Mise en œuvre d'un modèle formel pour l'extraction manuelle ou automatique de la sémantique de liens hypertextes

Verley Gilles , Al Hajj Moustafa

IUT – Université François Rabelais Tours
Département Information et communication
Rue du pont-volant -37000 Tours

gilles.verley@univtours, moustafa.ALHAJJ@telecom-bretagne.eu

Section de rattachement : 27
Secteur : Tertiaire

RÉSUMÉ Dans cet article, on présente une expérience relative au web sémantique en cours de réalisation avec des étudiants en sciences de l'information et de la communication dont l'objectif est triple. En premier lieu, vérifier la pertinence d'une méthode d'explicitation par l'homme de la sémantique de liens hypertextes selon un modèle formel original utilisant les ontologies. Deuxièmement, montrer l'intérêt qu'offrent les possibilités d'utilisation pour les internautes (lecteurs comme auteurs) des données sémantiques formalisées selon ce modèle en termes de navigation et de recherche. Troisièmement, constituer une base de données en vue de tester des algorithmes d'apprentissage capables d'effectuer totalement ou partiellement cette formalisation. Ce travail original se situe complètement dans les objectifs du web sémantique tels qu'ils sont définis par Tim Berners Lee et le W3C afin qu'émerge enfin le « web sémantique ».

MOTS-CLÉS : web sémantique, liens sémantiques, ontologies.

1. Introduction

Les liens hypertextes sont porteurs d'informations sémantiques qui, si elles étaient complètement formalisées, seraient exploitables par des programmes pour améliorer la navigation et la recherche d'information, mais cela n'est pas le cas actuellement. Cela est dû (1) à l'utilisation massive du langage HTML sur le web dans lequel la possibilité d'expliciter formellement la sémantique des liens est très limitée, (2) à la faible utilisation du langage XML et de ses dérivés notamment les liens XLink qui, depuis leur apparition, permettent une certaine explicitation formelle de la sémantique des liens notamment a travers les attributs sémantiques et (3) à la surcharge de travail que

demanderait cette explicitation formelle de la part des auteurs. L'utilisation d'informations sémantiques portées par les liens peut être au service des lecteurs à la fois pour guider leur navigation, par exemple, en distinguant les différents types de liens présents dans une page ou en proposant des liens non prévus par les auteurs mais néanmoins pertinents par rapport au contexte de lecture, mais également pour la recherche d'informations par les moteurs de recherche qui pourraient fournir des réponses plus exactes aux requêtes, c'est-à-dire comportant moins de silence et moins de bruit. En outre, cette formalisation de la sémantique des liens devrait faciliter l'érnergence du web sémantique parce que nous faisons l'hypothèse qu'elle peut être faite simplement par les internautes eux-mêmes ou, à terme, par des programmes.

Dans cet esprit, nous présentons ici une expérience relative au web sémantique en cours de réalisation avec des étudiants en sciences de l'information et de la communication dont l'objectif est triple :
- vérifier la pertinence d'une méthode d'explicitation par l'homme de la sémantique de liens hypertextes selon un modèle formel,
- montrer l'intérêt qu'offrent les possibilités d'utilisation pour les internautes (lecteurs comme auteurs) des données sémantiques formalisées selon ce modèle en termes de navigation et de recherche,
- constituer une base de données en vue de tester des algorithmes d'apprentissage capables d'effectuer totalement ou partiellement cette formalisation.

Dans une première partie, nous détaillerons le modèle formel de représentation de la sémantique des liens et la méthode d'explicitation de cette sémantique par l'homme. Ensuite, nous présenterons l'expérimentation en cours et son état d'avancement. Enfin, nous présenterons ses développements futurs en termes d'utilisation de la sémantique formelle extraite et des possibilités d'automatisation de cette formalisation.

