XEW-SS USPTO : la veille brevets de la BdD américaine

Amine EL HADDADI (*)(**), **Aziz LAKTIB** (***), **Anass EL HADDADI** (***)
{amine.elhaddadi, laktaib.mail, anass.elhaddadi}@gmail.com

(*) Faculté des Sciences et Techniques, Ancienne Route de l'Aéroport, Km 10, Ziaten. BP : 416. Tanger (Maroc),
(**) IRIT, Université Paul Sabatier, 118 route de Narbonne, F-31062 Toulouse (France),
(***) Ecole Nationale des Sciences Appliquées, BP 03, Ajdir Al-Hoceima (Maroc).

Mots clefs :
Agent, Collecte d'Informations, Crawler, Scraper, Système d'Intelligence Economique, Veille scientifique et technologique.
Keywords:
Agent, Information Gathering, Crawler, Scraper, Competitive Intelligence System, Scientific and technical observation.
Palabras clave:
Agente, Reunir de Información, Crawler, Raspador, Sistema de Inteligencia Competitiva, Escudriñar científico y tecnológico.

Résumé

Grâce à un chiffre de plus de 45 milliards de pages que représente Internet actuellement, l'intérêt des entreprises aux données contenues par le web s'est accru d'une manière assez remarquable. En effet, de plus en plus d'entreprises cherchent à collecter, analyser et exploiter cette ressource quasi-inépuisable. Pour répondre à ce besoin, plusieurs outils ont été proposés. Parmi eux, on distingue deux grandes catégories, les crawlers et les scrapers.
Dans cet article, nous allons mettre en évidence l'importance de ces outils, puis nous allons présenter notre nouvel outil « XEWAgent », qui est un agent intelligent de crawling et de scraping dédié à des fins d'Intelligence Economique.

1 Introduction

Internet, qui peut être défini comme étant un large éventail de données et de liens hypertextes interconnectés, représente une source quasi-inépuisable de données utilisables pour des fins de Data Mining. Ceci-dit, il est à savoir que pour pouvoir collecter des données du web, il faut faire face à certaines problématiques :

- La complexité des pages web due essentiellement à leur quantité et à leur structure diversifiée.
- Le fait que l'information utile ne soit contenue que dans une petite partie de la page web.

Plusieurs outils ont été proposés pour la collecte des données à partir du web et qui peuvent être catégorisés selon deux types majeurs : « Crawlers » et « Scrapers », les deux prochaines sections auront pour rôle expliquer ces deux concepts et les sections qui suivront présenteront notre nouvel outil « XEWAgent » qui est un agent intelligent de crawling et scraping dédié à des fins d'Intelligence Economique.

2 Crawlers

Le crawling, un concept aussi vieux qu'internet lui-même avait été utilisé originellement afin de parcourir et d'indexer les pages web et d'en établir une cartographie, un web crawler peut avoir plusieurs autres appellations telles que : web bot, web spider ou autres…

2.1 Principe

Le crawler commence par un ensemble défini d'URLs de pages web. Il les scanne et détecte la présence d'éventuels liens hypertextes pointant vers des pages non-prédéfinies qu'il scanne encore une fois dans un cercle vicieux. Les URLs nouvellement découvertes représentent un nombre de tâches en attente qui peut facilement monter exponentiellement, ce qui pourrait éventuellement causer un crash du crawler ou même du système. Pour éviter cela, il est de bon augure de sauvegarder ces tâches sur le disque et de libérer la mémoire [1]. Une autre solution à considérer est d'utiliser un crawler ciblé [2] qui a pour but de chercher l'information reliée à un domaine spécifique (l'Intelligence Economique dans notre cas).

Le processus de crawling commence par la lecture d'une liste d'URLs dits « candidats », qui peut être fournie soit par un utilisateur, soit par un autre programme. Une boucle reçoit dans chaque nouvelle itération un élément de cette liste, recherche la page web correspondante afin d'en extraire d'autres URLs et les ajouter à la liste des candidats. Il est possible d'ajouter un score à une page web visitée exprimant son degré de pertinence avant d'ajouter son URL à la liste. Le processus de crawling s'arrête lorsque la liste de candidats devient vide, le crawler n'a donc plus de pages web à scanner et se voit arrêté.

La figure 1 décrit le processus d'un crawler [1].

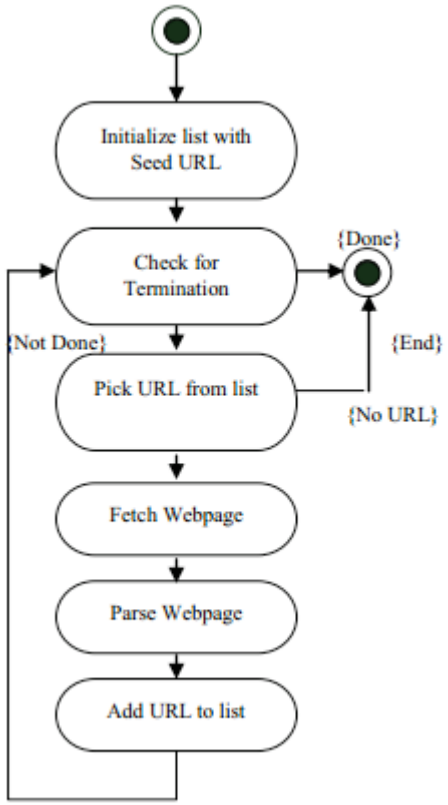


Figure 1 : Processus de Fonctionnement d'un Crawler [1]

2.2 Techniques de Crawling

Ouyang et al. [3] avaient distingué deux types de techniques ou de stratégies de crawling :

- Les stratégies heuristiques basées sur le contenu des pages web.
- Les stratégies basées sur l'évaluation des URLs.

