République Algérienne Démocratique et Populaire
Université Abou Bakr Belkaid– Tlemcen
Faculté des Sciences
Département d'Informatique
Mémoire de fin d'études

Pour l'obtention du diplôme de Master en Informatique

Option: *Système 'Information et de Connaissances (S.I.C)*

Thème

Synthèse d'approches pour la liaison d'information échangée dans le Web (application dans le cadre des réseaux sociaux).

Réalisé par :

- **Benali Medjahed Fatima Zohra**
- **Belkhodja Amina**

Présenté le 03 Juillet 2011 devant le jury composé de MM.

Président :	- *Mme Didi F.*		
Encadreur :	- *Mr Badr Banmammar*		
Examinateurs :	- *Mr Chouiti S.*		
	- *Mr Midouni S.D*	-	*Mme El Yebdri Z.*
	- *Mme Hlfaoui A.*	-	*Mme Khitri*

Année universitaire: 2010-2011

Remerciement

Remerciement

Notre remerciement va en premier lieu à ALLAH le tout puissant de nous avoir donné la foi et de nous avoir permis d'en arriver là.

Nous tenons à remercier particulièrement notre encadreur Mr Badr Benmammar pour son encadrement et pour l'intérêt qu'il a manifesté a notre travail.

Nous remercierons très sincèrement, les membres de jury d'avoir bien voulu accepter de faire partie de la commission d'examinateur.

Nous adressons également notre remerciements, à tous notre enseignants, qu'ils ont consentis pour nous permettre de suivre notre études dans les meilleures conditions possibles et n'avoir jamais cessez de nous encourager tout au long de nos années d'étude.

Nous tenons également à remercier tous nos collègues de promotion que nous avons eu le plaisir de les côtoyer pendant cette période de formation.

Nous remercierons tous ceux qui ont contribué de près ou de loin à la réalisation de ce travail.

Dédidaces

Dédidaces

A la mémoire de mes parents qui ont souhaités vivre pour longtemps juste pour nous voir Qu'est-ce que nous allons devenir.

A ma tante « yahiaouia » qui ma éclairé mon chemin et qui mon encouragé et soutenue toute au long de mes études.

A mes chers tantes : Houaria, Khaira, Khadija, Malika et sans oublier l'esprit de Fatima (que je me demande que dieu fait pitié).

A mes chers frères Mustapha, Saïd, Omar, Fethi, Omar, Sid Ahmed.

A m'adorable sœur Kamila.

A qui sont proche de mon cœur et que j'aime très fortes, a mes sœurs Lamia et Karima.

A qui ma compagne a tous moment pour réaliser ce modeste travail a mon binôme : Benali medjahed Fatima Zohra

A qui m'ont donné un magnifique modèle de bonheur, d'amitié et d'amour ; a tous qui dit unité *liberté*travail et plus spécialement le bureau de Wilaya, de Talab abd Rahman et de Chatouane.

A l'esprit de « Hassan El Bâna » qui a toujours été mon pays d'amour, de soutien et d'espoir.

À toutes les personnes

qui connaissent

« Amina » de prés ou

de loin.

Didicases

« Louange à Dieu, le tout puissant »

A mes très chers parents,

Pour leur soutien permanent et inépuisable,

Que Dieu les protège.

A mes chers frères : Ahmed, Ameur , Abdelouahab et Mustapha

Que Dieu les bénisse.

A mes chères sœurs Nacéra et Nassima

A mes neveux : Yasser, Alaa ,Loay, Djawad et Fares

A ma nièce Kawthare Ferdousse

A mon encadreur Mr Badr Banmammar

Que Dieu exauce ses vœux les plus chers.

A mes tantes, mes oncles, mes cousins et cousines.

A qui sont proche de mon cœur et que j'aime très fort, à mes copines Fatima et Nour El Houda

A qui m'accompagné a tous moment pour réaliser ce modeste travail a mon binôme : Belkhodja Amina

Et à tous ceux qui me sont chers.

A TOUS, je dédie ce travail en leur adressant tous mes sentiments D'affection et de considération.

Table de matière:

Liste des figures:

Liste des tableaux:

Liste des acronymes:

Acronyme	Signification
JSP	Java Server Pages
HTML	HyperText Markup Language
CSS	Cascading Style Sheets
XML	eXtensible Markup Language
RSS	Really Simple Syndication
AJAX	Asynchronous Javascript and XML
API	*Application Programming Interface*
REST	Representational State Transfer
XHTML	eXtensible HyperText Markup Language
JSON	JavaScript Object Notation
XML-RPC	eXtensible Markup Language- Remote procedure call
SOAP	*Simple Object Access Protocol*

W3C	World Wide Web Consortium
RDF	*Resource Description Framework*
SGML	*Standard Generalized Markup Language*
ISO	*Institut Supérieur d'Optique*
TMQL	Topic Maps Query language
UDDI	Universal Discription , Discovery and Integration
WSDL	*Web Services Description Language*
DAML-S	Semantic markup language for describing web services and related ontologies
XQuery	langage de requête
XDI	eXtensible Data Interchange
XRI	Extensible Resource Identifier
MySQL	système de gestion de base de données
HTTP	HyperText Transfer Protocol
FTP	*File Transfer Protocol*

I. Introduction :

1. Motivations :

Le terme "Web 2.0" est la "plate forme de données partagées via le développement d'applications qui viennent architecturer les réseaux sociaux issus de la contribution essentielle des usagers à la création des contenus et des formats de publication" (blogs, wikis…). La clef du succès dans cette nouvelle étape de l'évolution du web réside dans

l'intelligence collective. Le Web 2.0 repose sur un ensemble de modèles de conception : des systèmes architecturaux plus intelligents qui permettent aux gens de les utiliser, des modèles d'affaires légers qui rendent possible la syndication et la coopération des données et des services. Le Web 2.0 c'est le moment où les gens réalisent que ce n'est pas le logiciel qui fait le web, mais les services [31].

Parmi les services du Web 2.0, nous avons pris en considération les agrégateurs des réseaux sociaux (Netvibes,..) qui permettent à son utilisateur de retrouver sur la même interface un ensemble de réseaux sociaux. Cela évite à l'utilisateur de se connecter à chaque réseau social et d'être au courant de ce qui se passe dans tel ou tel réseau avec la seule interface de l'agrégateur, donc avec un seul coup d'œil, vous pouvez faire le tour de vos contacts.

Mais ces derniers ne donne pas la pertinence des informations agrégées, ce qui nous à ramener à les développés pour obtenir des agrégateurs évolués qui indique l'importance du besoin de l'utilisateur pour chaque réseaux sociaux.

2. Contribution :

Dans le cadre de ce PFE, nous avons développés un nouveau modèle qui consiste à orienter les utilisateurs vers la bonne source d'information. Ce modèle est basé sur l'utilisation des pages web dynamiques (JSP) ainsi que les parsers HTML.

Deux types de parsers HTML ont été nécessaires afin de réaliser notre application :

- Le premier parser a comme objectif de parser le contenue de l'agrégateur pour tester l'existence de la requête de l'utilisateur (à ce niveau, on traite la recherche d'information).
- Le deuxième parser sert à parser le code source de cet agrégateur pour calculer le nombre d'occurrence de cette requête dans chaque réseau social (à ce niveau on traite la fouille de texte).

Enfin, nous avons visualisé à l'utilisateur le résultat final qui va l'orienter vers le réseau social qui répond le mieux à son besoin.

3. L'organisation de notre travail :

Ce travail est structuré comme suit :

Le premier chapitre est consacré à l'introduction du web, commençons par le Web 2.0 qui désigne des nouveaux services liées à des nouvelles technologies et un nouveau

rôle pour l'internaute. Puis nous avons introduit le Web sémantique qui est une extension du Web qui facilite l'automatisation du traitement des connaissances disponibles. Finalement, le Web social qui est la partie la plus intéressante de tous le Web.

Le deuxième chapitre présente la fouille de texte qui définie comme un ensemble de traitements informatiques consistant à extraire des connaissances selon un critère de nouveauté ou de similarité.

Le troisième chapitre présente les agrégateurs de réseaux sociaux qui proposent au visiteur un large choix de sources d'information. Dans cette phase nous avons étudié quelque agrégateurs (Freindfeed, Netvibes, Spokeo et Seesmic).

Le quatrième chapitre présente notre application dans le cadre des réseaux sociaux, dans la quelle nous avons guidé l'utilisateur vers la bonne source d'information en fonction de son besoin.

Chapitre I:
Introduction du web

I.1. Web 2.0 :

I.1.1. Introduction :

Le Web 2.0 était annoncé comme une véritable révolution de l'Internet, une mutation qui allait influencer notre mode de vie, notre manière de communiquer et de travailler. Aujourd'hui, il commence à faire tomber son masque et le mythe est devenu un concept banal

qui n'a finalement rien d'exceptionnel. Le changement n'a pas été aussi brusque que le mot « révolution » pourrait laisser croire. La mutation s'est faite lentement, que ce soit pour la technologie, pour la pratique sociale et pour l'économie. Le Web 2.0 a déjà été pratiqué avant qu'il ne soit appelé ainsi. Le basculement technique a commencé depuis la démocratisation des outils de gestion de contenus pour créer et pour gérer des sites Web dynamiques. Cela a donné à tous, le moyen d'avoir une présence sur Internet. En quelques clics, on peut avoir non simplement un support Web mais aussi la possibilité de faire une mise à jour des contenus. Les autres outils de mise en relation existaient déjà mais le changement se trouve dans la possibilité de faire de l'agrégation et du partage de contenus et de services. Cette réorganisation de l'information respecte une certaine sémantique [1].

I.1.2. Définition :

Pour le concept de « Web 2.0 » il existe plusieurs définitions :

L'expression « Web 2.0 » désigne certaines technologies et des usages du *World Wide Web* qui ont suivi la forme initiale du Web, en particulier les interfaces permettant aux internautes ayant peu de connaissances techniques de s'approprier les nouvelles fonctionnalités du Web [2].

Considéré comme l'évolution naturelle du web actuel, le Web 2.0 est un concept d'utilisation d'internet qui a pour but de valoriser l'utilisateur et ses relations avec les autres. Le Web 2.0 était annoncé comme une véritable révolution de l'Internet, une mutation qui allait influencer notre mode de vie, notre manière de communiquer et de travailler [3].

Le terme Web 2.0 est généralement utilisé pour désigner une évolution d'un Web statique et unidirectionnel vers un réseau dynamique et interactif, caractérisé par une large participation des usagers à la création et à l'échange de contenus. En d'autres mots, nous pourrions dire que c'est un synonyme du Web participatif. Il est souvent associé à l'utilisation de logiciels sociaux en ligne, tels que les blogues, wikis, micro blogues ou autres réseaux sociaux [4].

I.1.3. Origine du terme :

L'expression a été médiatisée en août 2003 par Dale Dougherty de la société O'Reilly Media lors d'une conversation avec Craig Cline de MediaLive en vue de préparer une

conférence. Il a suggéré que le Web était dans une période de renaissance ou mutation, avec un changement de paradigmes et une évolution des modèles d'entreprise. Dougherty a donné des exemples au lieu de définitions : « Double Click, c'était le Web 1.0. Google AdSense, c'est le Web 2.0. Ofoto, c'était le Web 1.0. Flickr, c'est le Web 2.0. », et recruté John Battelle. Puis, O'Reilly Media, Battelle et MediaLive ont lancé la première conférence Web 2.0 en octobre 2004. La seconde conférence annuelle a eu lieu en octobre 2005. Cette dernière réalise les aspects suivants :

- Le Web comme plate-forme ;
- Les données comme « connaissances implicites » ;
- Les effets de réseau entraînés par une « architecture de participation », l'innovation comme l'assemblage de systèmes et de sites distribués et indépendants ;
- Des modèles d'entreprise poids plument grâce à la syndication de contenus et de services [2].

I.1.4. Limites ergonomiques du Web 1.0 :

- Pour chaque action : effacement de la page HTML en cours, réalisation d'une requête synchrone vers le serveur et affichage de la nouvelle page HTML.
- Ne permet pas de concurrencer les applications clients lourds en termes d'ergonomie.

Mais les applications clients lourds ont d'autres inconvénients (déploiement, maintenance, sécurité, etc...) [5].

I.1.5. Web 2.0 : un nouveau modèle de développement :

Amélioration de l'ergonomie en déplaçant une partie de la logique applicative vers le navigateur. Les outils et termes "Web 2.0" sont pour certains bien connus du grand public, parfois sans pour autant être rattachés à la catégorie Web 2.0. D'autres ont une diffusion plus confidentielle. Wikipédia, un des outils les plus emblématiques du Web 2.0, donne la définition de nombreux termes du Web 2.0 [6].

I.1.6. Les concepts de base :

I.1.6.1. Un blog ou blogue :

Est un site Web constitué par la réunion d'un ensemble de billets classés par ordre chronologique. Chaque billet (appelé aussi *note* ou *article*) est à l'image d'un journal de bord ou d'un journal intime, un ajout au blog ; le blogueur (celui qui tient le blog) y porte un texte,

souvent enrichi d'hyperliens et d'éléments multimédias, sur lequel chaque lecteur peut généralement apporter des commentaires [6].

I.1.6.2. Un wiki :

Est un système de gestion de contenu de site Web qui rend les pages Web librement et également modifiables par tous les visiteurs autorisés. On utilise les wikis pour faciliter l'écriture collaborative de documents avec un minimum de contraintes. Au milieu des années 2000, les wikis ont atteint un bon niveau de maturité ; ils sont depuis lors associés au Web 2.0. Créée en 2001, l'encyclopédie Wikipédia est devenue le wiki le plus visité au monde [6].

I.1.6.3. Really Simple Syndication :

Souscription vraiment simple ou le Rich Site Summary (sommaire développé de site), encore appelé flux RSS, fil RSS ou RSS feed en anglais, sous forme de sigles, est un format de syndication de contenu web, codé sous forme XML. Ce système est habituellement utilisé pour diffuser les mises à jour de sites dont le contenu change fréquemment, typiquement les sites d'information ou des blogs.

L'utilisateur peut s'abonner aux flux, ce qui lui permet de consulter rapidement les dernières mises à jour sans avoir à se rendre sur le site [6].

I.1.6.4. Un réseau social :

Quand peut aussi le qualifié de réseau humain est un ensemble de relations entre des individus. L'analyse des réseaux sociaux est l'approche scientifique en sciences sociales pour étudier les réseaux sociaux. Les réseaux sociaux sont aussi simplement considérés

comme étant la mise en relation de gens pour des fins amicales ou professionnelles. Il existe des applications internet aidant à se créer un cercle d'amis, de partenaires commerciaux ou autres [6].