2. Le modèle de représentation formelle des liens et la méthode d'explicitation par l'homme

Les auteurs posent des liens hypertextes dans leurs propos pour satisfaire différents besoins tels que (une typologie plus exhaustive est présentée en infra) :
- la spécialisation du propos de l'auteur, tel un lien situé dans un sommaire vers une partie détaillée,
- l'illustration du propos de l'auteur, tel un lien vers une photo,
- l'explicitation d'un terme, tel un lien vers une définition,
- le renforcement du propos de l'auteur, tel un lien vers un document d'un autre auteur partageant les même idées,
- la mise en situation du propos de l'auteur, tel un lien vers un exemple, etc.

Ces liens posés volontairement par les auteurs, nous les appelons *liens natifs* (Verley 2000) et nous nous intéressons spécifiquement à eux car nous faisons l'hypothèse

raisonnable qu'ils ont un minimum de sens, de par la volonté humaine dont ils sont le produit. Nous mettons de côté (pour autant que l'on puisse les identifier) les liens calculés par des automates car c'est justement l'objet du web sémantique et des travaux comme celui que nous présentons ici de permettre, à terme, de créer des liens calculés pertinents car fondés sur leurs sémantiques. Ne mettons donc pas la charrue avant les bœufs !

D'une manière générale, on considère qu'un lien relie deux nœuds constitués par la page où se situe l'ancre et la page cible. C'est un point de vue technique qui ne prend pas en compte l'aspect sémantique des choses. Nous préférons substituer au concept de nœud le concept de contexte. Ainsi, nous appelons *contexte appelant d'un lien*, l'ensemble minimal de multimedia (textes, sons, graphiques et images fixes on animées) autour de l'ancre du lien qui fournit une information suffisante pour comprendre le rapport avec la cible. De même, nous appelons *contexte appelé par le lien*, l'ensemble minimal de multimedia qui suit la cible du lien et qui fournit une information suffisante pour comprendre le rapport du contexte appelant avec le contexte appelé du lien. Ainsi, l'ensemble constitué du contexte appelant du lien et du contexte appelé par le lien forme une unité d'information autosuffisante.

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 le contexte appelé par le lien peut être explicité par la phrase suivante ; le contexte appelé *explique* « la convention de Lomé IV » qui est un élément du contexte appelant. 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 la partie de sa page 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 de l'utilisation des liens hypertextes pour améliorer l'ergonomie de son site 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.

Soit donc un *lien natif* , ayant donc un sens, reliant un *contexte appelant* et un *contexte appelé* de telle manière que ces deux contextes puissent être lus l'un derrière l'autre sans que cela choque l'entendement. Quel type de lien relie ces deux contextes ? En d'autres termes, qu'est-ce qui justifie intellectuellement que le contexte appelé puisse être lu (ou regardé, ou entendu) directement après le contexte appelant ?

De nombreux auteurs se sont intéressés à cette question. (Géry 2002) répartit les liens en deux grandes catégories :
- le lien de référence (ou de citation) uni ou bi-directionnel est celui qui établit des relations non hiérarchiques entre un nœud (contexte appelant) référence et un nœud (contexte appelé) référencé. Ce type de lien permet de mieux documenter une information ou d'approfondir un sujet.
- le lien organisationnel, dit aussi structurel, hiérarchique, spécialisation/généralisation, de subsomption et de composition, concerne la structure hiérarchique d'un hypertexte construit sous forme d'arbre. Ce type de lien permet de relier un nœud fils à son père dans le graphe structurel.

Les liens de référence sont plus variés et plus intéressant sémantiquement que les liens structurels mais ils sont aussi plus rares sans doute parce qu'ils demandent plus de travail pour être créés par les auteurs. C'est l'enjeu majeur du web sémantique de développer ce type de liens. Nous présentons ci-dessous une liste quasi-exhaustive, provenant essentiellement de (Trigg 1983), des types de liens, c'est-à-dire de ce que peut faire un contexte appelé par rapport à un élément du contexte appelant :
le contexte appelé *argumente* un élément (concept) du contexte appelant
le contexte appelé *caractérise* un élément (concept) du contexte appelant
le contexte appelé *commente* un élément (concept) du contexte appelant
le contexte appelé *cite* un élément (concept) du contexte appelant
le contexte appelé *contextualise* un élément (concept) du contexte appelant
le contexte appelé *critique* un élément (concept) du contexte appelant
le contexte appelé *décrit* un élément (concept) du contexte appelant
le contexte appelé *définit* un élément (concept) du contexte appelant
le contexte appelé *détaille* un élément (concept) du contexte appelant
le contexte appelé *exemplifie* un élément (concept) du contexte appelant
le contexte appelé *explique* un élément (concept) du contexte appelant
le contexte appelé *généralise* un élément (concept) du contexte appelant
le contexte appelé *illustre* un élément (concept) du contexte appelant
le contexte appelé *justifie* un élément (concept) du contexte appelant
le contexte appelé *réécrit* un élément (concept) du contexte appelant
le contexte appelé *réfute* un élément (concept) du contexte appelant
le contexte appelé *résume* un élément (concept) du contexte appelant
le contexte appelé *soutient* un élément (concept) du contexte appelant
le contexte appelé *spécialise* un élément (concept) du contexte appelant

