres sont soumises à différentes politiques de gestion des droits d'auteurs : licences Creative Commons, aux droits d'auteurs et nous devons respecter cela. L'aspect juridique est un point déterminant et un facteur clé afin de créer un moteur capable de centraliser des informations dans le respect des droits d'autrui.

Ainsi, le moteur se contentera d'indexer les informations, c'est-à-dire de citer les différentes sources où l'on peut lire tel ou tel article, œuvre, publication. Le respect des droits est ainsi géré par les sites en aval.

Parmi toutes les recherches que nous avons effectuées, nous n'avons gardé que les API « pures ». Nous les avons mises en relations avec des critères essentiels pour la recherche d'information.

Les APIs

arXiv

Qu'est-ce que c'est ?

arXiv est une base de données en ligne contenant des publications ainsi que des liens entre ces différents articles, comme par exemple des citations. Cette Base de données est gérée par la Bibliothèque de l'Université de Cornwall.

A quoi sert-elle ?

Cette API a pour but de permettre un accès facile à cette base de données. Elle est accessible par le protocole http, cette API ne nécessite donc pas l'installation d'un logiciel afin d'interroger cette BD.

Restrictions d'accès

Cette API est à but non lucratif. On peut donc l'utiliser gratuitement. Il est possible de créer un compte, mais cela n'est nécessaire que pour ajouter des publications.

Comment l'utiliser :

On dispose en ligne d'un formulaire permettant d'effectuer des recherches, mais il est aussi possible de l'interroger directement via une url, par exemple, pour effectuer une recherche dans tous les domaines contenant le mot « electron » on utilisera :

http://export.arxiv.org/api/query?search_query=all:electron

En réponse on obtiendra un flux XML Atom.

Contenu de la base de données :

Le 8/12/2010, la base de données contient 644,661 articles en libre accès dans les domaines de la Physique, les Mathématiques, les Sciences Informatiques, la Biologie, la Finance et les Statistiques.

Les avantages :

Accès libre et rapide au contenu de la base de données.

Les inconvénients :

Un nombre restreint de publications comparé à certaines API comme Springer qui rassemble plusieurs millions d'articles.

Springer

Qu'est-ce que c'est ?

Springer est une API regroupant des métadonnées pour plus de 4.8 millions de documents. Springer est ainsi un éditeur de presse scientifique.

A quoi sert-elle ?

Cette API fournit des accès sur les documents, les métadonnées et des images pour 80 000 articles dans le biomédical et les programmes libres de publication.

Que récupère-t-on ?

Cette API permet de fournir plusieurs variétés de documents aux formats XML ou JSON.

Comment l'utiliser ?

Exemple de requête : http://dev.springer.com/docs/read/Filters_Facets_and_Constraints

Il suffit de lancer une requête q, avec différents paramètres et filtres :

Exemples :

q=type:Book ➔ retourne les livres uniquement

q=name: «hughes, m"➔ retrouve les articles qui incluent le nom de l'auteur : Hughes, M".

Exemple avec filtres multiples :

q=Issn:1573-4838 year:2009. La réponse fournie (dans le lien précédent) nous donne un exemple de réponse au format XML.

Ce lien : http://dev.springer.com/docs/read/Example_API_Responses montre d'autres formats de réponse en JSON ou en PAM (genre de format XML).

Comment déterminer le type de document et son format de réponse?

Exemple :

http://api.springer.com/metadata/pam/doi/10.1007/s11276-008-0131-4?api_key=yourKeyHere

Ceci est une requête pour obtenir des métadonnées (« métadata ») avec un format de sortie « PAM ». Ainsi, la requête analyse les métadonnées recherchées, et les métadonnées retournées sont retournés en format PAM.

Avantage :

Fournit un grand nombre de documents sous plusieurs formats différents. Les filtres peuvent être multiples et les recherches assez précises.

Inconvénient :

La requête à envoyer est assez complexe et pas très intuitive.

