Aide à l'extraction de la sémantique des liens hypertextes

Moustafa Al-Hajj, Gilles Verley, Hubert Cardot

Aide à l'extraction de la sémantique des liens hypertextes

Moustafa Al-Hajj*, Gilles Verley*, Hubert Cardot*

Université François-Rabelais de Tours
Laboratoire d'Informatique (EA 2101),
64, Avenue Jean Portalis,
37200 TOURS – France
http://www.li.univ-tours.fr
***prenom.nom@univ-tours.fr**

Résumé :

Les auteurs qui publient sur le Web des connaissances sous la forme de documents électroniques lisibles sur un écran utilisent de plus en plus la technologie des liens hypertextes pour améliorer l'ergonomie de leur sites et pour les enrichir par des informations provenant d'autres sites Web. Nous nous intéressons à la sémantique des liens hypertextes, en termes d'extraction et d'exploitation, dans le but de faciliter la recherche d'information sur le Web. Dans cet article, nous proposons une méthodologie originale d'extraction de la sémantique des liens hypertextes par des moyens manuels et semi-automatiques. Dans une première partie, nous montrons comment nous avons constitué un corpus de documents sur le Web, qui sera par la suite notre base de test. Cette constitution consiste à extraire un sous-ensemble du Web, regroupant des pages ayant des critères utiles à l'étude de la sémantique des liens hypertextes. Ensuite nous proposons une méthode d'analyse de la sémantique des liens hypertextes. Celle-ci consiste à faire l'analyse sémantique du contexte appelant du lien et du contexte appelé par le lien, et à expliciter de manière formelle la relation sémantique entre le contexte appelant et le contexte appelé. La dernière partie est consacrée à l'élaboration d'outils d'aide à l'analyse, nous proposons une automatisation de la reconnaissance des formes littéraires des contextes appelant des liens et des contextes appelés par des liens avec les treillis de Galois.

Mots Clés : Constitution de corpus ; analyse sémantique de liens hypertextes ; ontologie ; RDF(S) ; treillis de Galois ; K-means.

1. Introduction

Les auteurs qui publient sur le Web des connaissances sous la forme de documents électroniques lisibles sur un écran utilisent la technologie des liens hypertextes pour améliorer l'ergonomie de leurs sites et pour les enrichir par des informations provenant d'autres sites Web.
La différence entre les approches d'écriture et de lecture hypertextuel et de celles pour les papiers réside dans la structure non linéaire de l'hypertexte. Le fait qu'à partir d'un nœud le lecteur peut se retrouver sur d'autres nœuds grâce à l'activation des liens hypertextes supposerait que l'hypertexte contienne une multiplicité de parcours de lecture. Cette multiplicité des parcours risque d'égarer le lecteur habitué à une approche traditionnelle de l'information écrite. Il est donc intéressant d'offrir aux utilisateurs un moyen de naviguer dans les réseaux hypertextes, qui limite les risques de désorientation et de surcharge cognitive. Il s'agit d'offrir aux lecteurs des outils lui permettant d'accéder à la sémantique de l'information.
La mise en évidence de la sémantique des nœuds d'un hypertexte et de relations sémantiques entre ces nœuds aide le lecteur à s'orienter dans sa recherche d'information. La sémantique des nœuds et la relation sémantique entre ces nœuds permettent au lecteur d'avoir un contexte de navigation, elles sont très utiles pour cibler l'information pertinente plus rapidement. En associant à chaque nœud un sous-ensemble de termes pertinents et à chaque lien la nature de la relation entre les nœuds qu'il lie, on peut offrir au lecteur un moyen de se déplacer sémantiquement dans le graphe hypertextuel.
Cela nous a motivé à nous intéresser à la sémantique des liens hypertextes. Dans cet article, nous proposons une méthodologie originale d'extraction de la sémantique des liens hypertextes par des moyens manuels et semi-automatiques. Pour vérifier la validité de la méthode, nous l'avons testée sur les liens hypertextes d'un corpus spécifique sélectionné sur le Web.
Dans une première partie, nous montrons comment nous avons constitué le corpus. Ensuite nous proposons une méthode d'analyse de la sémantique des liens hypertextes. Celle-ci consiste à faire l'analyse sémantique du contexte appelant du lien et du contexte appelé par le lien, et à expliciter de manière formelle la relation sémantique entre le contexte appelant et le contexte appelé. La dernière partie est consacrée à l'élaboration d'outils d'aide à l'analyse, nous proposons une automatisation de la reconnaissance des formes littéraires des contextes appelants des liens et des contextes appelés par des liens avec les treillis de Galois.

2. Constitution du corpus et ontologie du domaine

Le corpus va constituer notre base de test, on l'obtient en extrayant un sous-ensemble du Web, regroupant des pages ayant des critères utiles pour l'étude de la sémantique des liens hypertextes. Le thème retenu est la biographie de personnages célèbres.
Nous présentons d'abord les critères de sélection des documents et ensuite une partie de l'ontologie du domaine représentée en RDF(S) (Resource Description Framework Scheme).