En général, les travaux qui s'intéressent aux liens hypertextes se limitent à la détermination du type de lien tel que nous l'avons présenté en supra, c'est-à-dire à ce que fait un contexte appelé par rapport à un élément (concept) du contexte appelant. Nous considérons que cette seule détermination ne rend pas complètement compte de la sémantique du lien. D'une part, il faut évidemment préciser quel est l'élément (concept) du contexte appelant qui est concerné par le lien. En général, cet élément est l'ancre du lien. Mais il parait essentiel de comprendre formellement ce que fait cet élément par

rapport au sujet principal du contexte appelant. Si on reprend l'exemple de la biographie de Mitterrand et du lien qui *explique* « la convention de Lomé IV », il est légitime de se poser la question de savoir ce que vient faire « la convention de Lomé IV» dans la biographie de Mitterrand et en quoi cette « convention de Lomé IV » est suffisamment importante pour justifier que l'auteur de la biographie ait voulu qu'on puisse en avoir une explication dans une page associée par un lien hypertexte. Dans ce cas précis, c'est parce que l'auteur est en train de dire, dans le contexte appelant, que Mitterrand *a réalisé (a participé, a signé, etc.)* « la convention de Lomé IV ». Le modèle de représentation de la sémantique d'un lien comprendra donc également ces informations dans une phrase formelle simple composée d'un sujet, d'un prédicat et d'un objet.

Nous avons maintenant toutes les clés pour présenter notre méthode d'explicitation de la sémantique de liens hypertextes . Elle consiste donc à :
- délimiter les contextes appelant et appelé
- caractériser le type de lien, c'est-à-dire ce que fait le contexte appelé par rapport à un élément du contexte appelant selon la taxonomie présentée en supra.
- synthétiser la sémantique du contexte appelant dans une phrase formelle simple composée d'un sujet, d'un verbe et d'un complément dont les éléments constituent une ontologie du domaine traité pour le Web sémantique (Charlet et al 2003).

Soit une biographie de J.J. Rousseau, on y trouve le paragraphe suivant :
« Rousseau devient alors célèbre et se retire à Montmorency. En 1761, il publie La Nouvelle Héloïse, un roman épistolaire puis, en 1762, *Du Contrat social* et *Émile*. Cette même année, le Parlement condamne *Émile* pour ses idées religieuses. Rousseau s'enfuit alors en Suisse. Ses ouvrages sont brûlés publiquement. Il commence la rédaction de ses Confessions en 1765 et rentre à Paris en 1770, après avoir séjourné à Londres. Rousseau y écrit les Rêveries du promeneur solitaire ».

En activant le lien « les Rêveries du promeneur solitaire », on aboutit à un extrait de la cinquième promenade de ces rêveries. Le contexte appelé est donc cet extrait. Il fournit un exemple (exemplifie) de ces mêmes rêveries écrites par Rousseau. Le contexte appelant, c'est-à-dire l'ensemble minimal de textes autour de l'ancre du lien qui fournit une information suffisante pour comprendre le rapport avec la cible, peut se limiter à la phrase « Rousseau y écrit les Rêveries du promeneur solitaire ». Cette phrase peut être immédiatement formalisée de la manière suivante :
Sujet : « Rousseau », instance de la catégorie « Personnes » de l'ontologie
Prédicat : « a écrit », sous-catégorie de la catégorie « a réalisé » de l'ontologie
Objet : « les Rêveries du promeneur solitaire », instance de la catégorie « Œuvres littéraires », elle-même sous-catégorie de la catégorie « Œuvres »