2.2.1 Stratégies basées sur le contenu des pages web

Ces stratégies sont en principe basées sur les algorithmes Fish Search [4] et Shark Search [5] qui évaluent une URL en calculant le degré de similarité, d'un côté, entre le contenu de la page web et le thème de recherche, et d'un autre côté, entre les textes d'ancre et le thème de recherche.

Ce type de stratégies s'avère avantageux dans le sens où il possède, non seulement, une base théorique bien formulée, mais aussi une complexité réduite et une haute précision. Cependant, il représente certaines difficultés quant à la récupération efficace de la cartographie des liens hypertextes puisqu'il néglige cette information.
A l'époque où ces algorithmes ont été proposés, le nombre assez réduit de sites web présents permettait d'avoir des résultats très satisfaisants, mais vu l'explosion du nombre de sites web présents actuellement ainsi que la fréquence de leurs mises à jour, ces algorithmes se voient limités dans un ensemble restreint de sites web à visiter (Viscousness Phenomenon).
Le modèle proposé par Balaji et Sarumathi avec leur application Topcrawl [1], un exemple de ce type de stratégies, calcule le score de pertinence des URL selon 3 critères :

- La quantité de données extraites.
- Leur qualité.
- Leur fraîcheur.

La figure suivante montre le processus de crawling adopté par Topcrawl :

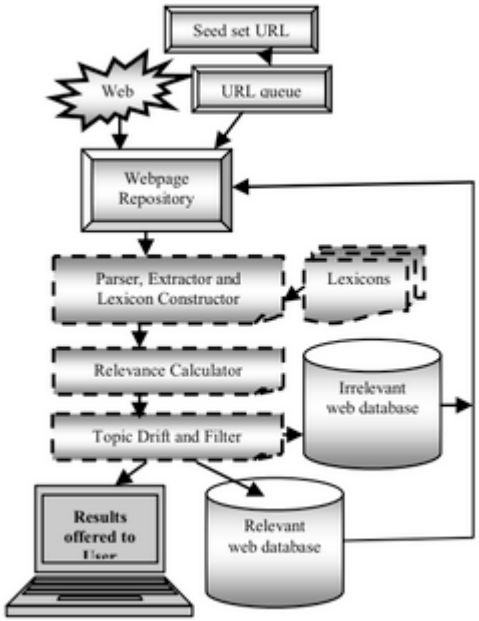


Figure 2 : Processus adopté par Topcrawl [1]

2.2.2 Stratégies basées sur l'évaluation des URLs

Ce type de stratégies est basé sur des algorithmes tels que PageRank [6] qui évolue la pertinence d'une page à partir du nombre des URLs qui lui pointent dessus. Une page est bien classée si la somme des classements des liens qui lui pointent dessus est élevée. Ceci peut être soit si le nombre des liens est grand, soit si le nombre de liens est petit mais que ceux-ci sont bien classés.

L'avantage de cette méthode réside dans la facilité d'établir la cartographie des hyperliens, mais l'inconvénient reste la haute complexité ainsi que le haut degré de dérapages en hors sujet, qui sont surtout causés par le fait d'ignorer l'analyse du contenu.

2.2.3 Stratégies combinant les deux précédentes

La combinaison des stratégies précitées a intéressé plusieurs chercheurs. En effet, une méthode avait été proposée [7] afin d'analyser la structure des liens combinée à une analyse de contenu à travers le Shark Search, sauf qu'elle ne parvient toujours pas à éviter le phénomène de Viscousness mais Liu [7] avait proposé une méthode inspirée par des stratégies de contrôle des vitesses d'accès pour crawlers pour pallier à cette problématique, Chen et al. [8] avaient aussi proposé une méthode basée sur l'algorithme génétique afin de créer une bouscule par mutation en cas de Viscousness.
Luo et al. [9] avaient proposé une amélioration du Shark Search combinant d'un côté l'analyse des poids et des structures des liens, et d'un autre côté une méthode de contrôle d'accès hôtes basée sur le Fish Search. Avec cette méthode, chaque accès de crawler est enregistré et le poids l'hôte est décrémenté s'il est fréquemment visité, et comme ça le phénomène de Viscousness pourrait être évité.

3 Scrapers

Le scraping, ou autrement appelé haversting, est une technique qui permet d'extraire du contenu web afin de le réutiliser pour d'autres fins, de Data Mining par exemple.