2.1. Critères de sélection

Les documents du corpus ont été sélectionnés par rapport à plusieurs critères :

2.1.1. Sujet et langue des documents

Notre corpus est constitué de documents traitant des biographies de personnages célèbres, ce choix est dû à la richesse de ces documents en liens hypertextes, les auteurs les utilisent en effet pour améliorer l'ergonomie de leur sites Web et pour les enrichir par des informations provenant d'autres sites sur le Web, etc. [VAN00].
Pour faciliter l'annotation sémantique des liens hypertextes, le corpus a été limité aux pages écrites en français.

2.1.2. Variétés des auteurs et de serveurs de documents et type de mise en page

Les documents du corpus sont issus de serveurs différents, les auteurs des biographies varient d'un serveur à un autre. Ceux-ci ont beaucoup de raisons différentes de poser des liens hypertextes dans leur propos, de plus les documents du corpus ont des formes littéraires très diverses, ceci constitue une grande richesse pour le corpus.
L'interrogation des outils de recherche francophones (annuaires et moteurs de recherche) comme Google (www.google.fr) et Altavista (www.altavista.com) et Yahoo (www.yahoo.fr), avec les requête « biographie » et/ou « nom d'un personnage célèbre » réduite aux pages satisfaisant les critères de sélection ainsi définis, a permis d'obtenir un ensemble de 140 biographies provenant d'au moins 19 serveurs différents, l'ensemble de biographies contient plusieurs milliers de liens hypertextes.
Les documents du corpus sont des documents HTML, ce choix a été fait car la majorité des documents sur le Web sont encore en format HTML, et permet la standardisation de certains traitements afin d'étudier la sémantique des liens hypertextes.

2.1.3. Liens natifs et répondant à des besoins variés

On entend par *lien natif,* un lien *voulu* par l'auteur, contrairement à un *lien calculé* par des automates.
Les auteurs posent des liens hypertextes dans leurs propos pour satisfaire différents besoins, ces besoins sont tels que :

– La généralisation du propos de l'auteur, comme un lien vers un sommaire.
– La spécialisation du propos de l'auteur, comme un lien situé dans un sommaire.
– L'illustration du propos de l'auteur, tel qu'un lien vers une photo.
– Etc.

Toute relation sémantique d'un propos de l'auteur avec un autre propos, peut être à l'origine de la pose d'un lien hypertexte dans le premier propos vers le second.

Nous avons sélectionné les sites constitués essentiellement de *liens natifs* et dont la sémantique était variée. Voici une typologie des relations sémantiques portées par les liens de notre corpus et que nous utiliserons dans la seconde partie.

– Une personne citée dans le propos de l'auteur *s'est opposé à, a connu, a été maître de, a fréquenté, est l'élève de, soutient, a été soutenu par, est la fille de, est la grand-mère de, est le fils de, est le grand-père de, est le mari de, est parent d'* une personne dont la biographie est traitée dans la cible du lien.

– Une personne citée dans le propos de *l'auteur a étudié, a écrit, a joué, a adapté, a fondé, a lu, a réalisé, a découvert, a traduit, a utilisé, a commenté, a détaillé, a retranscrit, s'est inspiré de, est à l'origine d'* une chose citée dans le propos cible du lien.

– Une chose citée dans le propos de l'auteur *a été étudiée par, a été écrite par, a été jouée par, a été adaptée par, a été fondée par, a été lu par, a été réalisée par, a été découverte par, a été traduit par, a été utilisée par, a été commentée par, a été détaillée par, a été retranscrite par* une personne dont la biographie est traitée dans la cible du lien.

– Une personne citée dans le propos de l'auteur *a participé à* un évènement cité dans la page ciblée par le lien.

– Une personne citée dans le propos de l'auteur *a travaillé, a vécu, a voyagé, est mort, est né* dans un lieu cité dans la cible du lien.

– Une chose citée dans le propos de l'auteur *a été jouée en* un lieu cité dans la cible du lien.

– Une chose citée dans le propos de l'auteur *donne accès à, est apparenté à, est comparé à, est détaillé par, est illustré par, a influencé, explique, fait partie de, illustre, a été commentée par, parle de, est représenté par, représente, est généralisée par, contient* une chose citée dans le propos cible du lien.

– Un évènement cité dans le propos de l'auteur *a eu lieu* dans un lieu cité dans la cible du lien.

– Un évènement cité dans le propos de l'auteur *est illustré par* une chose citée dans le propos cible du lien.

– Un lieu cité dans le propos de l'auteur *est l'endroit d'*un évènement cité dans la cible du lien.

Exemple d'un lien natif

Considérons une page de la biographie de François Mitterrand et une page qui a pour sujet la convention de Lomé IV. La dernière est, soit faite par le même auteur, soit par un autre.

Dans la première page, l'auteur cite les oeuvres économiques étrangères de François Mitterrand, et parmi elles, la convention de Lomé IV. Dans la partie qui cite la convention de Lomé IV, l'auteur pose un lien hypertexte vers la page qui a pour sujet la convention de Lomé IV. Le contexte appelant du lien est supposé être la partie de la page du lien qui cite les œuvres économiques étrangères, et le contexte appelé par le lien est supposé être la cible du lien qui traite la convention de Lomé IV.