I.1.6.5. L'intelligence collective :

Désigne les capacités cognitives d'une communauté résultant des interactions multiples entre des membres. Les éléments portés à la connaissance des membres de la communauté font qu'ils ne possèdent qu'une perception partielle de l'environnement et n'ont pas conscience de la totalité des éléments qui influencent le groupe. Des agents au comportement très simple peuvent ainsi accomplir des tâches apparemment très complexes grâce à un mécanisme fondamental appelé synergie. Sous certaines conditions particulières, la synergie

créée par la collaboration fait émerger des facultés de représentation, de création et d'apprentissage supérieures à celles des individus isolés [6].

I.1.6.6. Mashup :

Dans le cadre d'une superposition de deux images provenant de sources différentes, superposition de données visuelles et sonores différentes par exemple dans le but de créer une expérience nouvelle. Dans le cas de site Web, le principe d'un Mashup est donc d'agréger du contenu provenant d'autres sites, afin de créer un site nouveau. Pour ce faire, on utilise le plus souvent l'objet XMLHttpRequest, AJAX du côté client, et les API (ou les Services Web) des sites dont on mixe le contenu. Pour en savoir plus, consultez le glossaire du web 2.0 [6].

I.1.7. Comparaison du Web 1.0 et du Web 2.0 :

Le tableau suivant illustre une comparaison entre le Web 1.0 et la Web 2.0 :

	Web 1.0	**Web 2.0**
Interface	Pages web	Pages web, RSS, API REST
Type de données	Page	Objets
Format de données	HTML	XML
Système de liens	Liens hypertextes (HREF) reliant des pages	Flux RSS et API REST exposant des objets
Rôle du site	Concentrer un trafic d'utilisateurs	Concentrer un trafic de partenaires (mash-up)
Business model du site	Relation directe avec le consommateur	Grossiste: fournir une plate-forme technique à ses Partenaires
Paradigme	Sites HTML	Applications AJAX
Unité d'information	Site, page	Services, objets / flux /

		source de données (RSS)
Mode de navigation	De page en page via des liens	Ajout de composants et de sources d'information
	hypertexte	sur une page d'accueil personnalisée
Technologie	Pages HTML générées sur un serveur et affichées dans un navigateur	Client AJAX autonome s'exécutant dans le
		navigateur et puisant ses données dans des API
		et des flux RSS
Leaders du web	Entreprises, marchands	Internautes
Profil de l'internaute	Passif	Actif
Interactivité perçue	Sélection et lecture	Sélection, lecture et publication de données
	d'information	
Unité de recherche	Mot-clé	Tag

Tableau I.1 : Comparaison entre Web 1.0 et Web 2.0 [6]

I.1.8. Composants du web 2.0 :

Afin d'aller plus loin dans la compréhension du phénomène, une dissociation de composantes Web 2.0 semble indispensable.

- Trois familles d'applications bâties sur ce socle :

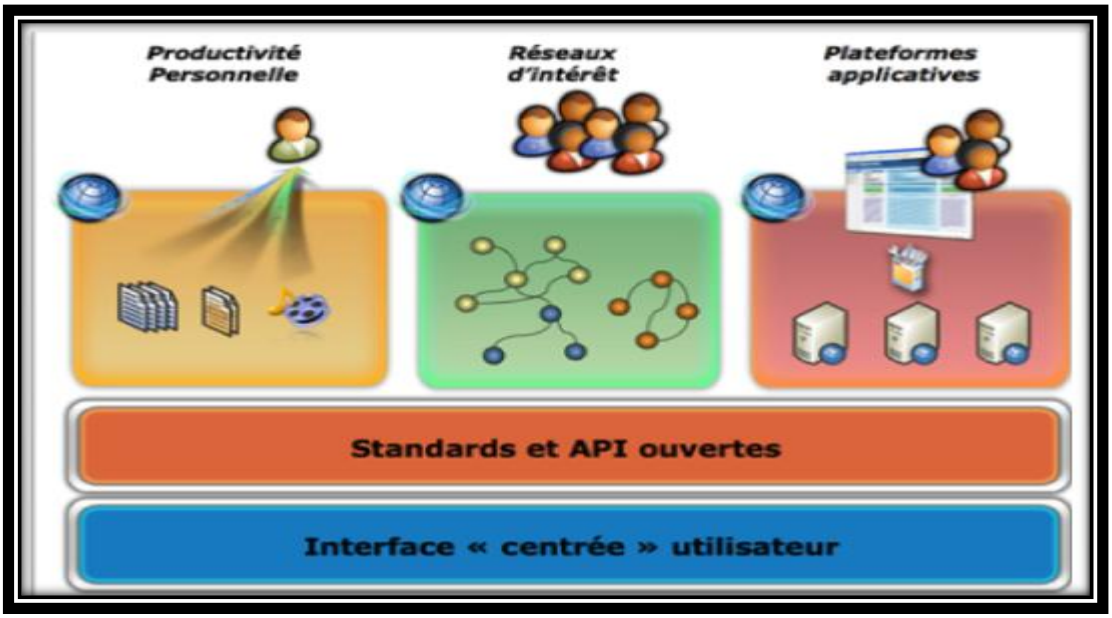


Figure I.1 : Composants du web 2.0. [7]

I.1.8.1. Interface « centrée » utilisateur :

L'architecture Ajax a popularisé deux changements majeurs dans la compréhension de ce que l'on peut faire avec du XHTML/CSS2/JavaScript :

❖ L'unité d'échange entre le client et le serveur n'est pas forcément la page entière. Il est ainsi possible d'échanger des petits flux d'information notamment suite à des évènements utilisateurs (touche pressée, sortie de champs…) comme en technologie client/serveur. Cela ouvre de nombreuses possibilités d'amélioration de l'expérience utilisateur sans aucun changement d'architecture : auto-complétion, glisser-déposer, édition en ligne des pages.

❖ Un navigateur peut agréger des informations provenant de plusieurs serveurs. Véritable claque au portail d'intégration, ce nouveau modèle d'architecture est validé à grande échelle par des acteurs tels que Netvibes et toute la démarche Open API dans laquelle s'engage les leaders du Web (Google, Yahoo, Flickr, Amazon …).

I.1.8.2. Standards et API ouvertes :

Pas de Web, sans respect des standards d'usage. C'est encore plus vrai avec le Web 2.0.Et ces standards d'usage sont disponibles :

- ✓ XHTML et le micro formats pour gérer des contenus structurés.
- ✓ JavaScript et Ajax pour gérer l'évènementiel des interfaces.
- ✓ RSS, Atom, XML, JSON comme format d'échanges.
- ✓ XML-RPC, SOAP et surtout REST comme protocole d'échanges [7].

I.1.8.3.Première catégorie d'application Web 2.0 : L'environnement de productivité personnelle :

Il s'agit ici de doter l'internaute de nouveaux moyens de consommer l'information, de possibilités de contrôle pour gérer sa réputation numérique ainsi que de facultés à participer activement au réseau.

I.1.8.3.1. Abonnement RSS :

La première et la plus indispensable des pratiques. Les utilisateurs peuvent personnaliser leur mode de consommation de l'information tout en gérant leur classement automatique (chaque info est « rangée » dans le flux d'abonnement). A noter que RSS est de plus en plus utilisé au sein des applications pour remonter des informations (notification d'une nouvelle commande, envoi de rapport synthétique, etc.). Aujourd'hui, il existe de nombreux lecteurs de fils RSS *standalone* ou en ligne (Outlook 2007, Safari, Google News Reader, etc.), le seul frein à l'adoption reste la formation des utilisateurs [7].

I.1.8.3.2. Les réseaux sociaux :

La participation à ces réseaux permet aux individus d'être mis en relation par l'intermédiaire des personnes qu'elles ont acceptées dans leur premier cercle. De multiples réseaux existent, aussi bien dans le domaine professionnel (LinkedIn, 6nergies, etc.) que dans le domaine privé (SuperLov (ex Friendset), Last.fm, etc.) [7].

I.1.8.3.3. Les blogs :

Cette pratique va permettre de satisfaire le besoin d'estime que ressentent certains internautes. Ils peuvent passer du statut de lecteur à celui de rédacteur. Hormis les blogs d'expertise (positionnement que nous avons retenu pour celui de Clever Age), l'intérêt reste néanmoins très limité à l'échelle de l'entreprise, voire même dangereux (validation des écrits, perte d'identité commune, dispersion des contenus …) [7].

I.1.8.3.4. La réputation numérique :

Partout nous laissons des « traces » nous concernant sur Internet (notation eBay, réactions dans des forums de discussions …). Il devient de plus en plus utile de pouvoir contrôler sa réputation sur le Web. Des services tels qu'énergies nous permettent de garantir un certain contrôle sur nos identité numérique [7].

I.1.8.4. Deuxième catégorie d'applications Web 2.0 : La constitution de réseaux d'intérêt :

Véritable creuset d'intelligence collective, ces applications reposent sur quatre principes :

1. **Abolition des limites** : Le Web permet de s'affranchir des limites physiques : c'est le cas des stocks pour les sites marchands, c'est le cas des tags pour classifier de l'information dans les bibliothèques... Il faut donc saisir cette opportunité et abandonner les vieux réflexes. C'est en exploitant cette liberté que l'on colle au mieux aux attentes des utilisateurs.

2. **Principe d'élévation** : Dans un réseau d'intérêt, plus un élément est pointé, plus il est mis en avant. C'est notamment ce qui permet de faire surgir des informations de l'infinité des possibilités. Ainsi les *tags* les plus fréquemment « collés » sont mis en avant dans les nuages de *tags* et apparaissent de plus en plus gros et de plus en plus gras. De la même manière, plus un visiteur contribue, plus il est valorisé par des mises en avant et/ou plus il a accès à de nouvelles fonctionnalités.

3. **Micro-contribution et autorégulation** : Dans un réseau d'intérêt, il faut prendre conscience que des capacités individuelles sont agrégées pour créer une capacité collective. Il est important que l'ensemble des participants adhère à une véritable culture de partage et de confiance. Ceci suppose que l'on laisse aussi libre que possible la publication d'informations sans validation, qu'on permet aux membres du groupe de juger leurs pairs et que l'on agira de manière directive qu'en dernier recours.

4. **Utilisation de mécanisme de ciblage apprenant** : Le principe est que plus un utilisateur utilise un service, plus ce dernier va apprendre de lui et va lui suggérer des informations, des produits ou services pertinents. En clair, c'est le retour du Marketing One to One [7].

I.1.8.5. Troisième catégorie d'applications Web 2.0 : Les plateformes applicatives :

Le Web est arrivé à maturité en tant que socle des applications à destination du grand public mais également des entreprises. On retrouve dans le grand bazar du Web 2.0 de véritables plateformes applicatives clé en main. De nombreux besoins génériques sont couverts par des services en ligne : solution groupware complète (Google Apps for Your Domain…), solution de collaboration (ning.com,Yahoo ! Groups…), suivi de projets (BaseCamp…), solution de gestion des temps [7].

I.1.9. Les Lacunes du web 2.0 :

Après cette catégorisation de la nébuleuse Web 2.0, quelques réflexions sur les lacunes du Web 2.0 :

• **Trop de contenus** : On laisse la possibilité à tous de publier et on s'en donne à cœur joie. Si la production du contenu était un des enjeux du Web 1.0 ; la cascade des contenus inutiles noie les contenus intéressants.

• **Trop d'identité numérique** : Il n'existe pas de standard de fait pour gérer les facettes de nos identités numériques (authentification, réputation,..). Certains projets tels que OpenID, Gravatar... sont des pistes mais leur adoption est bien trop limitée pour résoudre cette problématique.

• **Trop d'applications** : Le principe de la plupart des applications Web 2.0 est que plus les utilisateurs investissent du temps plus le bénéfice retiré est important (contribution dans les Wiki, enrichissement de son réseau dans un réseau social...). Or, l'émergence permanente de nouveaux services concurrençant les précédents incite les utilisateurs à la plus grande « frivolité » [7].

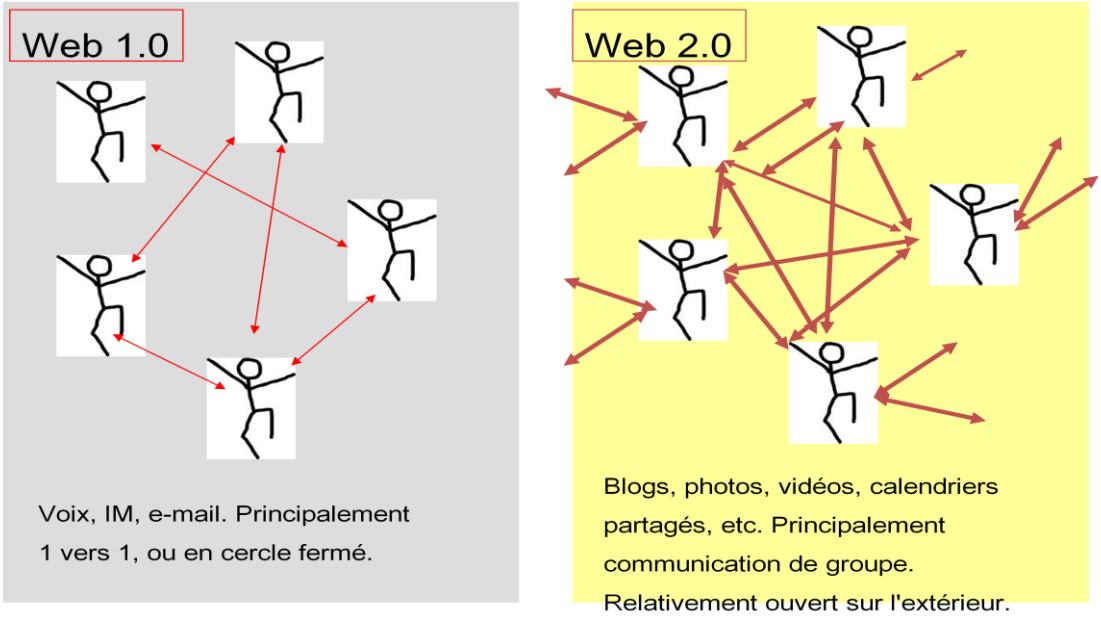


Figure. I.2 : Comparaison entre la communication dans le web 1.0 et le web 2.0 [8]

I.2.Web sémantique :

I.2.1. Introduction :

Le Web sémantique est une extension du Web qui facilite l'automatisation du traitement des connaissances disponibles. C'est une extension du Web classique. Les connaissances ne sont pas représentées dans une langue naturelle mais formalisées à l'aide de langages pouvant être interprétés par des machines [9].

I.2.2. Définition :

Le Web sémantique désigne un ensemble de technologies visant à rendre le contenu des ressources du World Wide Web accessible et utilisable par les programmes et agents logiciels, grâce à un système de métadonnées formelles, utilisant notamment la famille de langages développés par le W3C [9].