Pour être conforme à notre modèle, il faut ajouter à ces trois éléments, le type de lien : « exemplifie » (donne un exemple). En effet, la cible *exemplifie* « les Rêveries du promeneur solitaire »

3. Expérimentation en cours

Dans le cadre d'un module complémentaire intitulé « web sémantique » et proposé aux étudiants de deuxième année à l'IUT de Tours préparant le DUT information-communication dans les options communication des organisations, gestion de l'information et du document dans les organisations et journalisme, nous avons demandé aux étudiants de participer à une expérimentation. Chaque étudiant doit analyser les liens natifs présents dans des sites biographiques de leur choix selon le modèle et la méthode présentée en supra. Le produit de cette analyse est réalisé en ligne grâce à un formulaire de telle manière que l'ontologie du domaine (les biographies) se développe au fur et à mesure des besoins de chacun mais de façon à ce que l'on respecte un principe essentiel des ontologies (et thesauri), à savoir qu'il ne faut qu'un seul terme pour représenter un concept et réciproquement.

Le formulaire comprend huit champs dont seulement quatre correspondent réellement à la formalisation de la sémantique du lien, les quatre autres permettant la constitution de la base de données en vue des développements prévus ultérieurs.

L'ensemble des huit champs se retrouvent en italique dans le texte suivant :
L'auteur a posé dans un <<*contexte appelant*>> d'une page source identifiée par une <<*url source*>> un <<*lien*>> vers un contexte appelé d'une page cible identifiée par une <<*url cible*>>. Le contexte appelant est résumé dans la phrase formelle ; <<*sujet*>> <<*prédicat*>> <<*objet*>>. La pose du lien est justifiée parce que la cible <<*remplit un rôle*>> par rapport à l'objet de la phrase formelle qui résume le contexte appelant.

Parmi ces huit champs, trois ne demandent aucun travail de la part des analystes, il s'agit de l'<<*url source*>>, de l'<<*url cible*>> et du <<*lien*>> (plus précisément, le texte du lien) puisque ce sont déjà des éléments formels. Le texte du <<*contexte appelant*>> délimité par l'analyste servira comme base de test pour mesurer les performances de programmes ultérieurs écrits pour automatiser sa délimitation et son explicitation formelle selon notre modèle (sujet+prédicat+objet). A ces huit champs est ajouté, pour des raisons pédagogiques, le nom de l'étudiant qui a fait l'analyse du lien. D'autre part, des formulaires sont à la disposition des analystes pour enrichir l'ontologie au fur et à mesure de leurs besoins. Il est à noter que toutes les données du modèle formel présenté ici sont représentées dans les formalismes du W3C, à savoir XML, RDF, RDF'S, OWL.

La formation des étudiants à la méthode et à l'interface a duré une heure. Actuellement, 300 liens provenant d'une vingtaine de sites biographiques ont déjà été analysés selon notre méthode et notre modèle sans problèmes cognitifs particuliers. L'ontologie comprend aujourd'hui environ 200 termes pour les champs <<*sujet*>> et <<*objet*>>, une cinquantaine pour le champ <<*prédicat*>> et une vingtaine pour le champ <<*remplit un rôle*>>. L'objectif est de dépasser le millier de liens avant la fin de l'année universitaire afin d'obtenir une « masse critique » permettant :
- de montrer l'intérêt des possibilités d'utilisation des données ainsi modélisées,

- de tester des programmes d'automatisation des opérations de formalisation

4. Utilisations et automatisations

En supposant que suffisamment de liens natifs issus de sites biographiques aient été analysés selon notre méthode, nous montrer maintenant quelques possibilités d'utilisation de leurs sémantiques ainsi formalisées.