La relation sémantique entre le contexte appelant du lien et du contexte appelé par le lien est la suivante : « Les oeuvres économiques étrangères de François Mitterrand » *contient* « La convention de Lomé IV ».

Si l'auteur de la page biographique n'est pas le même que celle de la convention de Lomé IV, la découverte par l'auteur de l'existence de la page sur la convention de Lomé IV et de son adresse, l'aura motivé à poser un lien vers celle-ci dans sa page, plus précisément dans la partie qui traite des oeuvres économiques étrangères, dans la sous-partie qui cite la convention de Lomé IV.
Et si l'auteur de la page de la convention de Lomé IV est aussi celui de la page biographique de François Mitterrand, ce sont les avantages d'utilisation des liens hypertextes, pour améliorer l'ergonomie des sites Web, qui l'auront motivé à créer la page sur la convention de Lomé IV à part, et à poser un lien vers cette page dans le texte principal.

2.1.4. Importance de la relation sémantique portée par les liens natifs

Nous nous intéressons dans cet article à expliciter la relation sémantique des liens natifs car nous faisons l'hypothèse que cette relation sémantique est suffisamment importante pour avoir motivé la pose d'un lien par l'auteur lui-même et qu'ainsi elle est hautement pertinente, nous pensons également qu'il est plus simple d'expliciter formellement les relations sémantiques portées par les liens natifs que celles entre les autres phrases du texte.

2.1.5. Variété des formes littéraires des contextes des liens et des contextes appelés par les liens

Pour la suite, on nomme « contexte appelant d'un lien » l'ensemble minimal de textes, caractères et objets, autour du lien et qui constituent une seule idée, concept ou sujet [VIG01].
De même, on nomme « contexte appelé par un lien » l'ensemble minimal de textes, caractères et objets de la page ciblée par le lien et qui constituent un sujet en rapport avec le « contexte appelant du lien » [VIG01].
Les contextes, qu'ils soient contextes appelants ou appelés, peuvent être de diverses formes littéraires. On les a regroupés de la manière suivante :
Forme Sommaire ; Forme Illustration Graphique ; Forme Définition ; Forme Citation ; Forme Liste ; Forme Référentielle ; Forme Récit ; Forme Description ; Forme Détail ; Forme Résumé.
Cet aspect est traité en §4.2.2.

2.2. Ontologie des liens hypertextes du corpus : représentation par RDFS

Une fois notre corpus constitué, nous avons défini une ontologie des liens hypertextes du domaine considéré.
Cette partie a pour objectif de présenter une partie de l'ontologie, et de montrer qu'elle est représentable par les technologies RDF(S) (Resource Description Framework Scheme) [RDFS] [RDF] [CHAR03]. Les concepts de l'ontologie sont : Personne, Chose, Lieu, Evènement, Date, FormeContexte. Les relations sont celles du §2.1.3. La représentation de l'ontologie par RDFS est la suivante :

```
<?xml version="1.0" encoding="UTF-8" ?>
<rdf:RDF
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns:owl="http://www.w3.org/2002/07/owl#"
        xmlns:dc=http://purl.org/dc/elements/1.1/ >
<owl:Ontology
        rdf:about="URIOntologieLiensHypertextes"
        dc:title="l'ontologie de liens hypertextes"/>
<rdfs:Class rdf:about="URIOntologieLiensHypertextes#Personne">
          <rdfs:label>Personne</rdfs:label>
          <rdfs:comment>
                  Une personne décrite par son nom et son prénom
          </rdfs:comment>
</rdfs:Class>
…
<!-- Les autres concepts : Chose,  Lieu, Evènement, Date sont représentés de la
même manière -->
<rdfs:Class rdf:about="URIOntologieLiensHypertextes#FormeContexte">
<rdfs:label>FormeContexte</rdfs:label>
        <rdfs:comment>
                  La forme littéraire du contexte appelant du lien ou du contexte
                  appelé par le lien
          </rdfs:comment>
</rdfs:Class>
<rdf:Property rdf:ID="a travaillé">
         <rdfs:domain rdf:resource="#Personne" />
         <rdfs:range rdf:resource="#Lieu" />
</rdf:Property>
…
<!--Toutes les relations sémantiques du §2.1.4 sont représentées de la même
manière -->
</rdf:RDF>
```

3. Méthode proposée pour effectuer l'analyse manuelle sémantique d'un lien

Pour effectuer l'analyse sémantique manuelle d'un lien hypertexte, la méthode proposée consiste à faire l'analyse sémantique des deux contextes, contexte appelant du lien et contexte appelé par le lien, et à trouver la relation sémantique entre le contexte appelant et le contexte appelé.
On fait l'hypothèse que la raison pour laquelle l'auteur a posé ce lien est contenue dans ces trois analyses sémantiques.

3.1. Analyse sémantique des contextes appelants et appelés

L'analyse sémantique des deux contextes, contexte appelant du lien et contexte appelé par le lien, consiste à les décrire dans une phrase composée de trois parties :

– La première pour dire qu'il s'agit d'un contexte du lien ou d'un contexte appelé par le lien.

– La deuxième pour décrire la forme littéraire du contexte - appelant ou appelé - qu'on analyse.