I.2.3. Principe général :

Le Web sémantique est entièrement fondé sur le Web et ne remet pas en cause ce dernier. Le Web sémantique s'appuie donc sur la fonction primaire du Web « classique » : un moyen de publier et consulter des documents. Mais les documents traités par le Web sémantique contiennent non pas des textes en langage naturel (français, espagnol, chinois, etc.) mais des informations formalisées pour être traitées automatiquement. Ces documents sont générés, traités, échangés par des logiciels. Ces logiciels permettent souvent, sans connaissance informatique, de :

- Générer des **données** sémantiques à partir de la saisie d'information par les utilisateurs ;
- Agréger des données sémantiques afin d'être publiées ou traitées ;
- Publier des données sémantiques avec une mise en forme personnalisée ou spécialisée ;
- Echanger automatiquement des données en fonction de leurs relations sémantiques ;
- Générer des données sémantiques automatiquement, sans saisie humaine, à partir de règles d'inférences [9].

I.2.4. Les langages pour le web sémantique :

Les travaux visant la réalisation du Web sémantique se situent à des niveaux de complexité très différents. Les plus simples utilisent des jeux plus ou moins réduits de métadonnées dans un contexte de recherche d'information ou pour adapter la présentation des informations aux utilisateurs. Dans ce cas, des langages de représentation simples sont suffisants. Dans les travaux les plus complexes mettant en œuvre des architectures sophistiquées, pour permettre par exemple l'exploitation de ressources hétérogènes, des langages plus expressifs et plus formels issus des travaux en représentation et en ingénierie des connaissances, sont nécessaires.

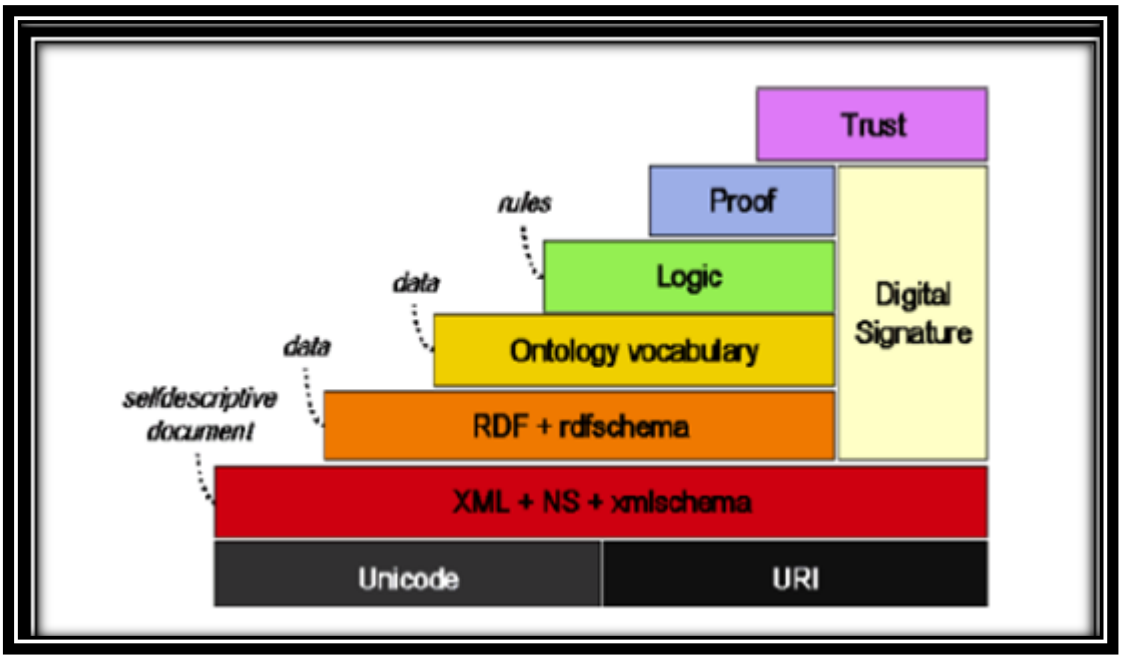


Figure. I. 3 : Les couches du Web Sémantique [10]

I.2.4.1. W3C :

La proposition du W3C s'appuie au départ sur une pyramide de langages dont seulement les couches basses sont aujourd'hui relativement stabilisées. La figure. I.3 montre une des versions de l'organisation en couches proposée par le W3C. Deux types de bénéfices peuvent être attendus de cette organisation. Elle permet une approche graduelle dans les processus de standardisation et d'acceptation par les utilisateurs [10].

I.2.4.2. RDF :

RDF est un langage formel qui permet d'affirmer des relations entre des «ressources». Il sera utilisé pour annoter des documents écrits dans des langages non structurés, ou comme une interface pour des documents écrits dans des langages ayant une sémantique équivalente (des bases de données, par exemple). RDF est muni d'une syntaxe, et d'une sémantique. Aucun mécanisme d'inférence n'est cependant proposé dans la recommandation [11].

I.2.4.3. Les Topic Maps :

Les cartes topiques (Topic maps) sont un standard ISO issu de HyTime dont le but était d'annoter les documents multimédia. Issu de SGML, il s'est vu récemment attribuer une syntaxe XML (XTM). Par ailleurs, un groupe de l'ISO s'occupe de définir un langage de requêtes pour les cartes topiques (TMQL). Les cartes topiques sont bâties autour de quatre notions primitives:

1. Les « topics » que l'on peut comprendre comme des individus des langages de représentation de connaissances.
2. Les noms donnés aux topics: l'une des originalités des cartes topiques est la séparation des concepts et de leurs noms. Cela permet d'avoir plusieurs noms pour le même concept (et donc d'avoir des cartes topiques multilingues) et des noms partagés par plusieurs concepts.
3. Les occurrences sont des « proxis » d'entités externes qui peuvent ainsi être indexés par les topics (où les entités littérales lorsque celles-ci sont représentables).
4. Les portées, qui sont parfois vues comme une quatrième dimension, permettent de spécifier le contexte dans lequel une relation est valide [11].

I.2.4.4. UDDI :

Le protocole UDDI (Universal Description, Discovery and Integration) est une plateforme destinée à stocker les descriptions des services web disponibles, à la manière d'un annuaire de style «Pages Jaunes». Des recherches sur les services peuvent être effectuées à l'aide d'un système de mots-clés fournis par les organismes proposant les services. UDDI propose également un système de «Pages Blanches» (Adresses, numéros de téléphone, identifiants…) permettant d'obtenir les coordonnées de ces organismes. Un troisième service, les «Pages Vertes», permet d'obtenir des informations techniques détaillées à propos des services et permettent de décrire comment interagir avec les services en pointant par la suite vers un PIP Rosetta Net ou un « service interface » WSDL. Le vocabulaire utilisé pour les descriptions obéit à une taxonomie bien précise afin de permettre une meilleure catégorisation des services et des organismes [11].

I.2.4.5. WSDL :

WSDL est un langage basé sur XML servant à décrire les interfaces des services Web, c'est-à-dire en représentant de manière abstraite les opérations que les services peuvent réaliser, et cela indépendamment de l'implémentation qui en a été faite. Il ne comporte pas de moyen de décrire de manière plus abstraite les services (tâche plutôt dévolue à DAML-S ou à UDDI), ni de moyen de conversation et de transaction de messages (tel que SOAP ou d'autres implémentations spécifiques), mais est en général utilisé comme passerelle entre ces représentations de haut niveau et de bas niveau [11].

I.2.4.6. DAML-S :

DAML-S est un langage de description de services basé sur XML, et utilisant le modèle des logiques de descriptions. Son intérêt est qu'il est un langage de haut niveau pour la description et l'invocation des services web dans lequel la sémantique est incluse, contrairement par exemple à UDDI. DAML-S est composé de trois parties principales :

- ***Service Profile***, qui permet la description, la promotion et la découverte des services, en décrivant non seulement les services fournis, mais également des pré conditions à la fourniture de ce service, comme «!avoir une carte bleue valide!» ou «!être membre d'un des pays de l'Union Européenne!». Les recherches sur les services peuvent se faire en prenant n'importe quel élément de Service Profile comme critère

- ***Service Model***, qui présente le fonctionnement du service en décrivant dans le détail et de manière relativement abstraite les opérations à effectuer pour y accéder. Certains éléments du Service Model peuvent être utilisés à la manière du Service Profile afin de fournir des informations supplémentaires à un utilisateur pour qui les opérations à effectuer seraient également un critère de choix. C'est le Service Model qui va permettre une composition des services s'il y a besoin. Il permet également d'effectuer un contrôle poussé du déroulement du service.

- ***Service Grounding*** va présenter clairement et dans le détail la manière d'accéder à un service. Tout type abstrait déclaré dans le Service Model s'y verra attribué une manière non ambiguë d'échanger l'information. C'est dans cette partie que le protocole et les formats des messages entre autres sont spécifiés [11].

I.2.4.7. XL :

XL est une plateforme destinée aux services web, axée sur XML, utilisant un langage propre de haut niveau (XL), et prenant en compte les technologies du W3C (WSDL, SOAP) afin de permettre une interopérabilité des applications XL avec d'autres applications écrites dans un langage autre que XL. Tout service web est considéré comme une entité recevant des messages XML et transmettant en retour des messages XML, avec (achat d'un livre) ou sans (consultation de la météo) modification du monde. Les types de données utilisés sont ceux de XQuery, développé lui aussi par le W3C, est dont est inspiré la syntaxe de XL [11].

I.3.Web social :

I.3.1. Introduction :

Avec l'avènement du Web 2.0, l'utilisateur est au centre des préoccupations des différentes technologies composant ce nouveau modèle comme les Mashup, les environnements collaboratifs, les réseaux sociaux, etc. Le principal ingrédient rajouté est le social qui consiste à mettre en relation les utilisateurs, à leur faciliter l'interaction et à la rendre plus riche et plus productive. Le Web social devient ainsi de plus en plus la partie la plus intéressante de tout le Web, au point de défier de grands acteurs bien établis sur le Web traditionnel comme le moteur de recherche Google. Ceci constitue une énorme avancée d'un point de vue utilisateur et ouvre aussi de grandes perspectives de recherche dans un environnement qui devient de plus en plus complexe, moins

structuré et plus hostile compte tenu de la grande masse d'information généralement cachée à l'utilisateur [12].

I.3.2. Définition :

Le Web social fait référence à une vision d'Internet considéré comme une socialisation, un lieu dont l'une de ses fonctions principales est de faire interagir les utilisateurs entre eux afin d'assurer une production continuelle de contenu, et non plus uniquement la distribution de documents.

Il est considéré comme un aspect très important du Web 2.0. En particulier, il est associé à différentes systèmes sociaux tels que le réseautage social, les blogs ou les wikis [13].

I.3.3. Historique :

En 1955, le terme "Social Web" apparait sous la plume de l'auteur C. Krey dans l'essai *History and the Social Web* publié par les presses de l'université du Minnesota.

Au début des années 1990, les idées associées à ce concept ont aussi été utilisées relative aux systèmes en ligne utilisées pour supporter les interactions sociales telles que les communautés virtuelles ou les MUD (ou Multi-user Dungeon, qui sont les jeux de rôle en ligne multiutilisateurs).

En 1998, le terme "Social Web" a été utilisé dans un article de Peter Hoschka qui décrivait le passage d'une utilisation des ordinateurs et du web comme de simples outils de coopération à un usage de l'ordinateur comme un médium social : From Basic Groupware to the Social Web.

En juillet 2004 ce terme a aussi été utilisé dans un article décrivant une utilisation de XDI pouvant intervenir dans le cadre de la conception d'applications web plus sociales.

Finalement, à partir de 2005, ce concept a aussi connu un développement très important avec l'arrivée du Web 2.0, avec lequel il est très fortement lié, du fait de l'importance qui est donnée à la participation des individus [13].

I.3.4. Objectif :

Les outils du web social permettent à leurs utilisateurs d'agir les uns pour les autres comme des filtres d'information. On peut contribuer à ce processus de filtrage aussi simplement qu'en signalant son intérêt pour certaines informations plus que pour d'autres, mais également de certaines façons plus sophistiquées. L'étiquetage est l'une de ces façons qui s'est dernièrement répandue comme une traînée de poudre. Cette pratique donne naissance à un mode de classification que l'on appelle communément la folksonomie [12].

I.3.5. Les aspects techniques du Web 3.0 :

La définition précise d'une application Web 3.0 est encore très débattue. Cependant, il est généralement admis qu'une solution Web 3.0 doit montrer certaines caractéristiques :

- On ne se réfère plus uniquement à un site Web (XHTML). Ce peut être aussi une solution Web SaaS (application: XHTML + base de données relationnelles (SQL Server, Oracle, MySQL…) ou XML ;
- Mobilité, elle doit être indépendante de tout type de support (taille d'écran, sortie imprimante, etc.) ;
- Universalité, elle doit être indépendante de tout système d'exploitation, et de tout matériel (fabricant, marque, logiciel, ou de plugin) ;
- Accessibilité, strictement en conformité avec le W3C, ce qui permet de rendre d'autres logiciels accessibles à l'aide de Microformat et ouverts aux bases de données diverses [14].

I.4.Conclusion :

La plupart des utilisateurs du Web connaissent les terminologies communes: Web 1.0, Web 2.0 et Web 3.0. Alors quels sont les usages de ces terminologies? Quelles sont les différences entre eux?

Les informations disponibles sous le Web2.0 sont très différentes de celle disponible sous le Web1.0 passé. Le Web 2.0 a commencé en 2002 avec de nouvelles idées à échanger, ainsi que de partager le contenu comme les Wiki, blogs, Widgets, et le marquage, etc. Dans le Web 1.0 c'est seulement pour lire. Mais dans Web 2.0 vous pouvez vous exprimer par écrit. Le premier était que pour les personnes morales. Et le second est sur vous et sur vos communautés.

En Web 2.0 vous pouvez non seulement interagir avec le site et le webmaster, mais également communiquer avec les autres qui ont l'accès à ce site.
Le Web 1.0 a dépendu de la publicité, le Web 2.0 a été popularisé par le bouche à oreille.

Dans le Web 1.0 il n'y avait rien à échanger. Tout allait dans un sens. Grâce à l'émergence du Web 2.0, vous pouvez échanger votre opinion avec les autres et facilement converser avec eux.

Le Web 3.0 pourrait être définie comme le Web sémantique, la personnalisation comme iGoogle, Mon Yahoo, etc.

Le Web sémantique est une extension plus développés de la WWW à l'aide de cette technologie du contenu Web peut être transmis non seulement sous forme de langage naturel, mais aussi être lisible par un agent logiciel qui les laisser pour trouver, partager et rassembler des informations plus facilement [15].