CatalogWS

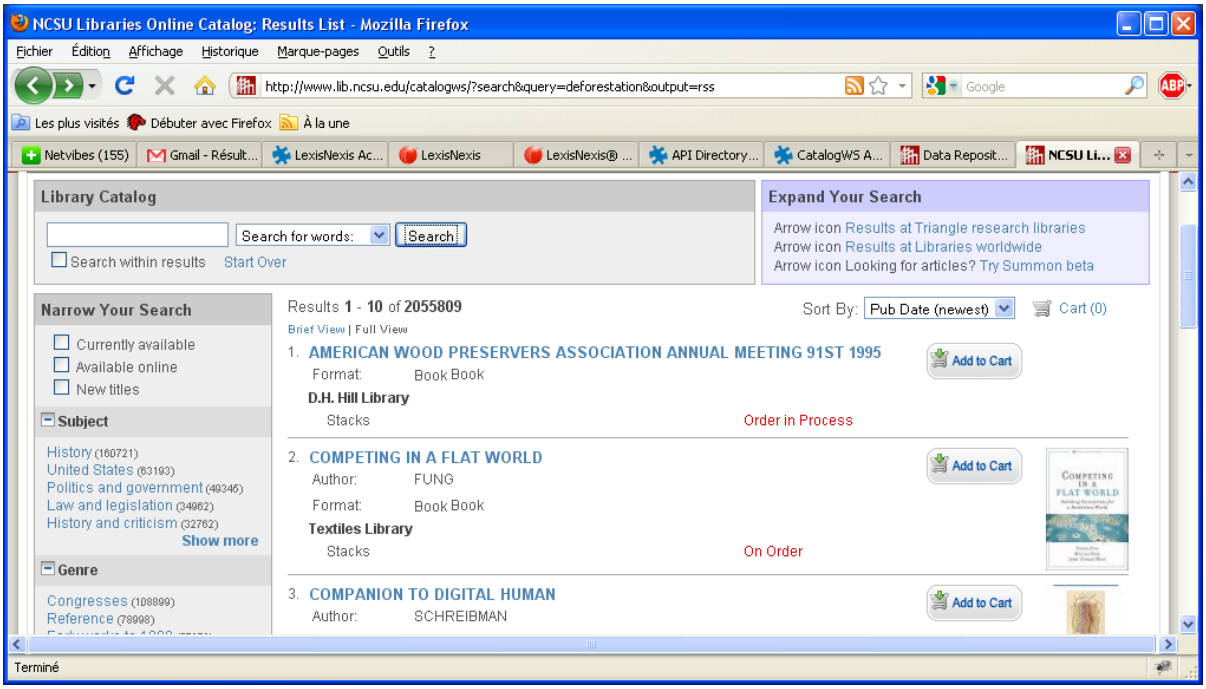
Qu'est-ce que c'est ?

CatalogWS est une API en ligne permettant de consulter la base de données du NCSU (North Carolina State University). Cette API possède une interface Web permettant d'effectuer des recherches.

Malheureusement, pour l'instant l'interface renvoie toujours une erreur quel que soit la recherche effectuée.

On peut tout de même interroger l'API via une requête http en passant des paramètres en GET (directement dans l'adresse).

Interface



Utilisation en ligne

Exemple de recherche sur le thème de la déforestation :

http://www.lib.ncsu.edu/catalogws/?service=search&query=deforestation&output=rss

Les paramètres

- Service = search spécifie que l'on souhaite effectuer une recherche dans la base de données.
- Query = définit les mots clés à utiliser
- output= permet de définir le format des résultats (par défaut XML). On peut obtenir les résultats sous forme d'un flux XML, RSS, XSL et PowerSearch.

Avantage

Utilisation gratuite.

Inconvénient

Très peu de documentation, uniquement une adresse mail de contact.

Source

http://www.lib.ncsu.edu/dli/projects/catalogws/

DeepDyve

Qu'est-ce que c'est ?

DeepDyve est le plus grand service de location pour la recherche scientifique, technique et médicale avec plus de 30 millions d'articles via des milliers de journaux.

Les recherches s'effectuent grâce à des mots clé.

Résultat :

Les articles qui correspondent au mot recherché sont générés. La recherche peut être affinée en fonction de l'auteur, date, éditeur, source ou contenu.

L'API offre des résultats propres au site DeepDyve mais aussi des résultats du Web comme Wikipedia.

Avantage :

On peut tester gratuitement l'API afin d'évaluer la qualité et l'importance du contenu avant d'acquérir quoi que ce soit.

Inconvénients :

L'accès est payant, de plus, il est variable. En effet le prix est fixé par le détenteur des droits sur l'article. En résumé il faut payer, consulter ou imprimer un article.

Mendeley

Qu'est-ce que c'est ?