– La troisième pour décrire, par quelques mots clés reliés dans une phrase dans langage naturel, le contexte appelant (resp. appelé) en cours d'analyse.

La forme littéraire peut être choisie dans la liste de formes parmi celles citées en §2.1.5. Les mots clés les plus représentatifs du contexte - appelant ou appelé - peuvent être dérivés de l'ontologie du domaine, c'est-à-dire, ils peuvent être des instances des concepts de l'ontologie que nous avons vu en §2.2. Sinon, c'est-à-dire au cas où un mot clé ne peut être dérivé d'aucun concept de l'ontologie, il faudra ajouter de(s) nouveau(x) terme(s) à cette dernière pour que le mot clé puisse être dérivé de celle-ci.

3.2. Relation sémantique entre les deux contextes

Une fois l'analyse sémantique des deux contextes, appelant et appelé, terminée, on cherche à trouver une relation entre ces deux contextes. On l'explicite dans une phrase selon le modèle présenté en § 2.1.3.

Pour vérifier la validité de notre méthode d'analyse sémantique des liens hypertextes, nous l'avons appliquée sur un échantillon de 130 liens hypertextes. L'expérience montre que la méthode est opérationnelle, voici un exemple d'analyse d'un lien hypertexte selon la méthode, il sera suivi par une représentation de la sémantique du lien par RDF (Resource description Framework).

3.3. Exemple

Soit le lien hypertexte situé dans la page biographique de George Bush dans la partie qui raconte son repêchage dans l'océan pacifique, la cible du lien est une illustration par une photo du repêchage (voir figure 1), la sémantique est la suivante :

a) Sémantique du contexte appelant du lien :

« La source est le récit du repêchage de George Bush dans l'océan pacifique le 2 septembre 1944 ».

– « Source » : il s'agit de l'analyse du contexte appelant.

– « Récit » est la forme littéraire du contexte source de lien.

– « Repêchage ; George Bush ; océan pacifique ; 2 septembre 1944 » sont les mots clés décrivant le contexte appelant du lien.

b) Sémantique du contexte appelé par le lien :

« La cible est une illustration par une photo du repêchage de George Bush dans l'océan pacifique le 2 septembre 1944 ».

– « Cible » : il s'agit de l'analyse du contexte appelé.

– « Illustration graphique » est la forme littéraire du contexte cible de lien.

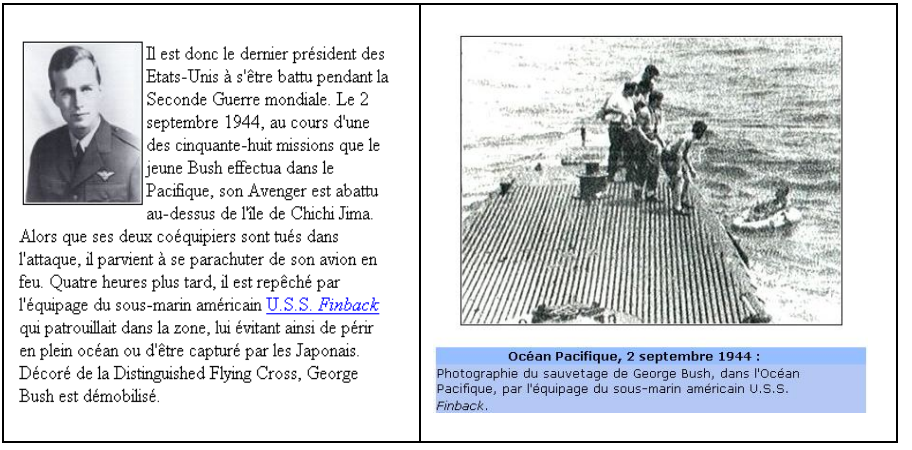
– « Photo ; Repêchage ; George Bush ; océan pacifique ; 2 septembre 1944 » sont des mots clés décrivant le contexte appelé par le lien.

c) Sémantique de la relation entre le contexte appelant et le contexte appelé : « Le repêchage est illustré par la photo ».

– « Repêchage » est un mot clé du contexte appelant.

– « photo » est un mot clé du contexte appelé.

– « est illustré par » est une relation sémantique entre les deux contextes.



Il est donc le dernier président des Etats-Unis à s'être battu pendant la Seconde Guerre mondiale. Le 2 septembre 1944, au cours d'une des cinquante-huit missions que le jeune Bush effectua dans le Pacifique, son Avenger est abattu au-dessus de l'île de Chichi Jima. Alors que ses deux coéquipiers sont tués dans l'attaque, il parvient à se parachuter de son avion en feu. Quatre heures plus tard, il est repêché par l'équipage du sous-marin américain U.S.S. *Finback* qui patrouillait dans la zone, lui évitant ainsi de périr en plein océan ou d'être capturé par les Japonais. Décoré de la Distinguished Flying Cross, George Bush est démobilisé.

