Premi re partie

Notions de base et tat de l'art

1

Quelques aspects sur le web sémantique

1.1 Introduction

Le web sémantique est un concept développé par le W3C, il a pour but d'ajouter du sens au web et de le rendre plus intelligent. Dans ce chapitre, nous allons présenter quelques aspects du web sémantique qui nous serons nécessaires pour la suite. Nous commençons par définir ce que c'est que le web sémantique.

1.2 Web sémantique

Selon Tim Berners-Lee [5] : *le web sémantique n'est pas un web distinct mais bien un prolongement du web que l'on connaît, dans lequel, on attribue à l'information une signification clairement définie, ce qui permet aux ordinateurs et aux humains de travailler en plus étroite collaboration.*

Et dans une autre définition : *c'est un immense espace d'échanges de ressources entre machines permettant à des utilisateurs d'accéder à de grands volumes d'informations et à des services variés* [5].

La structure du web actuel est essentiellement syntaxique, son contenu est lisible par des humains et par des machines, mais il n'est compréhensible que pour les humains. L'idée est de changer la structure du web actuel, que l'on appelle web "présentable" ou "syntaxique", vers une autre structure, que l'on appelle web "intelligent" ou "compréhensible" par les machines, (voir la Figure 1.1). C'est de là qu'est née l'initiative du web sémantique : *un web qui parle aux machines* [4]. Pour ce faire, il est nécessaire de standardiser des langages et des outils adaptables à un maximum d'applications tout en conservant des propriétés permettant leur emploi dans les conditions d'échelle et de performance requises pour le web [6].

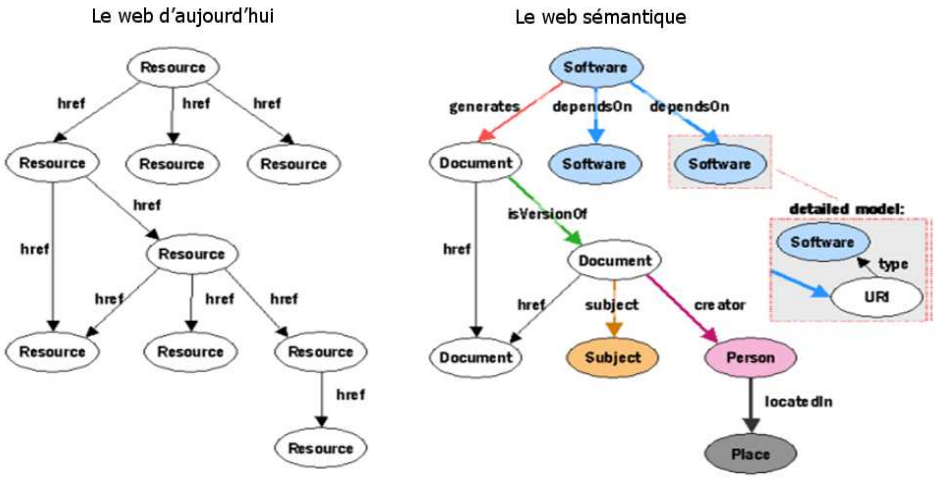

Fig. 1.1: Web d'aujourd'hui et le web sémantique

1.3 Services web

L'intérêt des services web est de favoriser une architecture orientée services, intégrant des systèmes hétérogènes complexes, fortement distribués et permettant la coopération et de nouvelles formes de collaboration entre les applications distantes. Un de ses intérêts est donc de faciliter l'interconnexion entre ces différentes applications, indépendamment des plateformes et des langages de programmation utilisés. Les services web semblent être la solution de l'avenir pour implémenter les systèmes distribués, aujourd'hui, ces services sont distribués à large échelle sur Internet [4].

En général, un service web se concrétise par un agent, réalisé selon une technologie informatique donnée. Un demandeur (utilisateur) utilise ce service à l'aide d'un agent de requête, il rentre alors des Inputs et attend des Outputs (saisir des entrées précises et avoir des sorties correspondantes). Le fournisseur et le demandeur partagent une même sémantique du service web, tandis que l'agent et l'agent de requête partagent une même description du service pour coordonner les messages qu'ils échangent (voir la Figure 1.2) [7] .