Mendeley est un logiciel de gestion bibliographique, destiné à la gestion et au partage de travaux de recherche. Il est composé d'un logiciel gratuit de bureautique (Windows/Mac/Linux) gérant notamment les PDF, les citations, et les références bibliographiques, et d'un réseau web. Il peut aussi être synchronisé avec un compte web utilisateur afin de présenter son identité numérique.

Une API est disponible afin de profiter de toutes les ressources bibliographiques fournies par cet outil.

Utilisation en ligne

Exemple de lien de recherche. Il faut remplacer la valeur <consumer_key> par la clé API obtenue après avoir créé un compte.

http://www.mendeley.com/oapi/stats/authors?consumer_key=<consumer_key>

Avantages

La base de données est imposante et il est possible d'effectuer des requêtes avec des filtres multiples**.**

Inconvénient

Il faut s'inscrire sur le site web avant de pouvoir essayer l'API.

Source

http://dev.mendeley.com/

Scopus

Qu'est-ce que c'est ?

Scopus est une API qui permet d'utiliser l'outil Sciverse. Sciverse est un outil simplifiant l'accès à l'information scientifique. Cet outil est destiné à différents acteurs, notamment les Chercheurs, les libraires et les développeurs.

L'API Scopus permet d'utiliser facilement Sciverse.

Cette API est destinée :

- aux Chercheurs, pour leur permettre de consulter des publications, consulter les articles citant leurs propres publications.
- aux Bibliothécaire, afin de mettre à disposition des utilisateurs des informations fiables et récentes. Cela permet aussi suivre les avancés dans certains domaines.
- aux Gouvernements afin de leur permettre d'évaluer les performances de leurs laboratoires de recherche, mais aussi d'établir le profil de candidat.
- aux Editeurs, afin d'améliorer la visibilité de leurs ouvrages.
- aux Industriels afin de leur permettre un accès fiable aux dernières informations sur la recherche.

Comment l'interroger et que récupère-t-on ?

On peut utiliser cette API par l'intermédiaire d'une interface web développée par Sciverse, ou on peut utiliser directement l'API.

Dans le cas de l'interface web les résultats sont sous format texte, sinon on peut récupérer ces informations sous forme XML afin de les réutiliser.

Comment obtenir accès à cette API ?

L'accès à ces informations est payant. Le coût dépend du type et de la taille de l'institution concernée. Il est aussi possible d'essayer cette API pendant une durée limitée, il faut pour cela remplir un formulaire qui permettra à un représentant de la société de prendre contact avec le client.

Inconvénient :

N'étant qu'étudiants, nous n'avons pu essayer cette API, de plus les frais d'inscription semblent être élevés.

Tableau comparatif

Afin de bien différencier ces API nous les avons présentés sous forme d'un tableau comparatif. Pour pouvoir établir ce tableau nous avons réfléchi à un certain nombre de critères présentés ci-dessous :

- **Le type d'information**
- **Le domaine des informations** et **spécialités** : cela nous permet d'identifier les API génériques et les API spécialisés. Ceci dépend uniquement de la base de données interrogée par l'API en question.
- **La méthode d'interrogation** : cela permet d'établir un premier constat technique quant aux technologies à utiliser pour utiliser les API. La plupart des API s'interrogent via une requête http sur une url. L'utilisation du protocole http garantit un niveau technique faible, il n'est pas nécessaire de se former longuement avant de pouvoir utiliser l'API.
- **Les formats de sorties** : Afin de permettre à plusieurs programmes d'interagir il est nécessaire d'établir des normes sur les modes de communication et les structures de données. Le format que l'on retrouve majoritairement est le XML, c'est un format largement répandu et une technologie mature. Il est donc facile de récupérer et de traiter ces informations.
- **L'accès :** Les conditions d'accès à l'API est une variable majeure dans le choix d'une API. En effet il existe de nombreuses différences entre les différentes API. Certaines sont totalement gratuites, d'autre partiellement ou complètement payantes.
- **Avantages et inconvénients :** Nous avons cherché à identifier les points forts et les points faibles pour chaque API. Dans certains cas cela n'a pas été possible, par exemple dans le cas de l'API Scopus. En effet il était nécessaire de prendre contact avec un responsable commercial afin de pouvoir se faire sa propre idée sur leur produit.
- **URL d'interrogation :** Nous avons proposé une URL d'interrogation déjà paramétrée afin de pouvoir visualiser rapidement un exemple des informations obtenues. Cela ne fonctionne que pour les API gratuites ne nécessitant pas une clé API.
- **URL de documentation :** C'est un lien vers les pages de documentation technique de l'API.