Figure 1 – **à gauche : contexte appelant du lien « U.S.S. Finback » ; à droite : contexte appelé par le lien**

Représentation de la sémantique du lien par la technologie RDF [RDF] :

```
<rdf:RDF
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:ol="URIOntologieLiensHypertextes"
  >
<rdf:Description about="URIduContexteAppelantDulien">
        <ol:FormeContexte>Récit</ol:FormeContexte>
        <ol:Evenement>Repêchage</ol:Evenement>
        <ol:Personne>George Bush</ol:Personne>
        <ol:Lieu>Océan Pacifique</ol:Lieu>
        <ol:Date>2 septembre 1944</ol:Date>
<rdf:Description/>
<rdf:Description about="URIduContexteAppeléParLelien">
        <ol:FormeContexte>
                Illustration Graphique
        </ol:FormeContexte>
        <ol:Chose>Photo</ol:Chose>
        <ol:Evenement>Repêchage</ol:Evenement>
        <ol:Personne>George Bush</ol:Personne>
        <ol:Lieu>Océan Pacifique</ol:Lieu>
        <ol:Date>2 septembre 1944</ol:Date>
<rdf:Description/>
```

```
<ol:est illustré par>
        <ol:Evenement>Repêchage</ol:Evenement>
        <ol:Chose>Photo</ol:Chose>
<ol:est illustré par/>
</rdf:RDF>
```

4. Aide à l'analyse de la sémantique d'un lien hypertexte

Dans une perspective d'automatisation de l'extraction, selon notre méthode, de la sémantique d'un lien hypertexte, nous proposons une automatisation de la reconnaissance des formes littéraires des contextes appelants des liens et des contextes appelés par des liens. Tout d'abord, nous présentons une réflexion sur la délimitation formelle des contextes appelants et appelés, puis nous choisissons une méthode permettant de faire une telle délimitation, ensuite nous présentons des travaux sur la classification des pages du Web par leurs profils, nous définissons nos profils de contextes appelants et appelés avec leurs paramètres, une expérience de classification des contextes sera ensuite menée avec les treillis de Galois.

4.1. Délimitation des contextes appelants et appelés : réflexion et choix de méthode

A partir des définitions données aux contextes au § 2.1.5, nous avons eu plusieurs idées de délimitation formelle des contextes appelants et appelés, nous en citons quelques unes :

1) La première est de considérer le contenu de la cible du lien qui correspond à l'écran, comme support du contexte appelé par le lien, cette idée est due au point de vue suivant :

Les auteurs posent des liens dans leurs documents pour améliorer la lecture à l'écran. L'écran cible serait révélateur de ce que l'auteur a voulu faire.

2) La deuxième est de considérer le paragraphe contenant le lien comme support du contexte appelant du lien, car un paragraphe représente une idée ou un sujet.

3) La troisième est de considérer le contenu entre les deux balises « a name » précédant et succédant immédiatement le lien comme support du contexte appelant du lien, et le contenu de la cible du lien entre le début de la cible et la première balise « a name » comme support du contexte appelé par le lien, cette idée est le fruit du raisonnement suivant :

La présence d'un « a name » dans une page Web, signifie l'existence d'un lien hypertexte, interne ou externe à la page, pointant vers la partie de la page qui commence par le « a name ». Cela signifie l'existence d'un sujet pour celle-ci, en rapport avec le sujet du contexte appelant du lien qui pointe vers elle. De ce fait, l'existence de deux « a name » successifs dans une page, signifie l'existence de deux sujets pour ces deux parties. Ces deux sujets, l'un par rapport à l'autre, sont :

1. Indépendants.
2. Dépendants et dissociables, comme un sujet et un sous-sujet.

3. Dépendants et indissociables, dans ce cas l'idée du sujet de la première partie n'est pas complète en l'absence de la deuxième. Ce cas est rarement rencontré.
Dans les deux premiers cas, la partie comprise entre les deux « a name » constitue un sujet, et donc un contexte, appelant ou appelé, selon que cette partie contient le lien ou sa cible.
Nous avons opté pour la troisième idée, étant donné qu'on ne dispose pas de moyens pertinents pour pouvoir définir le rendu-écran (qui dépend des tailles des écrans, des navigateurs, etc.), et qu'il est facile d'extraire automatiquement les parties entre les balises « a name ». La deuxième idée de délimitation, à savoir, considérer le paragraphe comme étant le support du contexte appelant du lien, est en cours d'application.

4.1.1. Discussion :

Notre méthode de délimitation formelle des contextes appelants et appelés donne des résultats satisfaisants dans le cas ou les parties de la page à découper sont des cibles d'autres liens. Cependant, dans plusieurs pages, il n'existe aucun « a name », donc le seul contexte que nous pouvons repérer par la méthode, est le sujet de toute la page, bien que ce contexte puisse être découpé en plusieurs contextes.

4.2. Classification des pages web par leurs profils

4.2.1. Généralités

Pour indexer les documents web, trois types d'information peuvent être utilisées :

– Le contenu lui-même des pages web : c'est-à-dire l'ensemble du code source de la page, le texte, les balises, les liens hypertextes, les liens vers les images ou d'autres ressources multimédias, la taille des fichiers, etc.
– Le graphe créé par les liens hypertextes reliant les pages les unes aux autres.
– Les données provenant de l'usage comme les fichiers de log, les "cookies", etc.