3.1 Principe

Contrairement au crawler qui parcoure et analyse les sites web pour des fins de reconnaissance et de cartographie du web, le scraper récupère les données, souvent non-structurées d'un site prédéfini et les transforme en données exploitables.
Vu la différence frappante des structures des sites web et la tendance de ceux-ci à changer régulièrement de structure, même si des fois, cela n'est visible à l'utilisateur final, il est nécessaire de faire une analyse préalable au site voulu avant de se lancer dans le processus de scraping. Ceci-dit, contrairement aux crawlers qui peuvent visiter n'importe quel site et en extraire les données nécessaires, les scrapers sont beaucoup plus dédiés et leur code peut se voir modifié si le site à scruter change de structure.
La figure 3 présente une capture d'écran d'un journal publié par Springer, un scraper peut, par exemple, récupérer le titre du journal, son ISSN et le nombre des articles qu'il contient et les stocker en base de données (Tableau 1).

Titre	ISSN	NbreArticles
Data Mining and Knowledge Discovery	1573-756X	558

Tableau 1 : Table récupérée depuis Springer par un scraper

Figure 3 : Capture d'écran d'un journal publié par Springer

3.2 Techniques

Vu les limitations auxquelles les scrapers doivent faire face, on peut en trouver différentes techniques de scraping. Cependant, on pourrait rassembler les techniques les plus courantes sous les catégories suivantes :

- Le « copier-coller » traditionnel : Malgré sa lourdeur, elle reste parfois la meilleure façon de récupérer l'information recherchée.
- « grep » : Cette commande Linux est un outil simple mais très performant pour l'extraction de l'information utile depuis un fichier HTML.
- Parseurs HTML : Technique très utilisée pour des fins de Data Mining dans laquelle des programmes appelés « wrappers » détectent la structure de la page HTML, la parcourent pour récupérer l'information utile puis la stockent dans une base de données.

Plusieurs tentatives ont été envisagées afin d'automatiser le processus de création de wrappers capables de s'adapter aux pages web qu'elles doivent scruter. Ces tentatives avaient fait l'objet d'une étude menée par Chang et al. [10] qui ont présenté et comparé plusieurs générateurs de wrappers selon leurs domaines de recherche, leur degré d'autonomie, ainsi que les techniques qu'ils utilisent. Ils ont pu distinguer quatre types de générateurs :

- Générateurs manuels : Avec cette approche, l'utilisateur est contraint de développer à chaque fois un wrapper spécifique pour chaque site web, comme par exemple avec TSIMMIS [11] qui représente l'une des premières approches de construction de générateurs de wrappers.
- Générateurs supervisés : Basés sur des algorithmes d'apprentissage supervisé, l'utilisateur leur fournit des exemples très exacts de pages contenant l'information à extraire, SRV [12] et RAPIER [13] sont deux exemples de ce type de générateurs.

- Générateurs semi-supervisés : Contrairement à leurs confrères supervisés, les générateurs semi-supervisés peuvent se contenter d'exemples moins stricts afin de lancer leur recherche. IEPAD [14], par exemple, n'a pas besoin de pages d'exemples d'apprentissage labellisées (un effort supplémentaire que l'utilisateur devait fournir avec les générateurs supervisés afin qu'il leur précise les données à extraire).
- Générateurs non-supervisés : Ces générateurs n'ont besoin d'aucun exemple d'apprentissage ni d'aucune interaction avec l'utilisateur pour générer le wrapper. RoadRunner [15] par exemple, a été conçu pour accomplir des tâches d'extraction au niveau des pages, tandis que DeLa [16] s'occupe des extractions au niveau des enregistrements.

4 XEW-SS USPTO

L'objectif de ce niveau architectural est de fournir une description complète de l'ensemble du processus de traitement de données issues des différentes sources. Pour cela, les techniques employées s'appuient sur des agents intelligents de crawling et scraping adaptés à chaque source d'information.
Les différents agents se basent sur l'architecture microservices pour faciliter la communication avec les autres services de l'architecture Big Data de XEW 2.0. La présentation détaillée des agents XEW fait l'objet du prochain chapitre.

Ce service permet la recherche, la collecte et le traitement des données issues de différentes sources. Il doit prendre en compte des techniques de fusion multi-modale. Cela, dans un souci de supporter l'hétérogénéité, l'imprécision et l'incertitude qui entachent les données multi-sources. Cette prise en compte de fusion assure une maîtrise des connaissances et des informations, et par conséquent, facilite amplement la prise de décision. SS-XEW traite l'hétérogénéité des informations, d'un point de vue contenu sémantique (scientifique, technique, ...), structurel (fortement structuré (brevet) à non structuré (e-mails)), linguistique (multilinguisme), format du support (Word, html, PDF, ...), taille : définition de l'unité d'information à analyser (granularité de l'information).