	arXiv	catalogWS	Deepdyve	Springer	Mendeley	Scopus
Type d'informations	Articles	Article + thèse + documents	articles + textes intégraux	Documents, métadonnées et images	Travaux de recherche	Travaux de recherche et articles
Domaine des informations	Physique, Mathématiques, Sciences Informatiques, Biologie, la Finance, les Statistiques	NCSU Libraries catalog	littérature scientifique, technique et médicale	Biomédical et les programmes libres de publication	x	x
Spécialités	Tout domaine scientifique	x	x	Biomédical	X	x
Méthode d'interrogation	Requête URL : utilisation du mot clé "électron"	requête URL	x	Nécessite une clé API, requête URL	requête http	x
Format de sortie	Flux XML, Atom	XML, RSS	x	XML, JSON	JSON	x
Accès	Gratuit	Gratuit	payant à l'inscription, consultation	Gratuit	Licence propriétaire, nécessite un compte Web.	Accès payant
Avantages	Gratuit, simple d'utilisation	2 services différents	x	Grand nombre d'informations + filtres multiples	Base de données imposante + filtres multiples	Accès à une vaste base de données d'éditeurs
Inconvénients	Peu de publications	x	Accès payant	Requête peu intuitive	nécessite un compte Web	Accès payant. Contact avec un représentant commercial pour tout essai et/ou informations. Le coût semble important.
URL d'interrogation	Url d'interrogation[1]	Url d'interrogation[2]	x	x	Url d'interrogation[3]	x
Documentation		Url de Documentation[4]	x	x	Url de documentation[5]	x

[1] http://export.arxiv.org/api/query?search_query=all:electron%20
[2] http://www.lib.ncsu.edu/catalogws/?service=search&query=deforestation&output=rss
[3] http://www.mendeley.com/oapi/stats/authors?consumer_key=%3cconsumer_key%3e
[4] http://www.lib.ncsu.edu/catalog/ws/
[5] http://dev.mendeley.com/docs/

Autres manières de rechercher de l'information

Notre sujet de travail, nous a permis d'explorer plusieurs pistes de recherche et de réflexion. Avant d'arriver à une synthèse des API vu ci-dessus, nous avons trouvé des « purs » moteurs de recherche. Il est intéressant d'en tenir compte, pour suivre leur évolution qui pourrait les amener à rentrer dans la catégorie API.

Voici la liste de ces moteurs que nous avons étudiés :

CSTB : Abréviation de Centre Scientifique et Technique du Bâtiment, il constitue une base de données qui regroupe une sélection de publications et divers écrits de nature scientifique ou technique rédigés par des ingénieurs et chercheurs du CSTB.

GoogleScholar : C'est le moteur de recherche de documents de Google. Il permet d'effectuer facilement une recherche étendue portant sur des travaux universitaires. Ces travaux peuvent provenir de sources telles que des éditeurs scientifiques, des sociétés savantes, des référentiels de prépublication, des universités et d'autres organisations de recherche.

Hal : C'est une archive ouverte pluridisciplinaire qui est destinée au dépôt et à la diffusion d'articles scientifiques de niveau recherche, publiés ou non, et de thèses, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

OJOSE (Online Journal Search Engine): C'est un moteur de recherche de publication scientifique gratuit. Une requête peut être soumise sur plusieurs bases de données (60 bases différente) juste en un clic. On peut télécharger ou acheter les publications scientifiques trouvé.

OpenDOAR : Il est différent des autres, car il ne propose pas des articles en tant que résultat, mais des URL de base de données. OpenDoar répertorie des bases de données académiques. Ces bases de données sont « Open Access ». Et donc une fois en possession de ces URL on peut parcourir ces différentes bases de données.

OpenSIGLE : C'est un site qui signale la littérature grise européenne. Divisé en deux communautés : aux pays ayant participé à SIGLE d'une part, et aux catégories principales du plan de classement SIGLE, de l'autre. SIGLE est une base de données - System of Information on Grey Literature in Europe – de la littérature grise en Europe. Dans chacune de ces communautés, un nombre illimité de collections peut être défini, qui peuvent comporter un nombre illimité de documents.

Persée: C'est un portail de revues en sciences humaines et sociales. Il nous permet de trouver des publications dans ces domaines.