Cette classification est proposée par la communauté du « web mining » [KOSA00].
Il existe plusieurs approches pour aider l'utilisateur à naviguer sur le Web mais aucune ne prend en considération la notion de profil syntaxique des documents. Pourtant ces profils permettent d'identifier les types de données qu'ils contiennent. Les balisages utilisés dans les documents écrits par exemple en HTML, fournissent ces types de données.
HTML définit un ensemble de balises de base. On cite les balises de structure, puis celles qui permettent d'agencer et de composer du texte. L'autre catégorie de balises est celle qui permet de mettre en place des hyperliens. Une page Web peut être définie par un ensemble de caractéristiques (domaine du site, structure (frames, etc.), liens internes, liens externes, quantité et poids des images intégrées, rapport balise/contenu, …)
On part de l'idée qu'une page HTML peut être intéressante par sa forme descriptive et par son aspect. Celle-ci est intéressante si elle contient des liens vers le site lui même, des liens externes vers d'autres serveurs. Une page Web peut contenir des formulaires ce qui permet de comprendre qu'il s'agit d'une interface de saisie.
Il est aussi important de signaler que le poids d'une page est un élément très significatif car il peut permettre de déduire l'importance du contenu de la page

quantitativement. La présence d'images dans une page est un élément qui permet aussi de dégager une idée sur la dimension esthétique de la page.
Les documents sur le Web sont hétérogènes (sites commerciaux, pages personnelles, livres, articles, annuaires), ne possèdent aucune véritable structure.
Les contenus des sites peuvent varier d'un site à un autre par rapport aux objectifs de chaque site.
Papy F. et Bounai N. [PAPY03] proposent une approche fondée sur la classification de pages. Ils prennent en considération les balisages utilisés dans les pages Web pour élaborer des profils des pages Web. Cette approche est fondée sur les caractéristiques de pages HTML. Cette catégorisation permet alors :

– d'améliorer les navigations en réduisant l'espace de recherche en montrant seulement les pages pertinentes par rapport aux souhaits de l'usager.
– d'éviter la situation de surcharge cognitive à laquelle l'usager est souvent confronté au fil de ses lectures.
– de signaler à l'usager les types de pages auxquels aboutit sa requête.
– de donner des possibilités à l'usager de filtrer et de choisir les types de pages qu'il désire consulter.

Ils distinguent trois catégories de sites Web par rapport à leurs contenus :

– Les sites textuels privilégient les contenus textuels avec plusieurs liens internes et des liens externes car leur objectif est de diffuser les informations auprès des utilisateurs (les sites institutionnels, bibliothèques, universitaires, entreprises). Dans ceux-ci, les images ou les illustrations offrent des informations complémentaires et n'interviennent le plus souvent qu'à un deuxième niveau de recherche.
– Les sites visuels : privilégient les contenus visuels (images, graphiques d'illustration, etc.). Ainsi, ils intègrent souvent des formulaires (champs de saisies), par exemple les sites commerciaux, publicitaires, commerces électroniques, musées. L'image joue un rôle important, elle participe à l'attractivité du site et pour les commerciaux, elle est une valeur ajoutée indispensable. Pour les sites "plus techniques", l'image a une fonction différente. Elle permet à l'utilisateur de mettre rapidement ses attentes en correspondance avec l'information présentée. Dans ces sites, les textes offrent des informations complémentaires et n'interviennent qu'à un deuxième niveau de recherche.
– Les sites portails (annuaires) : privilégient plutôt les liens externes.

Pour établir une catégorisation de classification automatique des pages, ils se sont appuyés sur les travaux d'Alain Lelu ([LELU99], [BALPE96]) en utilisant l'algorithme de K-means axiales.
Une fois la méthode de K-means appliquée sur leur corpus, cinq types de pages ont été distingués automatiquement, et leur degré de typicité visualisé par une échelle à trois degrés (*, **, ***). En effet, ces cinq catégories constituent des pôles flous, plus que des classes bien distinctes :

– Page informative textuelle : Le contenu de la page est un texte.
– Page informative avec texte illustré : Le contenu de la page est une illustration visuelle, ce peut être des images, des figures, des boutons, etc.
– Page carrefour interne au site : le contenu de la page est un ensemble de liens internes au site.
– Page carrefour externe au site : le contenu de la page est un ensemble des liens externes au site.

– Page interface à la saisie : le contenu de la page est un ensemble de champs de saisie.

4.2.2. Notre contribution

Nous nous sommes inspirés de ces travaux pour construire nos classes de formes littéraires de contextes appelants et appelés, nous en avons retenu certaines et en avons rajouté d'autres spécifiques au domaine des biographies de personnages célèbres. Nous avons opté pour les classes suivantes :

– Classe sommaire : Le contenu du contexte est un résumé qui comporte les titres des parties des sites, c'est la même chose que la page carrefour interne. On les reconnaîtra principalement grâce à l'adjacence des liens.

– Classe illustration graphique : Le contenu du contexte est une illustration graphique par une image. On les reconnaîtra principalement grâce à la présence d'images de taille importante dans le contexte.

– Classe récit : Le contenu du contextes est en majorité du texte, on les reconnaîtra principalement grâce à la présence de texte en grande quantité dans le contexte.