Chapitre II:
Fouille du texte

II.1.Introduction :

La fouille de textes ou l'extraction de connaissances dans les textes est une spécialisation de la fouille de données et fait partie du domaine de l'intelligence artificielle. Cette technique est souvent désignée sous l'anglicisme *text mining* [16].

La fouille de données est une extraction d'informations préalablement inconnue ou partiellement connue a partir des données [30].

II.2.Définition :

C'est un ensemble de traitements informatiques consistant à extraire des connaissances selon un critère de nouveauté ou de similarité dans des textes produits par des humains pour des humains. Dans la pratique, cela revient à mettre en algorithmes un modèle simplifié des théories linguistiques dans des systèmes informatiques d'apprentissage et de statistiques.

Les disciplines impliquées sont donc la linguistique calculatoire, l'ingénierie du langage, l'apprentissage artificiel, les statistiques et bien sûr l'informatique [16].

La fouille de textes est constituée d'un ensemble de méthodologies qui modifient le texte afin d'en préciser le sens dans le but d'améliorer la solution d'un problème spécifique. Classiquement, on essaie de transformer le texte en tableau de données afin d'appliquer des techniques de fouille de données [17].

II.3.Objectif de fouille de texte :

L'objectif du Text Mining (fouille de texte) est de trouver des informations importantes et des relations dans de grands volumes de textes comme il n'est pas rare d'en rencontrer dans des bases de données, dans des manuels techniques d'avions, dans le savoir global d'une entreprise, dans des demandes quotidiennes de clients ou même dans l'ensemble du 3W [17].

II.4. Etape de la fouille :

On peut distinguer deux étapes principales dans les traitements mis en place par la fouille de textes :

II.4.1. Analyse :

Consiste à reconnaître les mots, les phrases, leurs rôles grammaticaux, leurs relations et leur sens. Cette première étape est commune à tous les traitements. Une analyse sans interprétation n'a que peu d'intérêt et les deux sont dépendantes. C'est donc le rôle de la seconde étape d'interpréter cette analyse [16].

II.4.2. l'interprétation de l'analyse :

Permet de *sélectionner* un texte parmi d'autres. Des exemples d'applications sont la classification de courriers en spam, c'est-à-dire les courriers non sollicités, ou non spam, l'application de requêtes dans un moteur de recherche de documents ou le résumé de texte qui sélectionne les phrases représentatives d'un texte voire les reformule [16].

Remarque :

Le critère de sélection peut être d'au moins deux types : la nouveauté et la similarité. Celui de la nouveauté d'une connaissance consiste à découvrir des relations, notamment des implications qui n'étaient pas explicites car indirectes ou entre deux éléments éloignés dans le texte. Celui de la similarité ou contradiction par rapport à un autre texte ou encore la réponse à une question spécifique consiste à découvrir des textes qui correspondent le plus à un ensemble de descripteurs dans la requête initiale. Les descripteurs sont par exemple les noms et verbes les plus fréquents d'un texte [16].

II.5.La différence fondamentale entre la Recherche d'Informations (RI) et l'Extraction d'Information (EI)) :

Avant de présenter cette section, il faut bien montrer qu'il y'a une grande différence entre la recherche d'informations « RI » et l'extraction d'informations.

La tâche de recherche d'informations (RI) : étant donné une requête d'un utilisateur et une collection de documents, identifier un sous ensemble de ceux-ci répondant aux critères de la requête.

Ils sont différents dans :

✓ **Les objectifs :**

- **RI sélectionne** des documents potentiellement pertinents à partir d'un ensemble
- **EI extrait** l'information pertinente des documents.

✓ **Les résultats :**

- **RI** propose les **documents** potentiellement pertinents à l'utilisateur
- **EI** propose des **faits** pertinents à l'utilisateur ou à un programme informatique [30].

II.6.Applications :

II.6.1. Recherche d'information :

Les moteurs de recherche tels Google, Exalead ou Yahoo! sont des applications très connues de fouille de textes sur de grandes masses de données. Notons toutefois que les moteurs de recherche ne se basent pas uniquement sur le texte pour l'indexer, mais

également sur la façon dont les pages sont mises en valeurs les unes par rapport aux autres. L'algorithme utilisé par Google est PageRank [16].

II.6.2. Applications biomédicales :

Un exemple d'application biomédicale de fouille de textes est PubGene, qui combine la fouille de textes et la visualisation des résultats sous forme de réseaux graphiques. Un autre exemple d'utilisation d'ontologies avec la fouille de textes est GoPubMed.org [16].

II.6.3 Filtrage des communications :

Beaucoup de gestionnaires de courriers électroniques sont maintenant livrés avec un filtre anti-spam.

Il existe aussi des logiciels anti-spam qui s'interfacent entre le serveur de courrier et votre gestionnaire de courrier [16].

II.6.4. Applications de sécurité :

Le système mondial d'interception des communications privées et publiques Echelon est un exemple d'utilisation militaire et économique de la fouille de textes.

En 2007, la division de lutte anti-criminelle d'Europol a acquis un système d'analyse afin de lutter plus efficacement contre le crime organisé. Ce système intègre parmi les technologies les plus avancées dans le domaine de la fouille et d'analyse de textes. Grâce à ce projet Europol a accompli des progrès très significatifs dans la poursuite de ces objectifs [16].

II.6.5. Gestion des connaissances :

Les méthodes d'Intelligence économique ont pour objectif général d'apporter des informations à l'organisation [16].

II.6.6. Analyse du sentiment :

Une utilisation particulière de traitement de l'information non structurée peut déboucher sur une analyse du sentiment. Par exemple, ces documents montrent-ils que mon produit sera bien vu par les utilisateurs ? [16].

Dans notre travail en intéresse à la recherche d'information, le filtrage de communication et même la gestion des connaissances.

II.7. Disciplines connexes :

La fouille de textes se distingue du traitement automatique du langage naturel par son approche générale, massive, pratique et algorithmique de par sa filiation avec la fouille de données. Son approche est moins linguistique. De plus, la fouille de textes ne s'intéresse pas au langage oral comme le fait la reconnaissance vocale.

La fouille de textes recoupe la recherche d'information pour la partie requête sur un moteur de recherche de documents. Par contre, la recherche d'information s'intéresse a priori plus aux types de requêtes possibles et aux indexations associées qu'à l'interprétation des textes.

Et pour information, car on s'éloigne alors du domaine de la fouille de textes, l'interprétation de l'analyse peut aussi *générer* un nouveau texte. Des exemples d'applications sont la correction des fautes d'orthographe, la traduction, le dialogue homme-machine ou l'imitation d'un style d'écriture [16].

II.8.Processus globale de fouille de textes :

La figure suivante présente le processus globale de la fouille de textes :

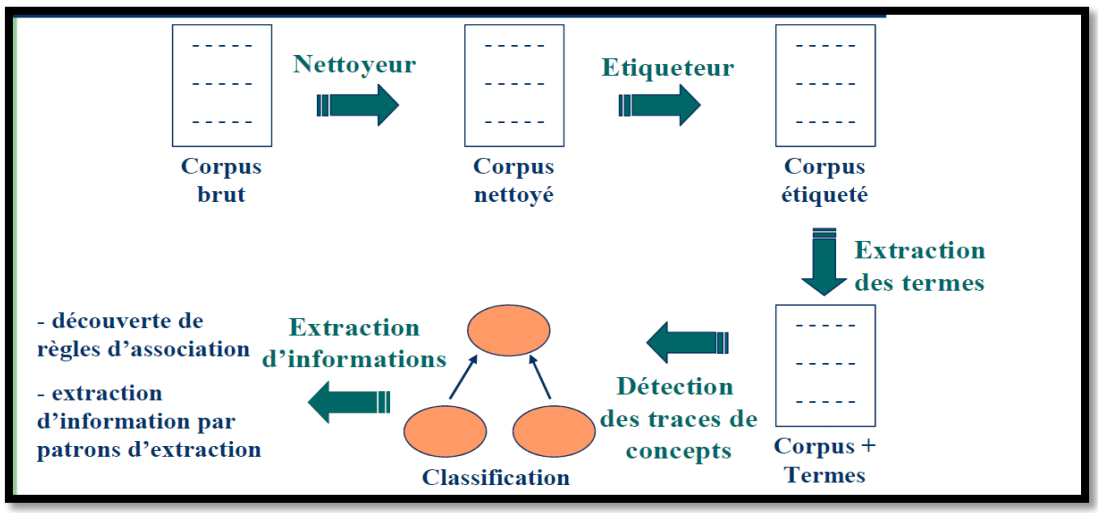


Figure. II.4 : Processus de la fouille de texte [18]

II.8.1. Etape 1 : Le nettoyage :

Exemples de corpus spécialisés :

● Corpus de 100 introductions d'articles en anglais écrits par des auteurs anglophones sur le domaine de la « fouille de données » (*369 Ko*).

● Corpus de plus de 6000 résumés d'articles en anglais sur la biologie Moléculaire (*9424 Ko*).

● Corpus en français de plus de 1000 Curriculum Vitæ (VediorBis, *2470 Ko*).

● Corpus en français relatif aux Ressources Humaines (PerfomanSe, *3784 Ko*)

Le but de nettoyage :

- Enlever les noms, prénoms, coordonnées, etc. (pour les articles et les CVs).
- Uniformiser les références.
- Généraliser certains noms.

II.8.2.Etape 2 : Etiquetage :

La figure II.5 illustre l'étiquetage :

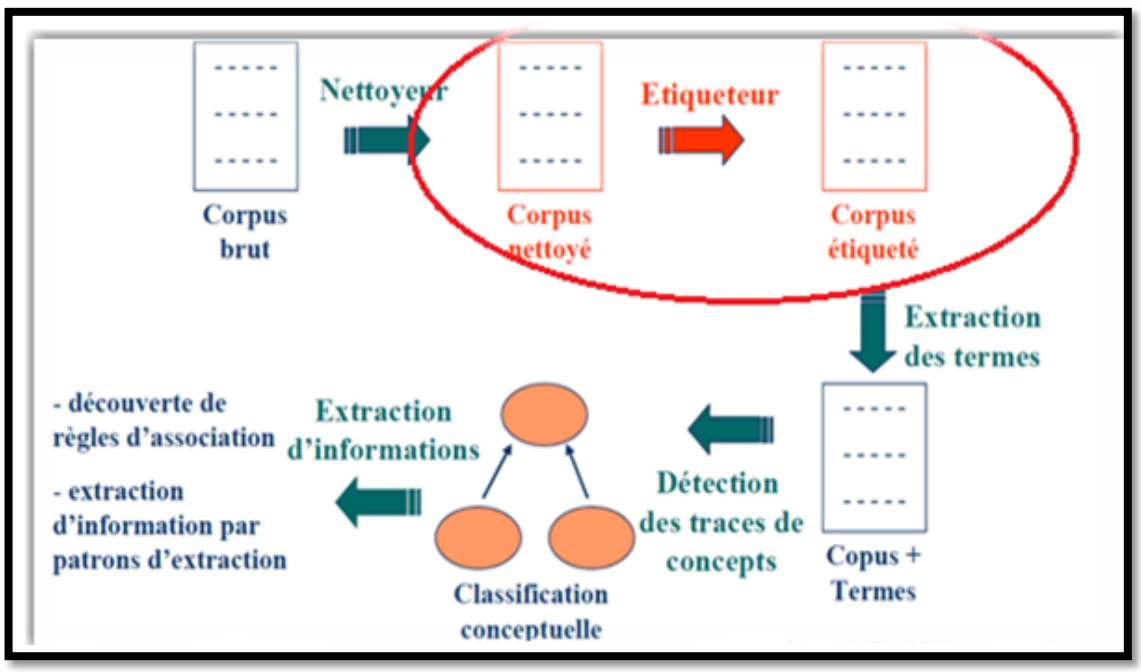


Figure. II.5 : Etape 2 de processus de fouille de texte **[18]**

La figure II.6 montre un type spécifique des étiqueteurs qu'il s'appelle : étiqueteur de Brill.

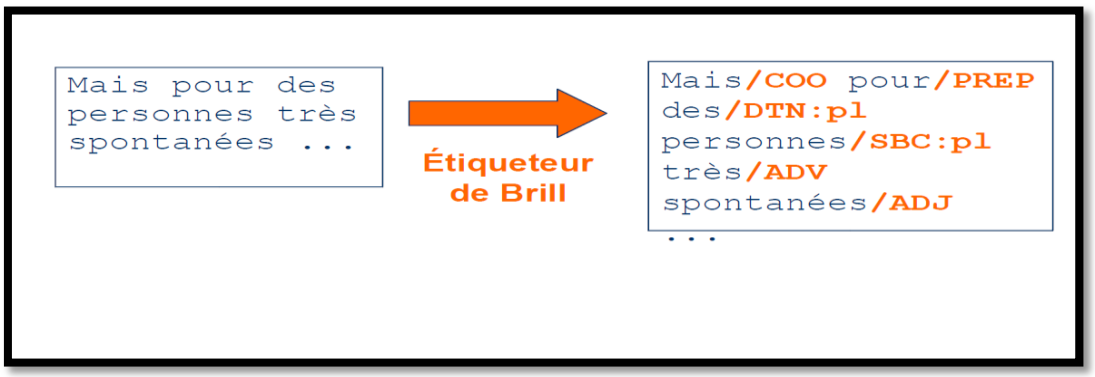


Figure .II.6 : Etiqueteur de Brill **[18]**

II.8.3.Etape 3 : Extraction de termes :

Cette figure en dessous présente la phase de l'extraction des termes quelle est divisé en deux étapes :

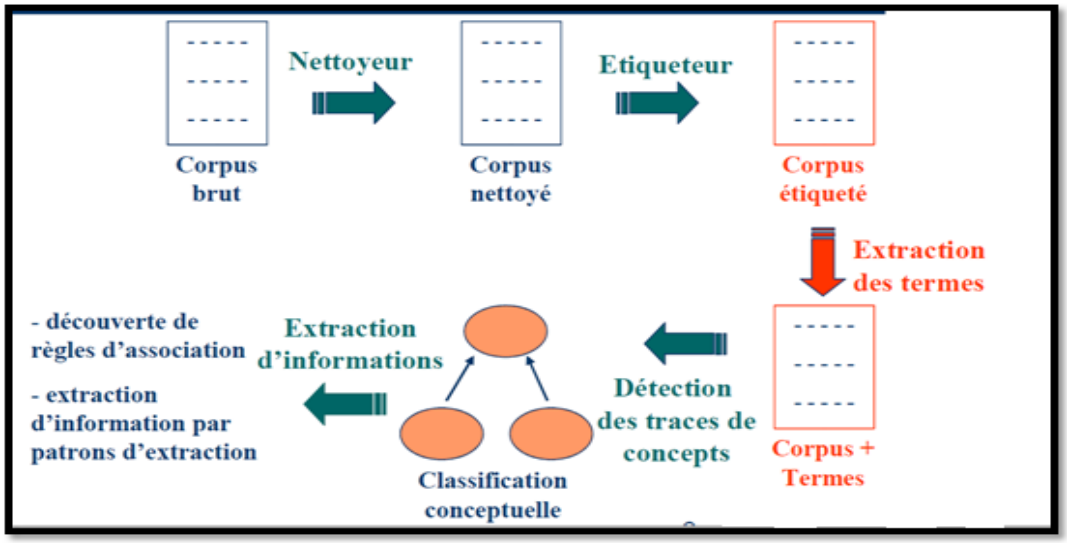


Figure. II. 7 : Etape 3 de processus de fouille de texte (extraction de termes) **[18]**

Les Etapes d'Extraction de Termes :

- ***1er étape :***

La première étape consiste à extraire tous les types de collections possibles.