Reader Meter : Ce site est un peu différent des autres, car il permet de mesurer l'impact de publieurs dans leur domaine. De plus, il fournit un classement des articles/livres en fonction de la popularité, les coauteurs des publications, ainsi que le profil des lecteurs et leur localisation.

Refdoc : Ce site nous fournit plus de 50 millions de références d'articles, ouvrages, rapports, actes de congrès, en science, technologie, médecine, sciences humaines et sociales, depuis 1847 à nos jours.

Scirus : C'est un outil de recherche scientifique sur le web. Avec plus de 410 millions d'articles scientifiques indexés au dernier décompte, il permet aux chercheurs de trouver des informations dans les sites web et le contenu des revues, mais également les pages d'accueil des scientifiques, les brevets et les dépôts institutionnels.

Dans le domaine de la recherche d'information, il existe aussi les méta-moteurs et les systèmes multi-agents.

Les méta-moteurs sont des outils de recherche sur internet qui permettent d'adresser simultanément une même requête à différents moteurs de recherche. Ils sélectionnent les réponses en fonction de leur pertinence. Les critères de pertinence diffèrent selon les méta-moteurs. Certains ne retournent pas la totalité des réponses fournies par un moteur de recherche en particulier.

Le vrai inconvénient de ces outils est qu'ils ne reprennent pas les fonctions avancées des moteurs de recherche (recherche dans les champs, dans les domaines etc.). De plus les requêtes valables pour un moteur ne le sont pas toujours pour un méta moteur. On identifie 3 générations différentes de méta-moteur, selon leur niveau de perfectionnement.

Méta-moteurs de 1ère Génération

Dogpile.com

Ce méta moteur permet d'interroger 5 moteurs de recherches simultanément, puis il effectue un tri afin de ne restituer que les meilleurs résultats.

Ces 4 moteurs sont :

- Google
- Yahoo
- Bing
- Ask

Search.com

Identique à dogpile, il permet d'interroger les moteurs suivants :

- Google
- Ask
- Msn
- DMOZ

Il n'y a pas d'API mais le site nous redirige vers un outil logiciel permettant d'effectuer des recherches : http://www.webferret.com/

Metacrawler

Il est basé sur le même principe que les précédents et interroge Yahoo, Google et Bing.

.

Méta-moteurs de 2nde génération

Ce sont des logiciels plus intelligents que les méta-moteurs de 1ère génération. Ils permettent d'éliminer les doublons, d'enregistrer les résultats pour une consultation hors ligne, etc.

On retrouve notamment Copernic, WebHaker/ WebSeeker et ECatch.

Méta-moteurs de 3ème génération

Les méta-moteurs de 3ème génération sont les plus sophistiqués, ils sélectionnent les sites dans différents moteurs, éliminent les doublons et affichent les résultats selon des critères de pertinence ou par type de document.

Les méta-moteurs permettent d'interroger simultanément plusieurs moteurs de recherche avec une même requête (du moins pour les méta-moteurs de génération 2 et 3). Les résultats de la requête sont issus de plusieurs bases de données, ce qui permet une plus grande couverture de l'Internet. Le principal avantage des méta-moteurs est donc l'exhaustivité.

Mais la même requête est envoyée à tous les moteurs, il est dès lors impossible de faire une requête complexe puisque chaque moteur utilise une syntaxe propre. L'usage des méta-moteurs se limite donc souvent à des recherches simples.

DigOut4U

DigOut4U est un système automatique de recherche sémantique d'information sur le Web conçu par la société Arisem. Il permet de poser une requête multilingue (anglais, français) en langage naturel sur plusieurs moteurs de recherche en même temps. Les résultats sont analysés, téléchargés sur votre ordinateur et classés. Les pages doublons comme celles comportant une erreur (404 par exemple) ne sont pas prises en compte. Une requête peut être redéfinie afin de trouver des résultats plus pertinents (recherche en entonnoir).L'analyse sémantique doit permettre de réduire les problèmes de bruit, ou de silence, ainsi que le volume d'information. Un de ces inconvénients, est qu'il ne fonctionne pas sous Windows Seven.

Un agent est une entité réelle ou virtuelle, évoluant dans un environnement (complexe), capable de le percevoir et sur lequel elle peut agir (en conséquence).

Les agents peuvent être un outil complémentaire à l'utilisation d'API. Grâce à leurs mécanismes d'intelligence, ils peuvent améliorer les résultats, affiner les recherches et mieux exploiter ces résultats.