– Classe citation : Le contenu du contexte est un texte qui fait référence directe à une oeuvre dans sa totalité ou en partie. On les reconnaîtra principalement grâce à la présence de texte en quantité moyenne et sans liens hypertextes.

– Classe liste : Le contenu du contexte est une suite d'articles inscrits les uns à la suite des autres. On les reconnaîtra principalement grâce à la présence des puces ou numéros aux débuts des articles.

4.3. Paramètres

En partant des caractéristiques citées auparavant et en observant une page Web sous ces deux angles, il est possible d'établir le profil d'un contexte appelant ou appelé en constituant un vecteur d'informations.

Le profil est construit par une analyse et un traitement statistique de balises *HTML*. Les données les plus significatives obtenues à partir de notre échantillon de contextes sont :

nbHref : nombre de liens, *nbImg* : nombre d'images, *TGimg* : taille de la plus grande image, *SMoyImg* : surface moyenne des images, *nbMot* : nombre de mots hors balise, *nbLEH* : nombre de lignes entre balises « a href », *nbLigne* : nombre de lignes hors balise, *nbBListe* : nombre de balises qui définissent des listes et/ou listes avec puces et/ou les énumérations, *nbBPg* : nombre des balises qui définissent les paragraphes, *nbBSLigne* : nombre de balises de saut de lignes, *cit* : prend 1 si des mots tels que « citation » figurent en balise méta name et 0 sinon, *def* : prend 1 si des mots tels que « définition » figurent en balise méta name et 0 sinon, *desc* : prend 1 si des mots tels que « description » figurent en balise meta-name » et 0 sinon, *sommaire* : prend 1 si des mots tels que « sommaire, résumé » figurent en balise meta-name et 0 sinon.

L'agent Web recueille les indicateurs quantitatifs, et les stocke sous forme d'une matrice (cf. tableau 1), chaque ligne correspond à un contexte, appelant ou appelé, et chaque colonne correspond à l'un des paramètres cités précédemment.

nbHref	nbImg	TGimg	SMoyImg	nbMot	nbLEH	nbLigne	nbBLiset	nbBPg	nbBSLigne	cit	Def	Desc	Sommaire
10	1	4628	4628	2770	23	239	40	47	0	0	0	0	0
9	2	0	0	308	0	40	0	0	0	0	0	0	0

Tableau 1. **Deux lignes de la matrice documents / paramètres**

4.4. Découpage de la base de données

Pour la phase d'expérimentation, nous avons choisi 1029 contextes parmi les contextes appelants de liens hypertextes et des contextes appelés par les liens hypertextes de notre corpus. Ensuite nous avons annoté ces contextes manuellement par leurs formes littéraires.
A partir de cet ensemble de contextes, nous avons tiré au hasard 852 contextes pour la base d'apprentissage et ce qui reste (177 contextes) sera pour la base de test.
Le tableau 2 est un récapitulatif des effectifs des formes littéraires dans les deux bases.

	Citation	Illustration	Liste	Sommaire	Récit
Base d'apprentissage	376	13	59	130	274
Base de test	80	3	14	18	62
% de classes dans les 2 bases	44,3	1,6	7,1	14,4	32,6

Tableau 2 - **Effectifs de forme littéraire dans les bases**

La classe citation est fortement représentée du fait du domaine d'application de biographies de personnages célèbres.
Ensuite nous avons mené une expérience de classification supervisée avec les treillis de Galois.

4.5. Classification supervisée avec les treillis de Galois

L'analyse formelle des concepts (AFC) [GAN99] offre un cadre théorique aux applications nombreuse et reconnues. Elle permet de représenter des données définies par une relation binaire entre deux ensembles, représentation encore appelée treillis de Galois [MEP02].

4.5.1. Discrétisation et Résultats

Notons que cette phase est un point très important pour l'efficacité du treillis de Galois. Il existe plusieurs méthodes permettant de faire une telle discrétisation. Nous avons d'abord appliqué la méthode de discrétisation que nous décrivons dans [HAJ03] et nous avons obtenu un ensemble d'environ mille cent intervalles, ce qui fait un très grand nombre d'attributs. En conséquence la complexité en temps de la construction du treillis de Galois est très grande car elle est exponentielle en fonction du nombre d'attributs.

Nous avons alors choisi une autre méthode de discrétisation. Nous avons défini pour chaque paramètre quantitatif cité auparavant quatre intervalles, le premier correspond à des valeurs très petites du paramètre, le deuxième à des valeurs petites, le troisième à des valeurs grandes, le quatrième à des valeurs très grandes. Cela nous donne une quarantaine d'intervalles, ce qui fait que la complexité en temps de la construction du treillis de Galois est assez raisonnable.
Pour passer de la matrice précédente « contextes de la base d'apprentissage »/ « attributs quantitatifs » (tableau 1) au tableau binaire « contextes de la base d'apprentissage »/ « attributs qualitatifs », qui sera utilisé comme entrée pour l'algorithme de construction du treillis de Galois, nous avons procédé de la façon suivante :
Dans ce tableau, les premiers attributs qualitatifs de chaque contexte sont obtenus par échantillonnage de chaque valeur des paramètres du contexte (§4.3) dans les quatre intervalles qui lui sont définis. A ces attributs s'ajoutent cinq attributs binaires dont chacun correspond à une de nos classes et prend la valeur 1 si le contexte est de la classe qui correspond à l'attribut et 0 sinon.
Les contextes de la base de test sont représentés de la même manière mise à part les cinq derniers attributs qui sont tous des zéro. Le problème de classification d'un contexte de la base de test revient alors à lui donner un attribut de classe.
Nous avons utilisé les deux techniques de reconnaissance se basant sur le treillis de Galois que nous avons déjà utilisées dans [HAJ03] : "Global Validation" et "Local Validation".
Par application de la méthode "Global Validation" sur les 177 contextes de la base de test, nous avons pu classer 108 contextes et ils sont tous correctement classés. Par application de la méthode "Local Validation" sur l'ensemble de test, nous avons pu classer 154 contextes dont 139 sont correctement classés. Le tableau 3 récapitule les résultats obtenus avec les treillis de Galois.