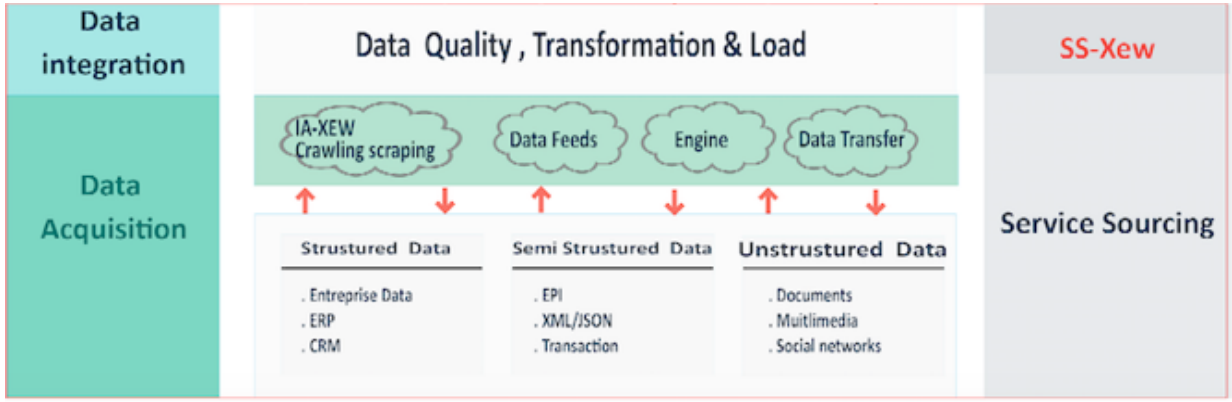


Figure 4 : Service de sourcing

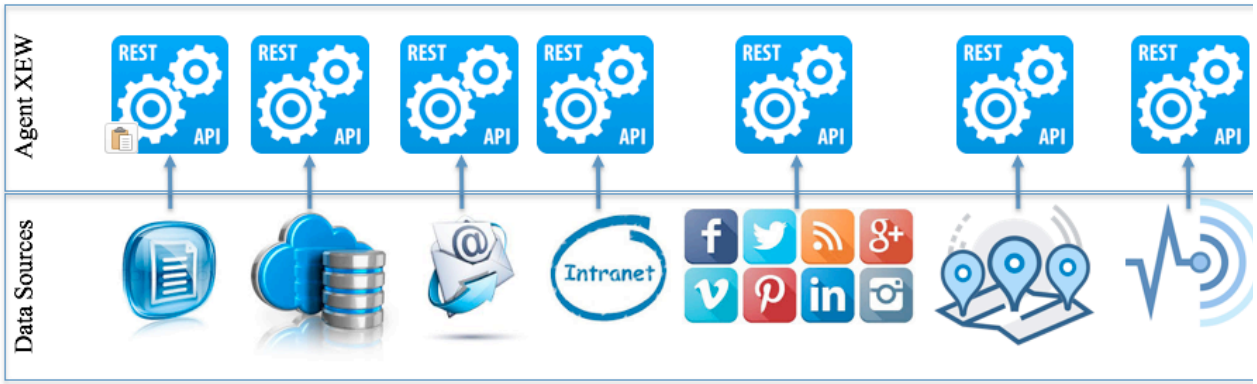


Figure 5: Gestion des multi-sources par l'agent XEW

XEW agent se base sur les web services REST et le modèle acteur de Framework AKKA (figure 6). Il nous permet un accès direct et en temps réel à des données tirées de l'exploration des milliers de sources en ligne, pour la veille scientifique ou technologique (figure 7).

Le Scraper XEW fournit un éditeur basé sur un navigateur pour configurer des robots et extraire des données en temps réels, il scanne constamment les sources désignées et trouve des mises à jours pour des données dans différents formats et langues. Le workflow de la 3.13 décrit en détail le flux d'information de XEW Scraper.
Après le choix des sources d'information pour un sujet d'analyse, l'agent scraper possède au choix de scénario de collecte des données, parmi l'ensemble des scenarii possible selon plusieurs critères. Un scénario est définie par :

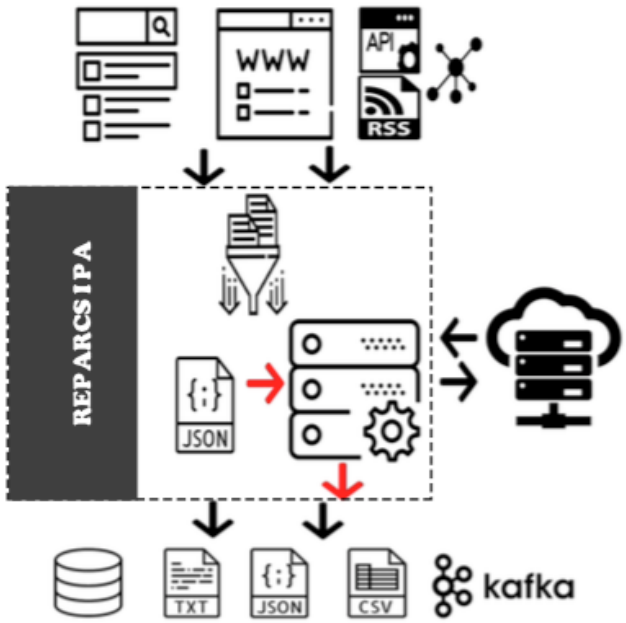


Figure 6: L'architecture de XEW Agent

Définition
Un scénario est un ensemble fini d'opérations organisées selon un ordre d'exécution bien établie, on note :

SC = {IN,OUT,ListOp,P}

Avec,

- IN : l'ensemble des entrées de scenarii.
- OUT : les résultats.
- ListOp : un ensemble d'opérations de 1..n
- P : politique d'exécution des opérations

Selon le type de la source, un scénario adéquat est établi pour la collecte des données qui se base sur les règles d'extraction définies dans les opérations, sous forme de fichier JSON. Ce dernier est envoyé a l'agent scraper pour le transformer en fonction puis l'encapsuler dans une tâche de création d'agent spécifique de scraping de la source cible.