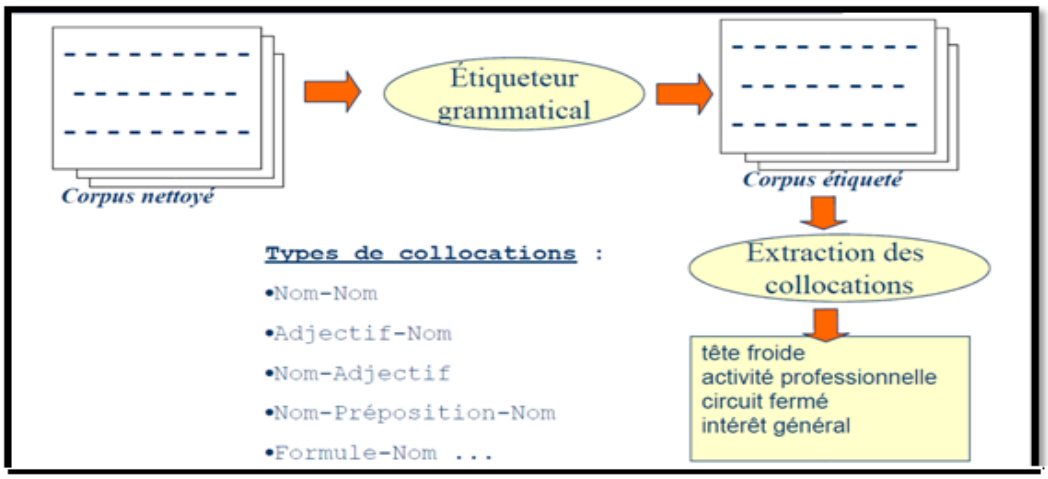


Figure.II. 8 :L'extraction des collocations **[18]**

- ***2éme étape :***

La deuxième étape sert à sélectionner les meilleures collections parmi les types collections possible retenu dans la première étape d'extraction des termes.

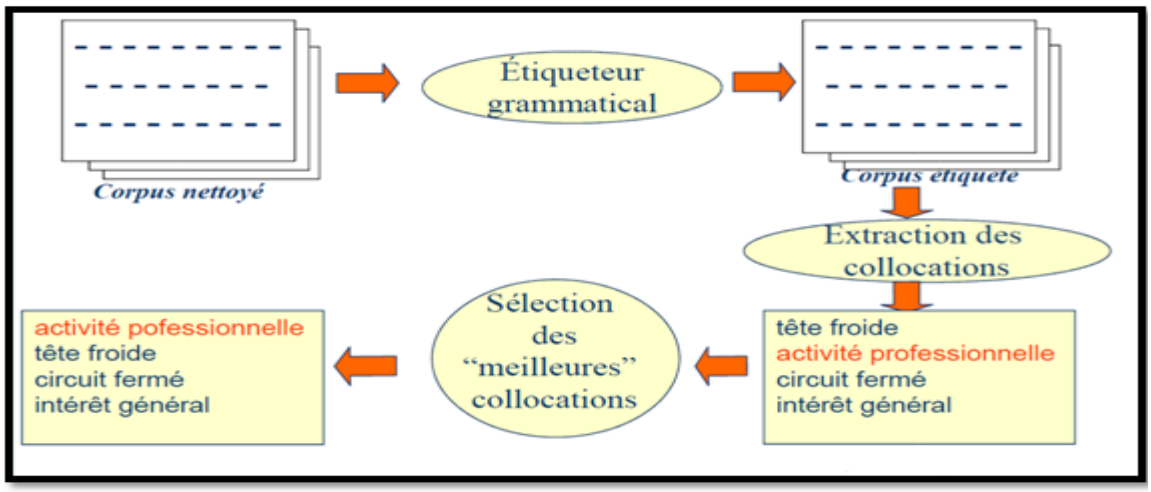


Figure. II.9 : Sélection de meilleures collocations [18]

II.8.4.Etape 4 : détection des traces de concepts :

Avant ça il faut tous d'abord définit la classification conceptuelle :

- Voila un exemple d'une classification des moyens de transportes :

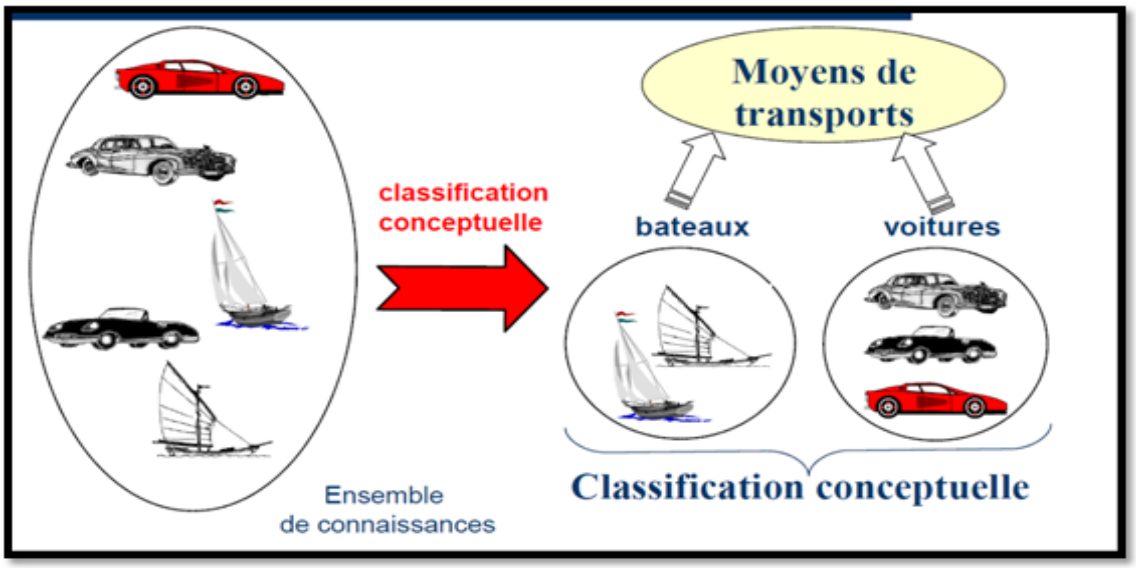


Figure.II.10 : La classification des moyens de transports **[18]**

La figure suivante présente la détection des traces de concepts pour les faires classifiés d'après une classification conceptuelle.

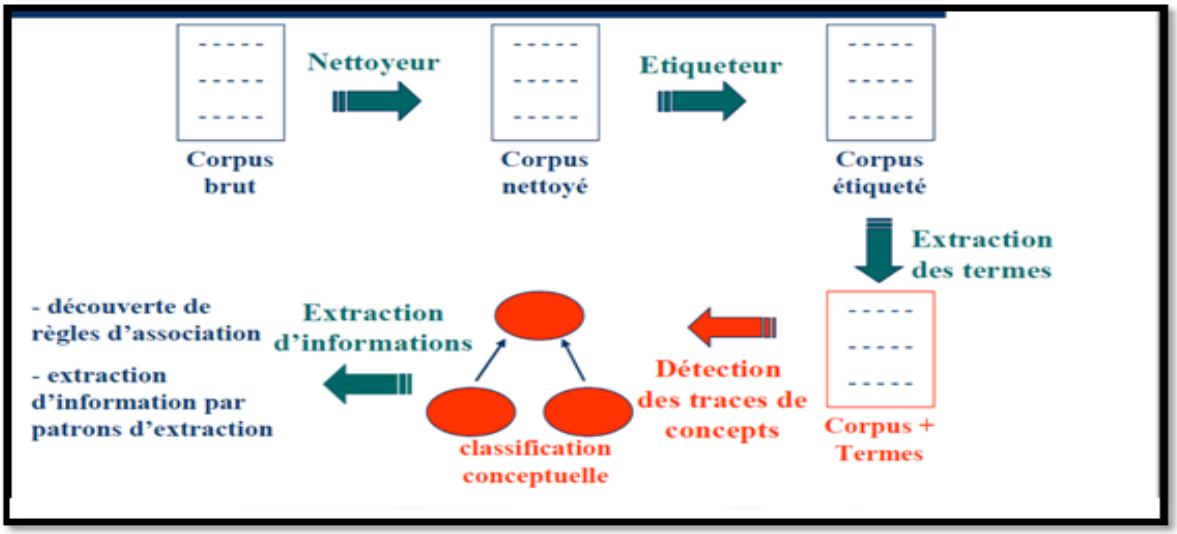


Figure.II.11 : Etape 4 de processus de fouille de texte **[18]**

La figure II.12 montre les détailles de la phase de détection des traces.

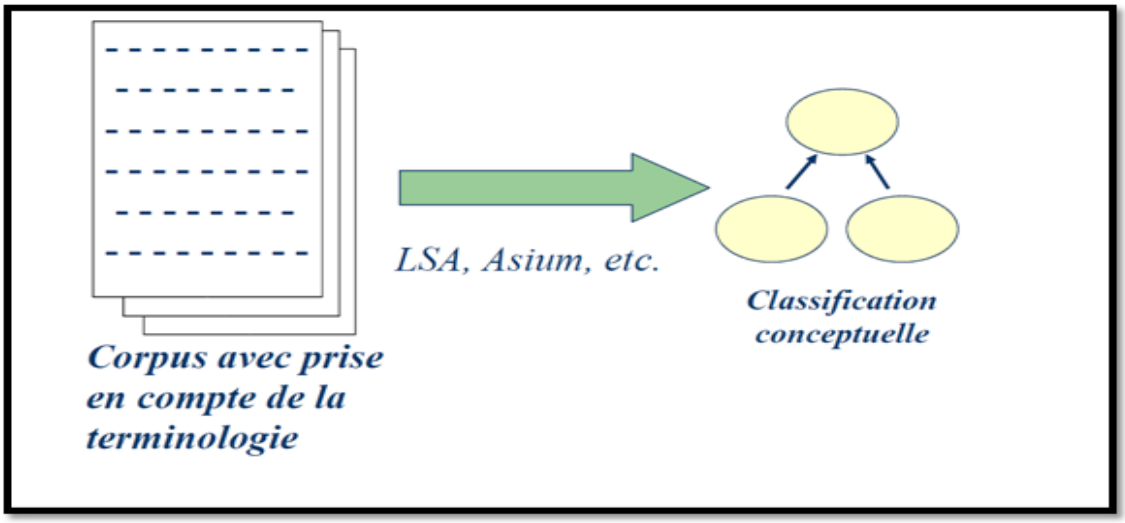


Figure.II. 12 : La détection des traces concepts **[18]**

II.8.5.Etape 5 : Extraction d'informations :

Cette figure présente la dernière phase de processus de la fouille de textes c.à.d. l'extraction d'informations qui peut être effectué soit par un patron d'extraction ou par les règles d'associations.

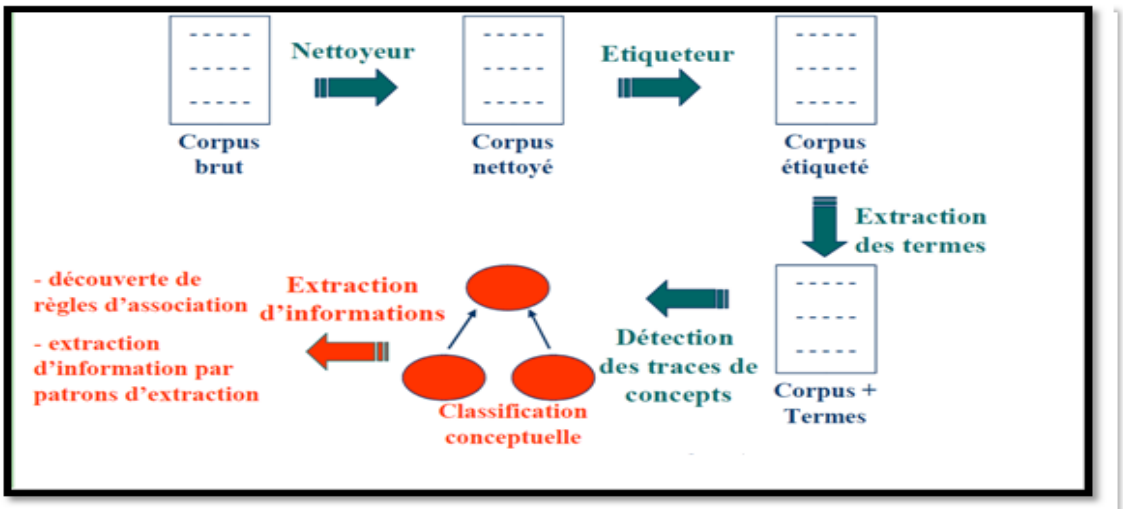


Figure.II.13 : Etape 5 de processus de fouille de texte (extraction d'information) [18]

II.9.Conclusion:

La fouille de textes est une thématique scientifique pluridisciplinaire mettant en relation la recherche universitaire, la recherche en milieu industriel, les préoccupations en sciences fondamentales et les enjeux politico-économiques. Elle fédère des thématiques issues des sciences de l'information, de la linguistique, de la statistique et des méthodes d'apprentissage (I.A). Dans un contexte où, grâce au Web, l'accès à l'information écrite est divers, abondant, multilingue, la fouille de textes est par excellence un domaine pour lequel il était nécessaire de faire un état des lieux

Chapitre III:
Agrégateurs de réseaux sociaux

III.1. Introduction :

Si vous être un utilisateur des réseaux sociaux comme twitter, facebook, linkedin, foursquare et autres, vous êtes souvent obliger de lancer plusieurs pages Web pour consulter vos messages ou pour suivre vos amis. C'est assez contraignant et pas toujours très simple à gérer. Pour éviter tout cela, il existe des agrégateurs pour vos réseaux sociaux. Ces solutions permettent de centraliser tous vos comptes et de n'avoir qu'une seule page web à lancer pour suivre vos amis [20].