		Total	Citation	Illustration	Liste	Sommaire	Récit
	Effectifs	177	80	3	14	18	62
VG	Classés	108	57	1	3	12	35
	Correctement classés	108	57	1	3	12	35
VL	Classés	154	70	2	8	18	56
	Correctement Classés	139	67	2	6	13	51

Tableau 3 - **Résultats obtenus avec les treillis de Galois**

La même expérimentation a été réalisée avec d'autres outils de classification supervisée (k-ppv, réseaux de neurones, arbres de décisions) avec des résultats moins bons [HAJ06].

5. Conclusion et perspectives

Cette étude visait à l'extraction de la sémantique de liens hypertextes natifs pour aider à la navigation et à la recherche d'informations sur le Web. Nous avons

proposé une méthode pour effectuer l'analyse sémantique des liens hypertextes, et nous avons montré la compatibilité de notre travail avec les formalismes RDF(S). L'analyse dans un premier temps se fait manuellement. Dans un souci d'automatisation, nous proposons une aide à cette analyse qui consiste en l'extraction par des outils de reconnaissance de formes, ici nous avons utilisé les treillis de Galois, des formes littéraires des contextes appelants et appelés. Nous avons discuté des délimitations formelles des contextes appelants et appelés, puis nous avons choisi une méthode permettant de les délimiter. Les résultats de classification obtenus avec les treillis de Galois montrent leur efficacité pour la classification de contextes appelants et appelés selon leurs formes littéraires. Parmi nos perspectives, nous envisageons une autre manière de délimiter les contextes en les limitant aux paragraphes contenant les liens. Concernant les mots clés décrivant les contextes, nous sommes en train de tester des approches d'extraction de mots clés sur les contextes comme le TFIDF [SALT88] et des approches exploitant les informations autour des liens entrant aux contextes.

6. Références bibliographiques

[BALPE96] Balpe J.P., Lelu A., Saleh I., Papy F., « Techniques avancées pour l'hypertexte », Editions Hermès, 1996.

[CHAR03] Charlet Jean, Bachimont Bruno et Troncy Raphaël (2003), Prié, Y., « Les ontologies pour le Web sémantique », Web sémantique, Rapport final - Action spécifique 32 CNRS/STIC.

[GAN99] Ganter B. & Wille R. (1999). "Formal concept analysis, Mathematical foundations". Springer Verlag, Berlin.

[HAJ03] Al-hajj M., Bertet K., Gay J., and Ogier J.-M.. "Using the Concept Lattice for Graphic Understanding". In the proceedings of the Fifth IAPR International Workshop on Graphics Recognition (GREC'2003) , pages 329-340, Barcelona, Spain July 2003.

[HAJ06] AL-HAJJ M., VERLEY G., CARDOT H., « Une approche de caractérisation des contextes appelants et appelés des liens hypertextes ». XIIIème Rencontres de la Société Francophone de Classification SFC'06.

[KOSA00] Kosala R., Blockeel H., "Web Mining Research: A Survey", SIGKDD Explorations, vol. 2 (1) 2000, p. 1-15.

[LELU99] Lelu A., Hallab M., Rhissassi H., Papy F., Bouyahi S., Bouhaï N., He H., Qi C., Saleh I., « Projet NeuroWeb : un moteur de recherche multilingue et cartographique », 5e conf. Int. H2PTM'99, Paris, France, septembre 1999.

[MEP02] Mephu Nguifo et Njiwoua, 2002, « Treillis de concepts et classification supervisée : un état de l'art ». CRIL rapport de recherche.

[PAPY03] Papy F., Bounai N., « Navigation et recherche par catégorisation floue des pages HTML », Actes des JET'2003, 2003.

[RDF] http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

[RDFS] http://www.w3.org/TR/rdf-schema/

[SALT88] Salton G. and Buckley C., "Term weighting approaches in automatic text retrieval". Inf. Process. Manage. 24(5): 513-523, 1988.

[VAN00] Vandendorpe C., « Du papyrus à l'hypertexte : essai sur les mutations du texte et de la lecture », Ed. la découverte, Paris, 2000, p. 113-138.

[VIG01] VIGNAUX G., « L'hypertexte. Qu'est-ce que l'hypertexte. Origines et histoire », Laboratoire Communication et Politique, CNRS-UPR 36 (juin 2001).