A titre d'exemple, voici l'exemple de l'agent USPTO

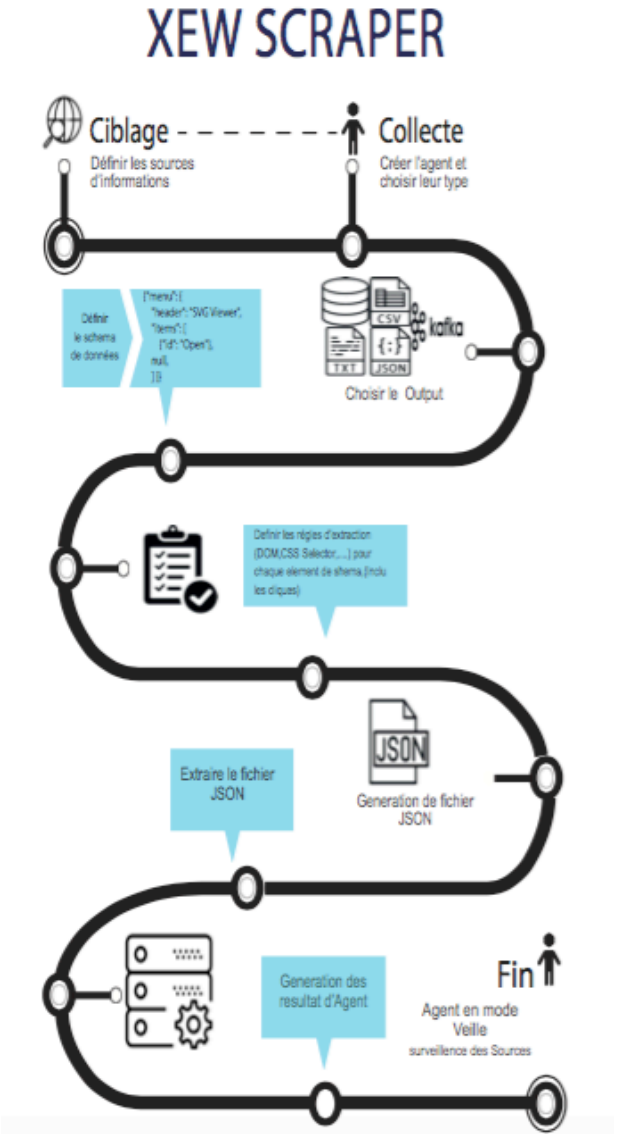


Figure 7: La workflow de XEW Scraper

Logout

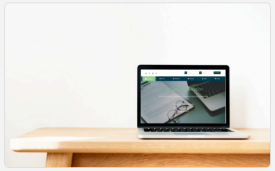
XEW 2.0 Explore Every Where

ACCUEIL | PROCESSUS | SERVICES | S'INSCRIRE | CONTACT

SS-XEW - Service de Sourcing

Le Scraper XEW fournit un éditeur web pour configurer des robots qui peuvent extraire des données en temps réels, ils scannent constamment les sources désignées et trouvent des mises à jours pour des données dans différents formats et langues. Jusqu'à maintenant nous avons développé et validé des agents pour des sources d'articles scientifiques comme IEEE, Springer, Sciencedirect, ACM, PubMed et des bases de données de brevets : USPTO, WIPO, Esp@cenet.

Termes d'utilisation



5+	10+	50+	20+
Partenaires	Projets réalisés	Clients satisfaits	Meetings

Choisir votre Source

Le Big Data, une source d'information sans limite pour l'Intelligence économique

USPTO

Patents & Trademarks

L'Office des brevets et des marques des États-Unis (USPTO) est une agence du Département du commerce des États-Unis qui délivre des brevets aux inventeurs et aux entreprises pour leurs inventions et l'enregistrement des marques.

Chercher

PubMed

Biomedical literature

PubMed contient plus de 28 millions de citations biomédicale de MEDLINE, des revues de sciences de la vie, et des livres électroniques. Les citations peuvent inclure des liens vers les texte intégral de PubMed ou des sites Web d'éditeurs.

Chercher

ScienceDirect

Scientific research

ScienceDirect est un site web offrant un accès par abonnement à une vaste base de données de recherches scientifiques et médicales. Il héberge plus de 12M de contenus provenant de 3500 revues académiques et 34000 e-books.

Chercher

WIPO

Patents & Trademarks

Organisation mondiale de la propriété intellectuelle (OMPI), c'est une organisation internationale destinée à promouvoir la protection mondiale de la propriété industrielle (inventions, marques et modèles) et des droits d'auteur.

Chercher

Coup de cœur

Créer une approche de recherche ouverte est une tâche difficile. Il ne suffit pas de planifier un processus ouvert, mais la mise en œuvre doit également suivre. En d'autres termes, si en théorie la recherche est participative, mais que personne n'y participe, nous ne pouvons pas vraiment dire que nous avons un processus de recherche ouvert. Mais chez XEW, nous l'avons déjà !

