Constitution et utilisation d'une terminologie en vue de l'extraction de la sémantique des liens hypertextes sur le web

Gilles Verley, Moustafa Al-Hajj, Hubert Cardot

Sémantique des liens hypertextes

Gilles Verley*, Moustafa Al-Hajj*, 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

Abstract :
The authors, who publish knowledge on the Web related to readable electronic documents on a screen, use the technology of hypertext links or the hypertext writing approach for making their sites more interesting and more attractive, and also to enrich it by information coming from other Web sites. However, hypertext links are not without posing problems for readers and search engines like disorientation and cognitive overload for readers and publicities links for search engines. We are interested in semantics of hypertext links, in terms of extraction and exploitation with the aim of facilitating the search of knowledge on the Web. In this article, we propose an original methodology for the semantic extraction of hypertext. Firstly, we show how we constitute a corpus of documents on the Web, which will be our data set. Then we propose a method of semantic analysis of hypertext links. This method consists in making the semantic analysis of calling context of link and context called by link, and explicit in a formal way the semantic relation between the two contexts.

Keywords: semantic web, corpus making, semantic of hypertexts links, formalism rdf(s), ontology.

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.

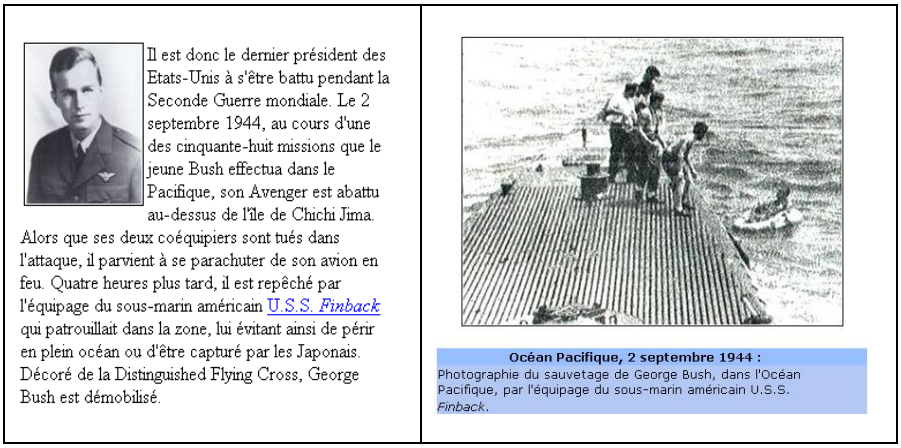

Figure 1 – **à gauche : contexte appelant du lien « U.S.S. Finback » ; à droite : contexte appelé par le lien**

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. 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).

5. 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).