III.2. Définition :

On appelle agrégateur un outil capable de recueillir et de synthétiser en temps réel des données et des contenus dispersés sur plusieurs sites. Concrètement, un agrégateur permet d'avoir une vision centralisée de ce qui se passe simultanément sur l'ensemble des réseaux dont on fait partie en permettant aussi de mieux gérer son identité numérique. Un agrégateur permet ainsi de gagner du temps en interagissant le cas échéant plus rapidement avec ses contacts.

Un agrégateur est un site qui va reprendre du contenu en provenance de blogs ou sites multiples. L'objectif de ce type de plateforme est de proposer au visiteur un large choix de sources d'information, afin de lui permettre de trouver celles qui l'intéressent, sur des thématiques précises. Car les agrégateurs apportent au lecteur une grille de lecture par le biais de différents classements : thématiques, popularité, etc...

En fréquentant ces sites, vous serez ainsi en mesure de découvrir quels sont les contenus qui ont le plus intéressé les lecteurs, ou ceux qui suscitent le plus de réaction.

Un agrégateur de réseaux sociaux permet à son utilisateur de compiler sur une seule interface tous ses réseaux sociaux. Cela évite principalement d'avoir à ouvrir plusieurs sites Web et à se connecter plusieurs fois afin d'être au courant de ce qui se passe dans tel ou tel réseau. En un seul coup d'œil, vous pouvez faire le tour de vos contacts [21].

III.3. Présentation générale des flux RSS :

Le standard RSS représente un moyen d'être tenu informé des nouveaux contenus d'un site web, sans avoir à le consulter. Le format «RSS» permet de décrire de façon synthétique le contenu d'un site web, dans un fichier au format XML, afin de permettre son exploitation par des tiers. Le fichier RSS, appelé également flux RSS, canal RSS ou fil RSS, contenant les informations à diffuser, est maintenu à jour afin de constamment contenir les dernières informations à publier. Pour pouvoir exploiter un fil RSS, un utilisateur doit disposer d'un outil spécifique, appelé «lecteur RSS» ou encore «agrégateur RSS», afin d'exploiter les fils RSS. Ainsi, il est possible de consulter en un seul endroit les dernières actualités de dizaines, et parfois de centaines de sites web, sans avoir à les visiter et sans avoir à communiquer d'informations personnelles [22].

III.4. Principe :

L'intérêt d'un agrégateur est triple. En une seule application :

- Il prévient de la mise à jour d'un site Web ou des actualités qu'il publie (par notification sonore, visuelle, etc.)
- Il importe le nouveau contenu en question.
- Il le fait pour un ensemble de sites.

C'est une sorte de « *facteur* » qui va chercher le courrier à l'extérieur, puis le dépose chez l'utilisateur, dispensant ce dernier d'aller régulièrement aux nouvelles en visitant de nombreux sites internet. Il fonctionne un peu comme une messagerie électronique

(quasiment en temps réel) mais (contrairement à un client de messagerie), l'utilisateur d'un agrégateur est souvent limité à la lecture passive des messages reçus (le « *fil* » de syndication). Il ne peut pas « répondre » aux éléments reçus. Il existe quelques exceptions dans le cas de billets blogs, où certains agrégateurs permettent de poster des commentaires.

Un agrégateur ne peut traiter qu'une information spécialement structurée, par une technologie particulière. Les sources de contenu (des sites web en général) offrent l'adresse d'un fil de syndication mis à jour plus ou moins régulièrement. Cette première phase, dite syndication de contenu structure les données pour l'agrégateur. L'agrégation consiste à s'abonner à un ou plusieurs de ces fils de syndication. L'agrégateur détecte leurs mises à jour et averti aussitôt l'utilisateur, sans qu'il ait à visiter périodiquement les sites internet diffusant les fils de syndication auxquels il s'est abonné. Chaque fil est associé à un dossier dans l'agrégateur, dossier qui contient les différentes entrées du fil le plus souvent par ordre chronologique inverse (les plus récentes entrées en premier). La détection de nouveaux éléments dans un fil est périodique, ou réalisée à la demande de l'utilisateur qui peut quand il le souhaite mettre à jour tout ou partie de ses abonnements [24].

L'objectif d'un agrégateur est de permettre l'agrégation de plusieurs sources de contenus internet en une seule application. Le suivi du contenu est réalisé quasiment en temps réel. Proche dans son fonctionnement de la messagerie électronique, l'agrégateur est le plus souvent un outil limité à la lecture des messages reçus.

En général, un agrégateur permet de visualiser une liste des fils enregistrés, classés alphabétiquement ou par thématique. Pour chaque fil, les *n*-derniers éléments sont listés (*n* choisi par l'utilisateur ou fixé). Pour chaque élément (billet, article...) peut être affiché un résumé ou son contenu complet. De ce fait, l'utilisateur peut être amené à quitter son agrégateur pour lire le contenu sur le site d'où il a été tiré, ou bien en faire l'entière lecture dans son logiciel [24].

III.5.Usages :

Les fils de syndication sont très utilisés sur les blogs : chaque nouveau billet posté est ainsi transmis en quasi-temps réel aux personnes abonnées au fil du carnet, qui peuvent le lire directement dans leur agrégateur. Ce mode de suivi commence à être adopté en

masse par les sites d'actualités, comme les quotidiens en ligne, dont le contenu renouvelé arbitrairement ou par cycles peut-être regroupé en thématiques par l'utilisateur. La plupart des agrégateurs permettent en effet de faciliter le suivi de ces fils en les catégorisant en dossiers et sous-dossiers.

Agrégation et syndication sont les deux facettes d'une même idée, qui veut proposer à l'utilisateur une décentralisation du contenu : créé en des points isolés d'internet, il doit pouvoir être transmis à travers les mailles du réseau de façon simple, et il doit également pouvoir être regroupé chez l'utilisateur et le lecteur, en des thématiques arbitraires, sans perdre sa cohérence. L'agrégateur essaye de faciliter l'organisation du contenu, en plus d'être un outil de suivi temporel.

III.6. Interface d'accès aux services d'agrégations :

Le stockage des fils se fait de deux façons :

1. Par des logiciels fonctionnant en local a installer.

2. Par inscription sur des sites de gestions et de partages de fils. L'agrégateur est alors une application Web, non un logiciel classique. Il évite a avoir a utiliser un logiciel de création de feed spécifique a chaque système d'exploitation (Feed Editor , RSS wizard , Feed Mix sous Windows) qui nécessite que le fichier XML soit déposer en ligne par le protocole FTP pour être accessible aux internautes par le protocole HTTP. Ils s'inscrivent dans le concept de travail a distance initié par le Web depuis les systèmes de gestion de contenu utilisant fortement le concept de clients légers en s'appuyant sur la technologie Ajax [24].

III.7. Types d'agrégateurs :

III.7.1. Agrégation en ligne :

Ces sites permettent de répondre aux problèmes de la gestion des signets qui par défaut sont privés et non collaboratifs.

Les sites suivants permettent de suivre, soit des flux RSS que vous aurez vous même choisis, soit toute l'actualité au travers d'un moteur de recherche dédié à l'actualité [24].

- Bloglines
- del.icio.us

- Netvibes
- NewsRSS
- Floobby

Parmi leurs avantages on peut citer :

- Accessible en ligne depuis n'importe quel poste
- Espace de stockage quasiment infini.
- Une surveillance continue des sources même hors ligne.

Mais, il existe des inconvénients comme :

- Moins de fonctionnalités avancées.
- Pas d'accès hors connexion.
- Visualisation des résultats moins efficace [34].

III.7.2. Agrégation en local (lecture) :

Ces logiciels permettent de s'enregistrer à certain format de flux primaire (atom ou RSS) en opposition au flux géré sur un site accessible par http (en ligne) [24].

Exemple :

- Ziepod
- Itunes
- Thunderbird
- FeedBurner
- AlertInfo

Parmi leurs avantages on peut citer :

- Permet de sauvegarder ses flux hors connexion,
- Permet d'avoir une visualisation plus globale des résultats.
- Plus de fonctions avancées disponibles.

Mais, il existe des inconvénients comme :

- Pas de consultation à distance possible,

- Espace de stockage limité sur un disque dur,
- Gourmant en mémoire source de l'ordinateur [34].

III.8. Quelques agrégateurs :

III.8.1. FriendFeed:

III.8.1.1. Historique :

Créé par des anciens de chez Google, FriendFeed est un service permettant de connecter ses différents profils sociaux (Facebook, Twitter, Gtalk…) sur une seule plateforme. On peut y voir une forme d'agrégateur d'amis .Vous pouvez ainsi centraliser les « lifestreams » de vos contacts sur les différents réseaux sociaux, et vous pouvez poster sur FaceBook, Twitter et autres à partir de FriendFeed. Le service sera donc très utile aux Community Managers. Le développement de FriendFeed est très rapide, l'effet viral semble être atteint [25].

III.8.1.2. L'objectif :

Leur objectif est donc de rendre le contenu sur Internet le plus utile possible et de vous le faire découvrir via vos diverses connexions sociales. Vous allez donc pouvoir choisir qui vous voulez suivre et savoir quels sont les différents services que celui-ci utilise [26].

Il permet de créer des messages avec éventuellement des ajouts de photos et de fichiers, de commenter les flux publiés (une discussion est associée à chaque message), de rassembler sur une même page des contenus publiés sur des autres sites, de partager ces contenus avec ses contacts comme on peut le faire sur d'autres réseaux sociaux [35].

La figure suivante montre l'interface de l'agrégateur FreindFeed :

Figure III.14 : Interface de l'agrégateur FriendFeed [36].

III.8.2. Spokeo :

Spokeo, un service d'agrégation de blogs et réseaux sociaux. Spokeo se positionne désormais comme un lecteur de flux RSS pour les réseaux sociaux, vous permettant ainsi de suivre en « temps réel » les mises à jour de vos contacts en un seul endroit (avec un système similaire aux mini-feed Facebook). Avec cet agrégateur, il est possible de connaître les vidéos favorites de vos amis sur YouTube, leurs dernières photos uploadées sur MySpace, ce qu'ils ont posté sur leurs blogs Friendster… Spokeo déclare être le plus simple des agrégateurs de réseaux sociaux [27].

La figure suivante montre l'interface de l'agrégateur Spokeo :

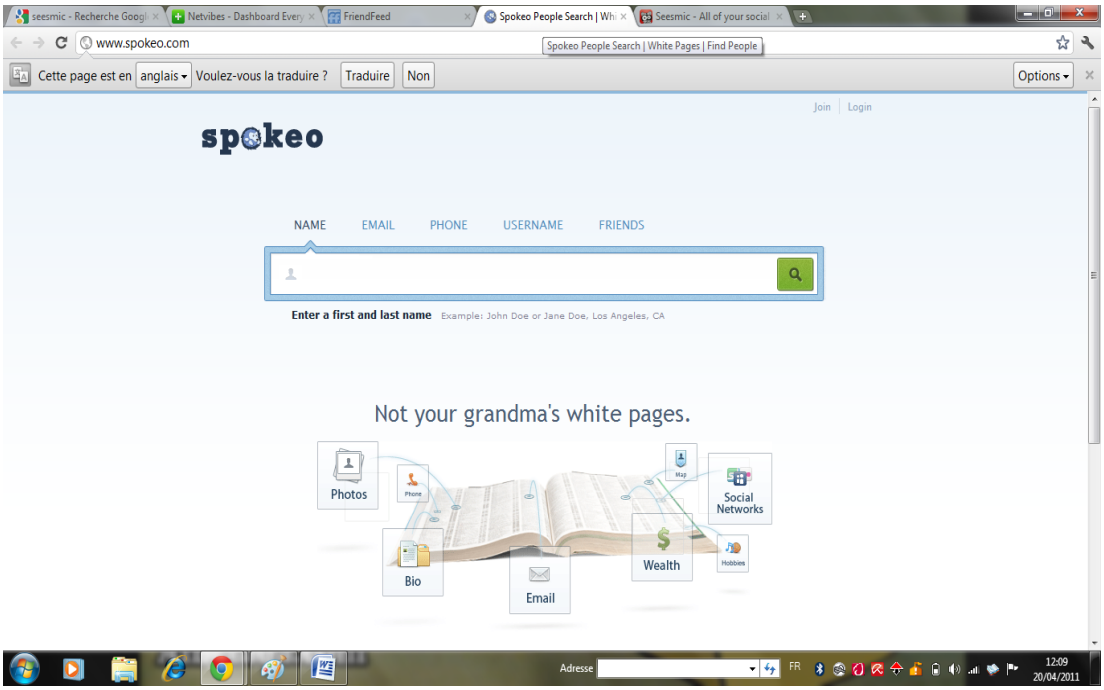


Figure III.15 : Interface de l'agrégateur Spokeo [37].

III.8.3. Netvibes :

III.8.3.1. Présentation :

Netvibes offre à ses utilisateurs un site Web personnel constitué par des pages onglets. Ce site est, à toute fin pratique, un portail Web individuel qui donne accès à une multitude de services. Chaque service se présente comme un bloc.

La page d'accueil de ce site se décompose en modules, représentés graphiquement par des blocs rectangulaires.

III.8.3.2. Historique :

Netvibes a été lancé officiellement le 15 septembre 2005 en concurrence avec Google (et son produit iGoogle, lancé en mai 2005) ou Microsoft (avec Live.com) sur le nouveau segment des pages d'accueil personnalisables. D'après Feedburner, le site est devenu en environ un an le troisième site de lecture de flux RSS au monde.

En octobre 2009, la nouvelle version nommée Netvibes Wasabi, accessible elle aussi de manière limitée grâce à des codes d'invitations limités en nombre d'utilisations,

apporte une nouvelle interface de lecture de l'information agrégée, en particulier des flux RSS [28].