Architecture SS-XEW

newsletters

Entrez votre email et recevez-vous toute news

Entrer votre e-mail

Subscribe

Follow Us

Contact

Pour toute question, Nous sommes là !

Al-Hoceima, Morocco
Phone: +212 670656881
Email: xplorew@contact.org

Contact

© All rights reserved for XEW 2.0.

Logout

XEW 2.0 Explore Every Where

ACCUEIL | PROCESSUS | SERVICES | S'INSCRIRE | CONTACT

United States Patent and Trademark Office

Le United States Patent and Trademark Office (USPTO), littéralement le Bureau américain des brevets et des marques de commerce, est l'instance administrative chargée d'émettre des brevets et des marques déposées aux États-Unis. Il est considéré comme le plus important bureau dans le domaine des brevets, surtout à cause de la taille économique du marché américain.

Termes d'utilisation

PROCESSUS XEW-SS

Veuillez suivre ces étapes pour avoir ce que vous désirez !

1 Recherche

Rechercher les brevets à analyser et visualiser .

2 Résultats

Une fois la recherche est terminée, vous pouvez afficher tous les brevets trouvés et les impimer aussi.

3 Téléchargement

Vous pouvez télécharger les brevets trouvés en format texte lisible, nettoyés et bien structurés.

4 Analyse

Revoir les brevets avec une autre philosophie

newsletters

Entrez votre email et recevez-vous toute news

Entrer votre e-mail

Subscribe

Follow Us

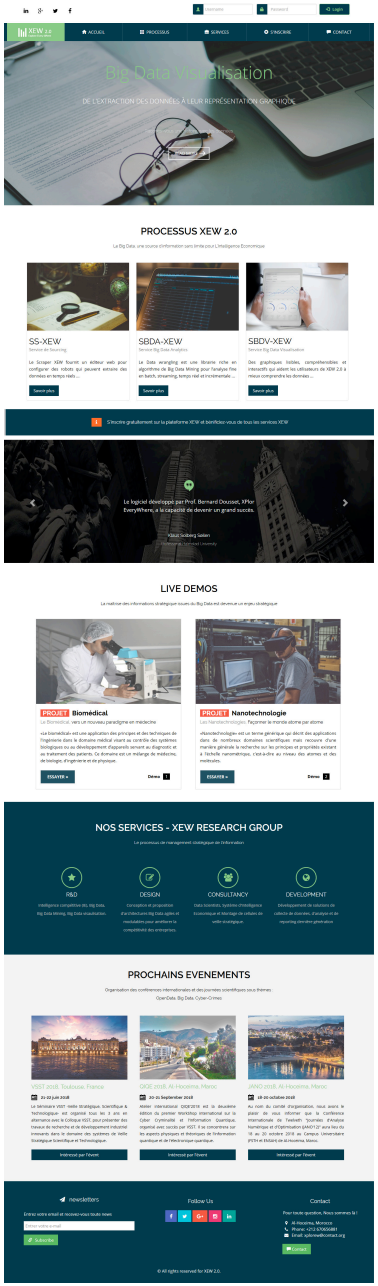
Contact

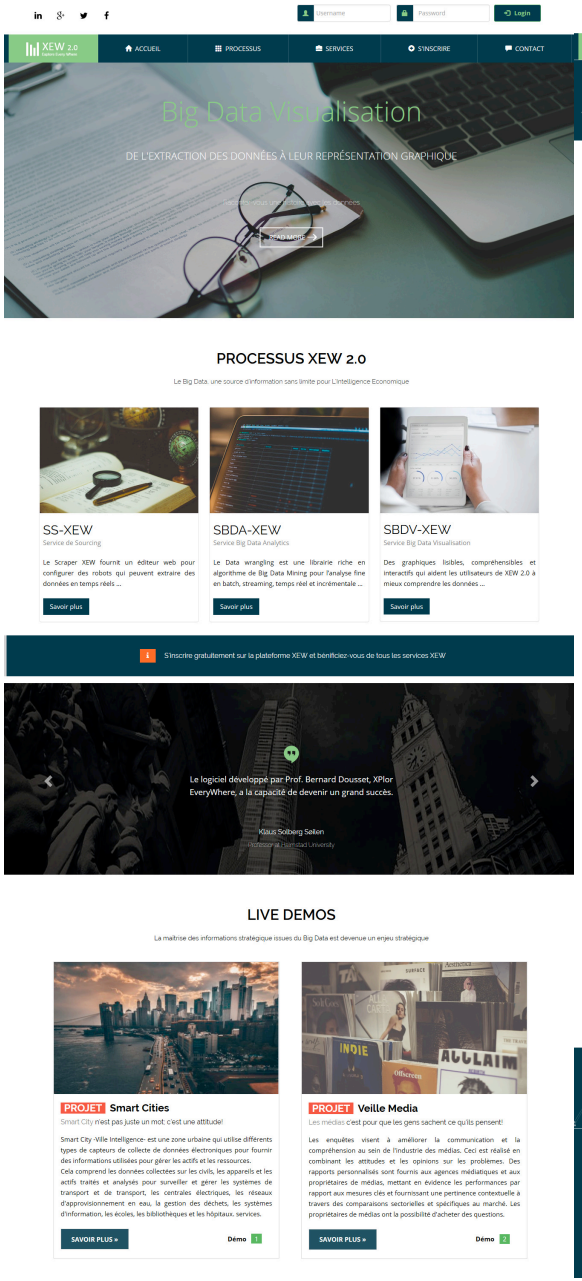
Pour toute question, Nous sommes là !