Les agents intelligents

BullsEye:

BullsEyesPro d'Intelliseek (éditeur du méta-moteur Profusion) permet de crawler et de surveiller le web.

Le module Search : permet de rechercher sur le web mondial ou spécialisé par pays. Le module permet également un choix sur des thématiques de recherche : Multimédia, groupes de discussion, informatique, logiciels, santé, livres, commerce, loisirs, éducation, ressource gouvernementales. On peut effectuer une simple recherche, ou bien une "analyse" (contrôle des liens morts, élimination des doublons, téléchargement des résultats avec mise en valeur des mots-clés).

Le module Tracker : permet de surveiller les pages d'un site ou de paramétrer une recherche récurrente. Dans les 2 cas, les options de paramétrage sont nombreuses : fréquence d'alerte, type de modifications à surveiller (textes, liens, images), type de reportinou autre.

Pratique : il est possible de configurer des filtres pour éviter d'être averti pour des modifications sans intérêt (publicités, compteurs, etc.).

Strategic Finder:

C'est un peu le concurrent de Copernic. Edité par la société Digimind, cet agent permet d'interroger plus de 4000 sources d'informations. Son originalité : la possibilité de rajouter des plugins (payants) qui permettent une recherche ciblée sur un secteur d'activité précis (plus de 38 disponibles).

Exemple : le module AerospaceFinder vous permettra de rechercher sur plus de 200 sources spécialisées dans l'industrie aéronautique.

L'objectif de Strategic Finder est de rendre possible la veille stratégique sur Internet. Il surveille l'apparition de produits de substitution, l'émergence de nouveaux concurrents, de nouvelles technologies, etc.

Il interroge plusieurs centaines de moteurs de recherches professionnels sur Internet, en même temps et à partir d'une seule requête. Plus qu'un simple méta-moteur, Strategic Finder s'affirme comme un outil de gestion des sources d'informations spécialisées sur Internet. Il utilise les meilleures bases de données spécialisées dans votre secteur d'activité. Il récupère les informations sur votre disque dur et les filtre automatiquement.

Idées de création d'entreprises

Nous avons eu 2 idées de création d'entreprise autour de la recherche d'information.

Détective privé

Quoi?

Un logiciel utilisant les APIs pour la recherche scientifique et académique peut aider comme fonction support d'une entreprise pour la recherche de personnes, type détective privé ou profiler.

Comment?

Cette entreprise pourrait modifier les paramètres des APIs pour rechercher des informations qu'elle juge importantes. Nous parlons ici surtout de l'origine des bases de données et des types de recherches.

Démarches

1. Pour créer une telle entreprise les créateurs doivent obtenir l'accès à des bases de données dites confidentielles, comme celle d'Interpol ou autre agence de recherche de personnes, qui ont une API.
2. Avec tous ces accès, créer un programme qui va synthétiser les résultats des recherches et leur présenter un profil de la personne recherchée.

Améliorations éventuelles

Les améliorations seraient en matière de configuration de logiciel, améliorations qu'une société de service en informatique fait pour un ERP qu'elle vend à plusieurs entreprises, mais qui configurée différemment sert à plusieurs utilisateurs avec des besoins différents.

Ici on peut rajouter au nom dans la recherche, la date de naissance pour trouver la bonne personne, inclure dans les résultats des photos ou des vidéos et aussi éventuellement des coordonnées GPS pour retrouver la personne et retracer son parcours depuis une date demandée.

Avantages

Avoir développé une application qui synthétise les informations sur une personne, rendre la configuration plus facile pour l'utiliser dans le domaine de la recherche des personnes. Ce qui reste à faire est référencer les bonnes sources d'information et espérer qu'elles ont une API.

Eventuels problèmes

Un tel projet risque de rencontrer quelques problèmes de sécurité et cryptage, car des mesures de sécurité sont appliquées sur les bases de données des autorités.

Il ne faut pas oublier aussi le droit à l'image qui doit être respecté, ou au moins la loi pour les détectives privés en ce qui concerne les recherches.

Pour que les résultats portent sur la vraie personne, il faut aussi prendre en compte l'existence des plusieurs personnes avec le même nom et donc arriver à filtrer nos résultats.