La figure suivante montre l'interface de l'agrégateur Netvibes :

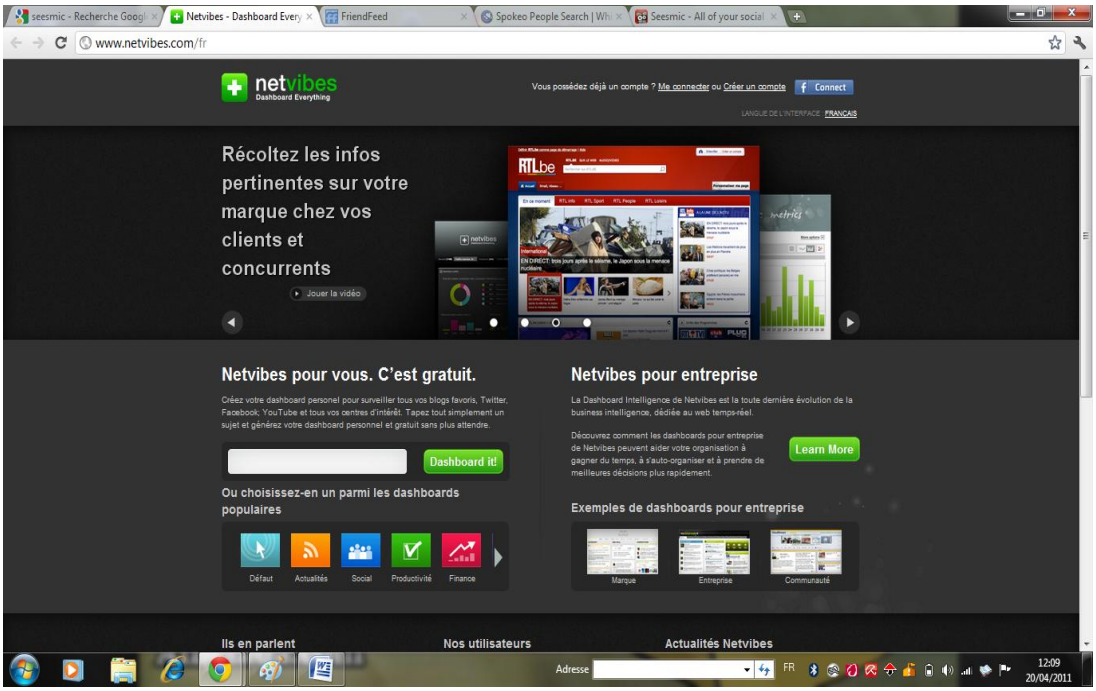


Figure III.16 : Interface de l'agrégateur Netvibes [38].

III.8.3.3. Utilisations documentaires :

Les pages Netvibes permettent d'utiliser des fonctionnalités apportées par le Web 2.0 comme l'agrégation de flux en ligne, la diffusion contenu multimédia et hypermédia ainsi que les réseaux sociaux. L'utilisateur peut :

- Créer une partie privée accessible par authentification (page personnelle) et une partie publique ;
- Développer une veille informationnelle et documentaire (facilitée par l'accès aux flux d'informations) ;
- Contribuer à un réseau social intégré à ce site, qui permet notamment de repérer le travail documentaire d'autres utilisateurs et faire de ce système un espace où l'information peut être récupérée et transmise entre les différents utilisateurs,
- Mettre à disposition des informations ;
- Informer d'autres usagers ou clients ;

➢ Créer un bureau virtuel.

Chaque partie publique agit donc comme un réservoir de ressources dans lequel chaque utilisateur peut prendre des références et alimenter sa propre page par le principe du mixage (mashup) [28].

III.8.4. Seesmic :

Seesmic fourni par la start up de Loïc Le Meur, a pour objet la création de passerelles entre différents médias sociaux. Après récupération des données des comptes Facebook, Twitter, Linkedin, Google Buzz et autres, on peut accéder à partir d'un seul écran aux informations des différents réseaux, avec possibilité de poster, de suivre les flux d'activités et d'accéder aux profils des utilisateurs. Seesmic permet d'aller un cran plus loin et d'interagir avec des réseaux plus professionnels [29].

La figure suivante montre l'interface de l'agrégateur Seesmic :

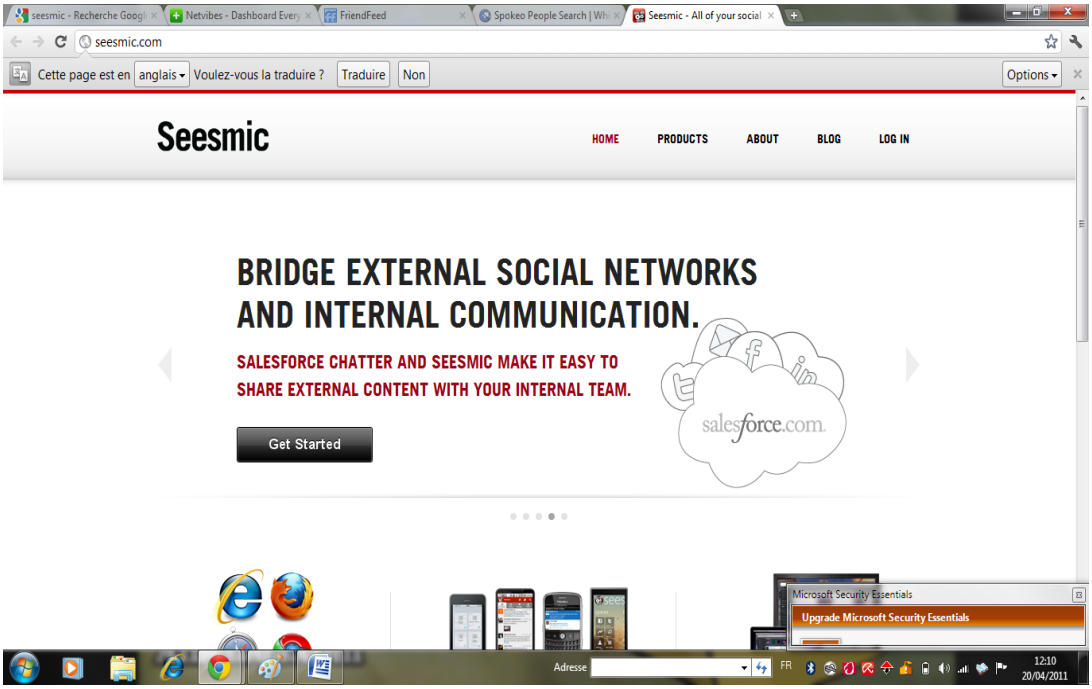


Figure III.17 : Interface de l'agrégateur Seesmic [39].

III.9. Conclusion :

D'après notre étude d'un certain nombre d'agrégateurs (Freindfeed, Netvibes, Spokeo et Seesmic), nous avons vu que chaqu'un de ces derniers est spécifié par des caractéristiques comme suit :

- Freindfeed, est l'agrégateur d'amis, son objectif est d'agréger touts les informations des contacts de l'utilisateur sur les différents réseaux sociaux.
- Spokeo, est un service d'agrégation de blogs et de réseaux sociaux, il s'intéresse par exemple aux vidéos favorites de vos amis sur Youtube.
- Netvibes, est un agrégateur qui offre à ses utilisateurs un site Web personnel constitué par des pages onglets.
- Seesmic, c'est un agrégateur qui a pour objectif de créer des parsselles entre les différents médias sociaux.

Chapitre IV:
Approche retenue

IV.1. Introduction :

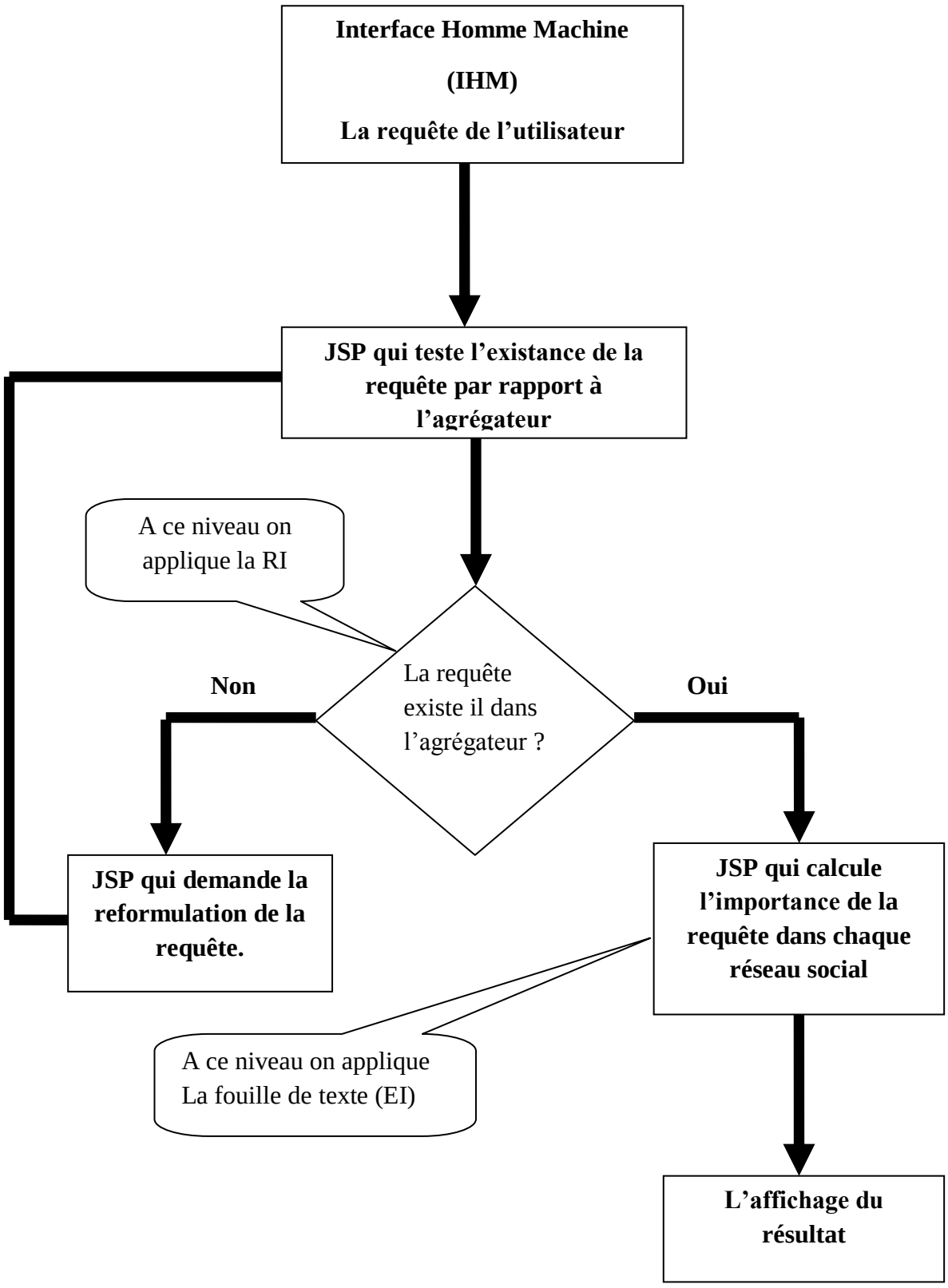
Dans ce chapitre, nous allons décrire notre contribution dans le cadre de ce projet de fin d'étude. Cette contribution se base sur deux étapes, la première étape est consacrée à la Recherche d'Information (RI), la deuxième étape s'intéresse à la fouille de texte.

Dans la première étape, on propose à l'utilisateur une interface homme machine (IHM), afin de formuler la requête de son besoin par rapport à un agrégateur, à ce stade on parle de la recherche d'information. Dans la deuxième étape, on a calculé l'importance de la requête utilisateur dans chaque réseau social (Facebook, Twitter, Myspace….), pour lui aider à trouver la bonne source d'information, à ce niveau on a appliqué l'extraction d'information (la fouille de texte).

Dans ce qui suit, nous allons présenter un exemple de notre approche :

Un utilisateur qui souhaitait trouver des informations sur une requête bien précisé (« Grève universitaire ») par exemple, peut formuler sa demande dans l'IHM et au lieu d'aller ouvrir chaque page web qui traite cette requête, notre application lui guide directement vers le réseau social qui répond à ses attentes.

IV.2. Schéma de notre application :



IV.3.Résultat de l'application :

IV.3.1. Interface homme machine (IHM) :

Dans cette phase, nous avons réalisé une IHM pour faire dialoguer l'utilisateur avec l'agrégateur Netvibes en ligne. Pour ce la nous avons utilisé une JSP qui contient d'une partie de HTML pour afficher notre interface, et d'autre part du Java pour récupérer la requête de l'utilisateur.

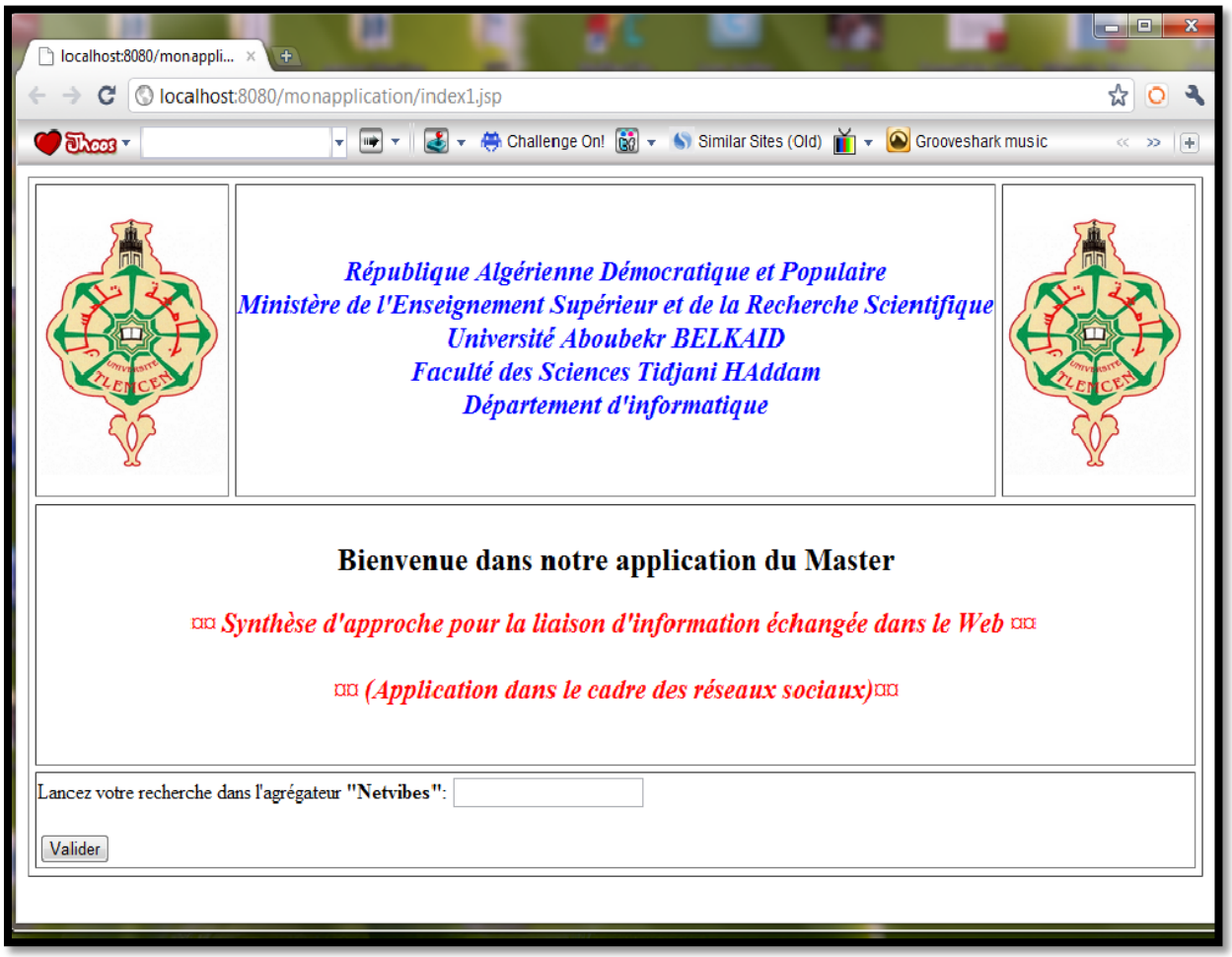


Figure IV.18 : Interface IHM de notre application.