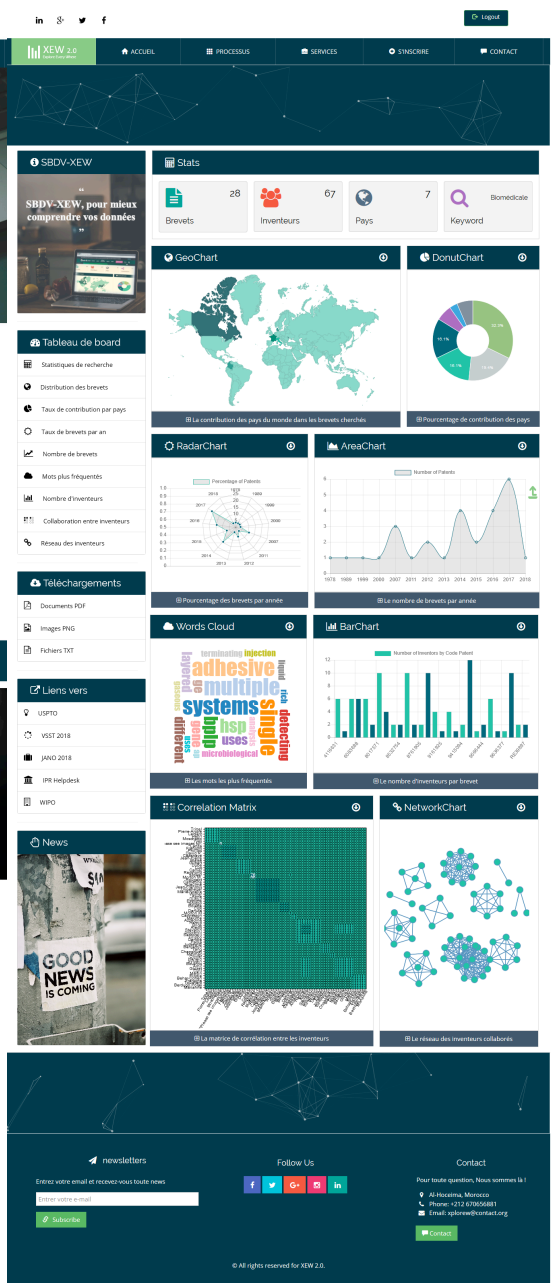
Al-Hoceima, Morocco
Phone: +212 670656881
Email: xplorew@contact.org

Contact

© All rights reserved for XEW 2.0.


XEW 2.0
ACCUEIL
PROCESSUS
SERVICES
S'INSCRIRE
CONTACT
Big Data Visualisation
DE L'EXTRACTION DES DONNÉES À LEUR REPRÉSENTATION GRAPHIQUE
PROCESSUS XEW 2.0
SS-XEW
SBDA-XEW
SBDV-XEW
S'inscrire gratuitement sur la plateforme XEW et bénificiez-vous de tous les services XEW
Le logiciel développé par Prof. Bernard Dousset, XPlor EveryWhere, a la capacité de devenir un grand succès.
LIVE DEMOS
PROJET Biomédical
PROJET Nanotechnologie
NOS SERVICES - XEW RESEARCH GROUP
R&D
DESIGN
CONSULTANCY
DEVELOPMENT
PROCHAINS EVENEMENTS
newsletters
Follow Us
Contact


XEW 2.0
ACCUEIL
PROCESSUS
SERVICES
S'INSCRIRE
CONTACT
Big Data Visualisation
DE L'EXTRACTION DES DONNÉES À LEUR REPRÉSENTATION GRAPHIQUE
PROCESSUS XEW 2.0
Le Big Data, une source d'information sans limite pour L'Intelligence Economique
SS-XEW
Service de Sourcing
Le Scraper XEW fournit un éditeur web pour configurer des robots qui peuvent extraire des données en temps réels ...
Savoir plus
SBDA-XEW
Service Big Data Analytics
Le Data wrangling est une librairie riche en algorithme de Big Data Mining pour l'analyse fine en batch, streaming, temps réel et incrémentale ...
Savoir plus
SBDV-XEW
Service Big Data Visualisation
Des graphiques lisibles, compréhensibles et interactifs qui aident les utilisateurs de XEW 2.0 à mieux comprendre les données ...
Savoir plus
S'inscrire gratuitement sur la plateforme XEW et bénificiez-vous de tous les services XEW
Le logiciel développé par Prof. Bernard Dousset, XPlor EveryWhere, a la capacité de devenir un grand succès.
Klaus Solberg Sellen
LIVE DEMOS
La maitrise des informations stratégique issues du Big Data est devenue un enjeu stratégique
PROJET Smart Cities
Smart City n'est pas juste un mot: c'est une attitude!
SAVOIR PLUS »
Démo 1
PROJET Veille Media
SAVOIR PLUS »
Démo 2