Reprenons l'exemple du lien de la biographie de Mitterrand qui explique la « convention de Lomé IV ». La formalisation de la sémantique de ce lien selon notre modèle est :
« Mitterrand » *a participé à* «la convention de Lomé IV» qui *est expliquée* » dans la cible.

D'après l'ontologie, on sait que :
- « la convention de Lomé IV» est une instance de la catégorie « conventions économiques », elle-même sous-catégorie de « accords internationaux»
- « Mitterrand » est une instance de la catégorie « personnes ».

En plus de ce lien posé par l'auteur dans la biographie de Mitterrand, il est alors facile de proposer des liens calculés (issus du même site ou d'autres sites) vers :
- les documents *expliquant* les autres « accords internationaux» auxquels « Mitterrand » *a participé*,
- les biographies des « personnes » *ayant participé à* «la convention de Lomé IV»,
- les documents *expliquant* les « conventions économiques », etc.

Ces liens calculés peuvent être proposés au lecteur lorsque celui-ci survole un lien natif ou dans un cadre réservé à cet usage dans la fenêtre du navigateur. Ce sera à l'expérimentation de déterminer les meilleures solutions ergonomiques. Il est remarquable de noter que cet enrichissement se fait totalement à l'insu de l'auteur de la page « courante » en train d'être consultée. Sur le plan technique, les principaux programmes permettant ces nouvelles fonctionnalités ont déjà été écrits en utilisant la technologie AJAX et attendent d'être utilisés dans le cadre de cette expérimentation.

Une autre utilisation des données sémantiques issues des liens natifs et présentes dans la base de connaissances est l'inférence obtenue à partir de règles générales applicables à l'ontologie du domaine traité. Ainsi si on sait, par deux propositions distinctes, que (1) X est frère de Y et (2) Z est père de X, on peut inférer (3) Z est père de Y, (4) Y est fils de Z, etc. Il existe d'autres possibilités inférences moins triviales.

Un dernier point concerne les possibilités d'automatisation totale ou partielle de la formalisation des liens hypertextes selon notre modèle. Actuellement, cette formalisation est entièrement effectuée manuellement mais il est peu raisonnable d'envisager qu'elle puisse être effectuée spontanément par les auteurs ou les lecteurs à cause de la relative complexité du modèle, complexité dont la contrepartie est sa

richesse. Par contre, nous avons envisagé de faciliter le travail des futurs analystes grâce à des programmes qui automatiseraient certaines parties du processus d'analyse et de saisie. Ces programmes nécessitent d'être validés sur une base de test obtenue manuellement dont la constitution est l'un des objectifs de l'expérimentation présentée ici. Certains de ces programmes sont présentés dans (Al-Hajj et al 2006).

5. Conclusion

A u stade actuel de l'expérimentation présentée dans cet article, nous pouvons faire les constatations suivantes :
- le modèle formel a été pensé dans l'objectif de représenter le plus complètement et le plus simplement possible la sémantique des liens natifs, il se résume finalement à quatre éléments dont les valeurs sont issues d'une ontologie du domaine traité,
- la méthode de formalisation « manuelle » de la sémantique des liens hypertextes présents dans les sites biographiques selon notre modèle ne pose pas de problèmes cognitifs particuliers chez des analystes sommairement formés,
- les possibilités théoriques d'utilisation des données formelles semblent prometteuses sur le papier et sont réalisables techniquement. Leur intérêt pratique reste encore à être vérifié expérimentalement,
- les données obtenues dans le cadre de cette expérimentation constituent une base de test utilisable pour valider ou améliorer des programmes d'automatisation de la formalisation.

Bibliographie

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, Metz, France, 2006, p. 32-36.

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

Géry M., « Indexation et interrogation chemins de lecture en contexte pour la recherche d'informations structurées sur le web », Thèse de doctorat, 2002, Université Joseph Fourier, Grenoble, France.

Trigg R., « A network-based approach to text handling for the on-line scientific community », Thèse de doctorat, 1983, University of Maryland, USA.

Verley Gilles, Rousselle J.J., *« An evolved link-specification language for creating and sharing documents on the web»*, CRIS 2000 Current Research Information Systems, 25-27 Mai 2000, Helsinky.