IV.3.2. Le test d'existante de la requête utilisateur :

A ce niveau, nous avons utilisé le parser HTML (htmlcontenentparser [32]) qui a comme objectif de parser le contenue de l'agrégateur « Netvibes » pour tester l'existante de la requête utilisateur (A ce stade nous avons appliqué la recherche d'information « Voir Annexe A »). Donc si la requête existe au moin une fois, il passe à la dernière partie (parser le code source), sinon le JSP demande à l'utilisateur de reformuler sa requête comme suit :

Figure IV.19 : Interface de reformulation de la requête.

IV.3.3 : Le résultat final de notre application :

Dans la dernière étape, nous avons utilisé le parser HTML (Jéricho [33]) qui parse le code source de l'agrégateur, ainsi nous avons appliqué le « matching » pour chercher et calculer le nombre d'occurrence de besoin de l'utilisateur dans chaque réseau social et a ce niveau nous avons appliqué l'extraction d'information (la fouille de texte « Voir Annexe B »), afin de guider l'utilisateur vers la bonne source d'information.

Figure IV.20 : Interface du résultat de notre application.

IV.4.Conclusion :

Dans ce chapitre, nous avons présenté notre contribution dans le cadre de ce PFE. L'application réalisée fait appelle à la technologie JSP ainsi que parser HTML. Le mécanisme classique qui est utilisé avec les agrégateurs des réseaux sociaux a été amélioré à l'aide de notre application en utilisant la recherche d'information et l'extraction des connaissances. Les résultats obtenus avec Notre approche sont prometteur car on arrive à diriger l'utilisateur au bon réseau social qui répond le mieux à son besoin.

IV.5.Annexes :

IV.5.1.Annexe A :

```
<%@page import="java.io.*, java.net.*"%>

<%@page import="java.util.regex.*"%>

<HTML>

<BODY>

<%   String sourceLine;

       String content = "";

  URL address = new URL("http://www.netvibes.com/privatepage/1#algerie_actualites");

InputStreamReader pageInput = new InputStreamReader(address.openStream());

BufferedReader source = new BufferedReader(pageInput);

while ((sourceLine = source.readLine()) != null)

      content += sourceLine + "\t";

Pattern style = Pattern.compile("<style.*?>.*?</style>");

Matcher mstyle = style.matcher(content);

while (mstyle.find()) content = mstyle.replaceAll("");

Pattern script = Pattern.compile("<script.*?>.*?</script>");

Matcher mscript = script.matcher(content);

while (mscript.find()) content = mscript.replaceAll("");

Pattern tag = Pattern.compile("<.*?>");

Matcher mtag = tag.matcher(content);

while (mtag.find()) content = mtag.replaceAll("");

Pattern comment = Pattern.compile("<!--.*?-->");

Matcher mcomment = comment.matcher(content);

while (mcomment.find()) content = mcomment.replaceAll("");

Pattern sChar = Pattern.compile("&.*?;");

Matcher msChar = sChar.matcher(content);

while (msChar.find()) content = msChar.replaceAll("");

Pattern nLineChar = Pattern.compile("\t+");

Matcher mnLine = nLineChar.matcher(content);
```

```
while (mnLine.find()) content = mnLine.replaceAll("\n");
String requet = request.getParameter("requet");
boolean rep1 = content.contains(requet);
if (rep1  == true) { %>
<jsp:forward page="Nextepage3.jsp">
   <jsp:param name="requet" value="requet"/>
</jsp:forward>
<% } else { %>
<jsp:forward page="index3.jsp"/>
<% } %>}
```

IV.5.2.Annexe B :

```
<%@page import="java.util.regex.Pattern"%>
<%@page import="java.util.regex.Matcher"%>
<%@ page import="java.io.IOException" %>
<%@ page import="java.net.URL" %>
 <%@ page import="java.util.List" %>
<%@ page import="net.htmlparser.jericho.*" %>
<html>
<body>
      <%!
    public String affichage () throws IOException {
       String sourceUrlString="http://www.netvibes.com/privatepage/1#algerie_actualites";
       if (sourceUrlString.indexOf(':')==-1) sourceUrlString="file:"+sourceUrlString;
                MasonTagTypes.register();
                Source source=new Source(new URL(sourceUrlString));
    String chaine="";
       List<Element> elementList = source.getAllElements("div");
       for (Element element : elementList) {
          chaine=chaine+"\n"+element.getContent();
```

```
    }
    return chaine;
}
public double[] conteurRS (String requet) throws IOException{
    String chaine=affichage();
    double[]tab={40,20,15,15,10};
    String motif="<a href=www.facebook.org>" +requet+ "";
    String motif2="<a href=www.twiter.com>" +requet+ "";
    String motif3="<a href=www.Myspace.com>" +requet+ "";
    String motif4="<a href=www.linked.com>" +requet+ "";
    String motif5="<a href=www.orkut.com>" +requet+ "";

     Matcher matcher = Pattern.compile(motif).matcher(chaine);
     Matcher matcher1 = Pattern.compile(motif2).matcher(chaine);
     Matcher matcher2 = Pattern.compile(motif3).matcher(chaine);
     Matcher matcher3 = Pattern.compile(motif4).matcher(chaine);
     Matcher matcher4 = Pattern.compile(motif5).matcher(chaine);
       int occur = 0;
       int occur2 =0;
       int occur3=0;
       int occur4=0;
       int occur5 = 0;
       while (matcher.find()) {
           tab[0]=occur++;}
      while (matcher1.find()) {
         tab[1]= occur2++; }
      while (matcher2.find()) {
           tab[2]= occur3++;}
       while (matcher3.find()) {
           tab[3]= occur3++;}
```

```
        while (matcher4.find()) {
          tab[4]= occur3++; }
return tab ;  }
%>
      <%
  String requet = request.getParameter("requet");
  double[]tab= conteurRS(requet);
  %>
```

Conclusions générales

Le travail réalisé dans ce PFE sert à évoluer le mécanisme classique des agrégateurs des réseaux sociaux afin d'orienter l'utilisateur vers la bonne source d'information qui répond a son besoin.

Notre mémoire est articulé autours de deux partie ; la première est théorique, il s'agit de l'état de l'art et elle contient trois chapitres , le premier introduit le Web depuis le Web 2.0 jusqu'au le web social , le seconde parle de la fouille de textes qui consiste a extraire les connaissances pertinentes selon un ensemble de méthodologies et le troisième présente les agrégateurs qui ont comme objectif de proposer a l'utilisateur un large choix de sources d'information et pour cela nous avons étudié quelque agrégateurs de réseaux sociaux comme Netvibes, Spokeo , FiendFeed, Seesmic,...

La seconde partie de notre mémoire est la partie pratique ou l'application du projet dans laquelle nous avons appliqué tous ce que nous avons déjà étudié dans la phase précédente (la phase théorique). Notre application est une application Web qui est basée sur l'utilisation des JSP et encore l'utilisation de deux types de parsers HTML ; le premier qui a comme objectif de parser le contenu de l'agrégateur « Netvibes » et tester si la requête utilisateur existe (nous avons appliquer la recherche d'information « RI ») et le deuxième qui a comme objectif de parser le code source de l'agrégateurs pour extraire la bonne information de chaque réseau social (nous avons appliquer la fouille de textes ou l'extraction d'information « EI ») et en fin le résultat de notre travail est visualiser a l'utilisateur pour lui dirigé vers le réseau social qui répond le mieux à son besoin.

Comme perspective de ce travail, il serait intéressant de tester notre application sur n'importe contenu web afin d'extraire l'information pertinente selon le besoin de l'utilisateur. Aussi une amélioration possible serait intéressante dans le cadre du web sémantique.

Références

[1] http://www.web-libre.org/dossiers/web-2-0,7251.html

[2] http://fr.wikipedia.org/wiki/Web2.0

[3] http://www.web-libre.org/dossiers/web-2-0,7251

[4] http://charlespauze.wordpress.com/2010/08/05/le-web-2-0-cest-quoi-au-juste/

[5] www.ossir.org

[6] www.dialog.ac-reims.fr/ecogestion/IMG/pdf/articleweb2

[7] http://www.clever-age.com/veille/clever-link/web-2.0-definitions-et-composantes-partie-1-sur-2-.htmlweb 2.0

[8] www.slideshare.net/.../powerpoint-sur-le-web-20-presentation

[9] www.info.univ-angers.fr/pub/genest/.../m2.../ws_chap1.pdf

[10] www.emse.fr/~beaune/websem/03-WebSemantique.pdf

[11] www.inrialpes.fr/exmo/cooperation/asws/ASWS-Langages.pdf

[12] http://eric.univ-lyon2.fr/~social-web/

[13] http://fr.wikipedia.org/wiki/Websocial

[14] http://fr.wikipedia.org/wiki/Web_3.0

[15] http://www.allbestarticles.com/internet/web-development/difference-between-web1.0-web2.0-and-web3.0.html

[16] http://fr.wikipedia.org/wiki/Fouille_de_textes

[17] paris13.fr/A3/AAFD06/slides/AAFD06_J123-kodratoff.pdf

[18] http://users.info.unicaen.fr~brunopapersLucasCremilleuxDNdef

[19] Httpwww.spss.chupload1055262370_tm

[20] http://zincdesliens.wordpress.com/2011/01/06/les-agregateurs-de-reseaux-sociaux/

[21] http://www.placedesreseaux.com/Dossi ers/reseau-relationnel/agregateurs-de-reseaux-sociaux-1.html

[22] http://www.commentcamarche.net/faq/3339-agregateurs-rss-lecteurs-de-fils-rss

[23] http://www.techno-science.net/?onglet=glossaire&definition=604

[24] http://fr.wikipedia.org/wiki/Agrégateur

[25] http://www.blueboat.fr/reseaux-sociaux-friendfeed-lagregateur-damis

[26] http://www.logiste.be/blog/beta-friendfeed-suivre-vos-amis-sur-le-web/

[27] http://fr.mashable.com/2007/07/14/spokeo-un-agregateur-de-reseaux-sociaux/

[28] http://fr.wikipedia.org/wiki/Netvibs

[29] http://pro.01net.com/editorial/524504/seesmic-passerelle-entre-les-reseaux-sociaux/

[30] http://biologie.univ-mrs.fruploadp93B.JACQ_Text_Mining

[31] http://www.ladocumentationfrancaise.fr/dossiers/internet-monde/web2.0.shtml.

[32]http://code.google.com/p/goaround/source/browse/branches/GoAroundWeb/src/br/com/fiap/goaround/parser/HtmlContentParser.java?r=42

[33] http://jericho.htmlparser.net/docs/index.html

[34] http://www.inaativ.com/images/stories/e-books/PDF/guide_recherche_en_ligne.pdf

[35] http://www.placedesreseaux.com/Dossiers/reseau-relationnel/agregateurs-de-reseaux-sociaux-5.html

[36] http://www.freindfeed.com

[37] http://www.spokeo.com

[38] http://www.netvibes.com

[39] http://www.sesmic.com

Résumé

Dans le cadre de ce PFE, nous avons proposé une nouvelle approche afin d'évoluer le mécanisme classique des agrégateurs des réseaux sociaux. En effet, l'utilisateur est guidé à travers notre application vers la bonne source d'information qui répond le mieux à son besoin. Notre approche est basée sur une application Web qui utilise les pages web dynamiques (JSP) ainsi que deux types de parsers HTML ; le premier parser a comme objectif de parser le contenu de l'agrégateur et tester si la requête utilisateur existe (nous avons appliquer à ce niveau la recherche d'information « RI »), le deuxième parser a comme objectif de parser le code source de l'agrégateurs pour extraire la bonne information de chaque réseau social (nous avons appliquer à ce stade la fouille de textes ou l'extraction d'information « EI »). L'agrégateur choisi pour le test est Netvibes. Le résultat final obtenu à l'aide de notre approche est très satisfaisant vu que l'utilisateur est guidé vers le réseau social qui répond le mieux à son besoin.

Abstract:

In this report which is the result of our master thesis project, we propose a new approach to evolve the traditional mechanism of social network aggregators. Indeed, the user is guided through our application to the pertinent source of information that satisfies his/her needs. Our approach is based on JavaServer Pages and two types of HTML parsers, the first one is used to parse the content of the aggregator in order to find the user's query (we apply at this level, information retrieval IR), the next one is used to parse the source code of the aggregator in orer to retrieve the correct information for each social network (in this stage we apply the text mining or extraction of information "IE"). The aggregator chosen for the test is Netvibes. The final result obtained using our approach is very satisfactory because the user is guided to the social network that satisfies his/her needs.

ملخص:

في إطار هاته المذكرة اقترحنا منهجا جديدا لتطوير الآلية الكلاسيكية لتجميع الشبكة الاجتماعية. في الواقع هو توجيه المستخدم من خلال تطبيقنا للمصدر الحقيقي للمعلومات التي تناسب احتياجاته. ويستند منهجنا على تطبيق ويب الذي يستخدم صفحات الويب الديناميكية ، ونوعين من موزعي [محللي] اللغة الرمزية الإعلامية النصية , الأول هدفه تحليل محتوى للمجمع و اختبار إذا كان طلب المستخدم موجود (نطبق في هذا المستوى :بحث المعلومات)، الثاني هدفه تحليل شفرة المصدر لتجميع و إسترداد المعلومات الصحيحة لكل شبكة اجتماعية (نطبق في هذا المستوى :التنقيب النصي أو إستخراج المعلومات). النتيجة النهائية المحصل عليها باستخدام منهجنا مرضية للغاية نظرا لتوجيه المستخدم إلى الشبكة الاجتماعية التي تناسب احتياجاته.