XEW 2.0
ACCUEIL
PROCESSUS
SERVICES
S'INSCRIRE
CONTACT
Logout
SBDV-XEW
SBDV-XEW, pour mieux comprendre vos données
Stats
Brevets 28
Inventeurs 67
Pays 7
Keyword Biomédicale
Tableau de board
Statistiques de recherche
Distribution des brevets
Taux de contribution par pays
Taux de brevets par an
Nombre de brevets
Mots plus fréquentés
Nombre d'inventeurs
Collaboration entre inventeurs
Réseau des inventeurs
Téléchargements
Documents PDF
Images PNG
Fichiers TXT
Liens vers
USPTO
VSST 2018
JANO 2018
IPR Helpdesk
WIPO
News
GOOD NEWS IS COMING
GeoChart
Le contribution des pays du monde dans les brevets cherchés
DonutChart
Pourcentage de contribution des pays
RadarChart
Pourcentage des brevets par année
AreaChart
Le nombre de brevets par année
Words Cloud
Les mots les plus fréquentés
BarChart
Le nombre d'inventeurs par brevet
Correlation Matrix
La matrice de corrélation entre les inventeurs
NetworkChart
Le réseau des inventeurs collaborés
newsletters
Follow Us
Contact
© All rights reserved for XEW 2.0.

5 Conclusion

Le SS-XEW génère des données dans différents format (XML, JSON, TXT, script SQL, script NoSQL) et dans différents modèles de données. Ceci nous a conduit à l'amélioration de XEW et à son adaptation aux Big Data, avec la possibilité de générer des modèles de données pour plusieurs utilisations qui incluent différents systèmes : OLTP, OLAP, NoSQL et MPP dans un environnement distribué. Le SDW-XEW nous permet une gestion des données massives et évolutives grâce à deux possibilités de stockage complémentaires : sur le serveur XEW2.0 et dans le Cloud.

6 Bibliographie

[1] BALAJI S. et SARUMATHI S., *TOPCRAWL: Community Mining in Web search Engine with emphasize on Topical Crawling*, Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, IEEE, March 2012, p 20-24

[2] EFIMOVA L. et FIEDLER S., *Learning webs : Learning in weblog networks*, Proceedings of the IADIS International Conference Web Based Communities Lisbon, IADIS Press, March 2004, p 490-494

[3] OUYANG L. B., LI X. Y., LI G. H. et al., *A survey of web spiders searching strategies of topic-specific search engine*, Computer Engineering, 2004, p 32-46

[4] BRA D. P. et POST R., *Searching for arbitrary information in the WWW: the fish-search for mosaic*, In: Second WWW Conference, ACM Press, 1994, p 45-51

[5] HERSOVICI M., JACOVI M., MAAREK Y. S., PELLEG D., SHTALHAIM M., UR S., *The Shark-Search Algorithm. An Application: Tailored Website Mapping*, Computer Networks and ISDN Systems, Elsevier 1998, p 317-326

[6] PAGE L., BRIN S. et MOTWANI R., *The PageRank Citation Ranking: Bring Order to the Web*, Stanford University, 1998

[7] LIU Y. F., *Focus crawler searching in research engine*, SUN Yat-Sen University, Guangzhou, 2005

[8] CHEN Y. F., ZHAO H. K., YU X. Q. et WAN W. G., *Improvement of focused crawling strategy based on genetic algorithm*, Computer Simulation, 2010, p 87-90

[9] LUO L., WANG R. B., HUANG X. X. et CHEN Z. Q., *A Novel Shark-Search Algorithm for Theme Crawler*, WISM, LNCS, Springer, 2012, p 603-609

[10] CHANG C. H., KAYED M., GIRGIS M. R. et SHAALAN K., *A Survey of Web Information Extraction Systems*, IEEE Transactions on Knowledge and Data Engineering, 2006, p 1411-1428

[11] HAMMER J., MCHUGH J. et GARCIA-MOLINA, *Semi-Structured data: the TSIMMIS experience*, Proceedings of the 1st East-European Symposium on Advances in Databases and Information Systems (ADBIS) St-Petersburg Russia, 1997, p 1-8

[12] FREITAG D., *Information Extraction from HTML: Application of a general learning approach*, Proceedings of the Fifteenth International Conference on Artificial Intelligence (IJCAI), 1997, p 729-735

[13] CALIFF M. et MOONEY R., *Relational learning of pattern-match rules for information extraction*, Proceedings of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing, Stanford California, 1998

[14] CHANG C. H. et LUI S. C., *IEPAD: Information extraction based on pattern discovery*, Proceedings of the Tenth International Conference on World Wide Web, Hong-Kong, 2001, p 223-231

[15] CRESCENZI V., MECCA G. et MERIALDO P., *RoadRunner, towards automatic data extraction from large Web sites*, Proceedings of the 26th International Conference on Very Large Database Systems (VLDB), Rome Italy, 2001, p 109-118

[16] WANG J. et LOCHOVSKY F. H., *Data extraction and label assignment for Web databases*, Proceeding of the Twelfth International Conference on World Wide Web, Budapest Hungary, 2003, p 187-196