Faisabilité

Le projet est très faisable, l'important est d'avoir accès aux bonnes bases et l'accord légal. Des moteurs de recherche de personne existent déjà sur internet. Par exemple :

- PeekYou
- Ex.plode.us
- InfoSpace
- Spock
- Spokeo
- Wink
- Zabasearch.com
- ZoomInfo
- X-Recherche (en français)
- Freefind
- Free-Site-Search (maximum 500 pages)

Pourtant ces exemples ne filtrent pas les résultats pour donner la bonne personne et parfois ne la trouvent pas du tout, on peut donc penser qu'ils n'ont pas les bons algorithmes de recherche ou les bonnes sources.

Aide aux maisons d'édition

Quoi?

Le travail d'édition littéraire consiste à choisir des textes, les imprimer et les commercialiser. Mais pour cela il faut étudier l'œuvre proposée pour ses qualités ou bien la conformité à la ligne éditoriale de ses collections et aussi la crédibilité de (des) l'auteur(s). Cela pourrait se faire à l'aide d'un moteur de recherche académique et scientifique adapté ayant deux modules.

Comment?

Cette étude de l'œuvre et l'auteur peut se faire de manière automatique d'une part pour l'éventuel plagiat de l'œuvre et aussi pour la crédibilité de l'auteur. Le plagiat étant étudié dans un module de recherche dans les publications déjà existantes sur internet et l'auteur dans un module de recherche des personnes mais surtout des bases académiques et scientifiques, type ReadMeter, pour avoir une idée de son succès et des réseaux sociaux pour juger sa crédibilité.

Démarches

1. Pour créer une telle entreprise il faut, dans un premier temps, développer une application avec un cahier de charges assez pointu qui mentionne toutes les bonnes bases de données à interroger pour l'œuvre, ainsi que pour l'auteur.
2. L'application une fois mise en place, il faut la mettre à jour périodiquement, au cas où il y aurait des nouvelles bases de données qui doivent être rajoutées ou bien d'autres modules souhaités.
3. La personne qui l'utilise doit être polyvalente, pour choisir objectivement entre les résultats donnés par le logiciel et la qualité de l'œuvre.

Améliorations éventuelles

Cette fois le logiciel doit avoir deux modules, mais les deux basés sur des moteurs de recherche, pour l'œuvre, et les bases de données de l' API pour l'auteur, en rajoutant des recherches dans les réseaux sociaux pour avoir une image d'ensemble sur la personne et l'impact de son image pour la maison d'édition.

Avantages

Une application qui donne des informations sur une personne est un des modules de cette application, donc déjà un bon avancement par rapport au résultat voulu. Le reste peut être adapté.

Eventuels problèmes

Les éventuels problèmes seraient encore ceux par rapport à l'identité de la personne et de ce qu'elle veut qu'un éditeur sache sur elle. Il serait aussi plus difficile si les auteurs publient sous un autre nom, ce qui rend leur recherche plus difficile.

Faisabilité

De nos jours les éditeurs choisissent leur publications en fonction de leur qualité et authenticité mais nous ne savons pas comment ils jugent cela, nous avons donc pensé à faire un logiciel qui peut centraliser ces informations qui circulent sur internet pour les aider.

Vision prospective du domaine

Moyen terme : 6 mois à 1 an

Le domaine de la recherche d'information scientifique et technique est un domaine qui peut évoluer assez rapidement. Il est donc difficile de prévoir sur 6 mois.

Long terme : Quelques années

A long terme, le développement de la recherche d'information va se situer au niveau des différents types des moyens de recherche.

Les différents types d'agents : logiciels, intelligence artificielle, programmation, agents utilisés en e-commerce, simulation d'environnement, robotique, mais aussi les moteurs de recherche intelligents - Un projet d'avenir.

Conclusion

La recherche d'information scientifique et technique est fondamentale pour toutes les personnes désirant trouver de la documentation dans un domaine spécifique. La recherche d'information est en grande majorité utilisée par des chercheurs, des étudiants, des publieurs. C'est donc un public assez ciblé.

Nous pouvons chercher de l'information sur les sites des institutions via leur formulaire de recherche, ou leur API si elles en possèdent une. Aussi on peut utiliser des méta-moteurs, ou encore des systèmes multi-agents.

Lors de la recherche d'information sur les publications d'une personne, il existe 2 problèmes qui sont : l'homonymie et le changement de nom de famille au cours de la vie. Une seule solution est totalement efficace. En effet, la seule alternative assurant une identification fiable à 100% est l'utilisation de researcherID.

Il reste à voir dans les années à venir si cet outil se démocratise ou si d'autres possibilités émergent.