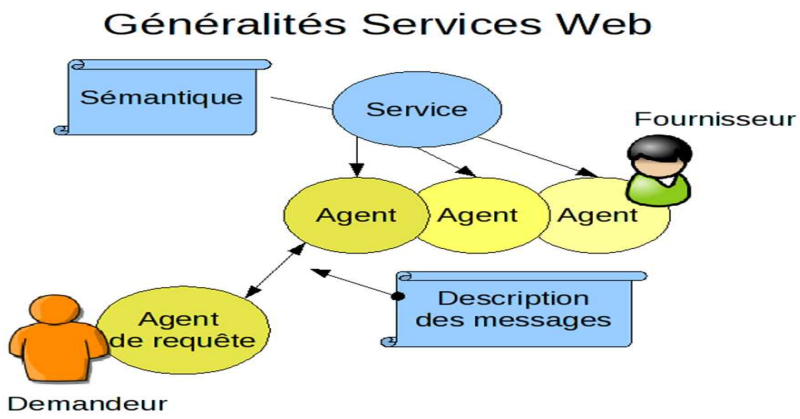

Fig. 1.2: Description du service web

1.4 Systèmes de médiation

Un système de médiation est un système intermédiaire, c'est une interface entre l'utilisateur et les services du web d'un domaine donné. Il doit donner l'impression à l'utilisateur qu'il n'utilise qu'un unique système alors que la satisfaction de sa demande peut exiger de composer plusieurs services.

Parmi les différentes grandes catégories d'applications de ces systèmes de médiation, on peut citer les applications de recherche d'information, celles d'aide à la décision en ligne et celles, de manière plus générale, de gestion de connaissances au sens large [1].

A titre d'exemple, on peut donner l'illustration du premier type d'applications. Supposons qu'un utilisateur pose la requête suivante : quels sont les maladies cancéreuses traitées à l'hôpital de Tlemcen ? lesquelles sont curables ?
Supposons l'existence de deux sources d'information. La première, Internet Medical Data Base, utilise un SGBD relationnel et contient une liste de maladies, précisant pour chacune le type, la partie du corps atteinte par cette maladie et les symptômes. La seconde source d'information, peut utiliser des fichiers XML contenant, par maladie, les différents cas traités et, pour chaque cas, le nom du patient, son état, et l'adresse de l'hôpital.
La réponse à la requête devra être construite en interrogeant chacune d'elles et en combinant les résultats de l'interrogation de façon à offrir à l'utilisateur une réponse globale.

1.5 Ontologies

Considérées comme des éléments fondamentaux du web sémantique, les ontologies y sont utilisées pour déterminer les index conceptuels décrivant les ressources sur le web. Les ontologies peuvent être définies comme des spécifications d'un vocabulaire de représentation pour un domaine partagé du discours qui peut inclure des définitions de classes, des relations, de fonctions et d'autres objets [8]. Les ontologies sont

utilisées en général pour permettre aux machines de raisonner et d'interpréter des informations, ainsi que d'améliorer la pertinence des recherches. Les ontologies permettent aux humains et aux machines de partager les connaissances du domaine et de coopérer ensemble.

Le principe consiste à définir une interprétation commune d'une partie du monde réel, et modéliser les concepts et les relations entre concepts par des classes et des relations entre classes, exemple dans la Figure 1.3.

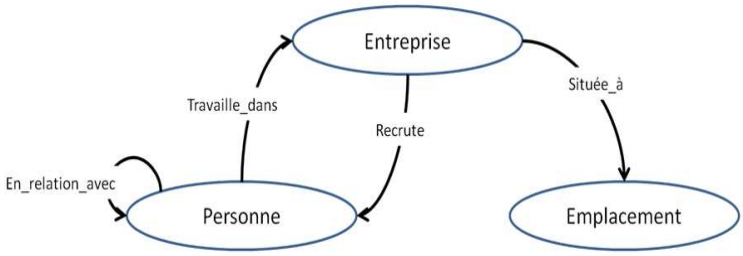

FIG. 1.3: Exemple d'ontologie

Les deux rôles des ontologies sont donc [9] :

– Définir une sémantique formelle pour l'information permettant son exploitation par un ordinateur (et donc des inférences) ;

– Définir une sémantique d'un domaine du monde réel, fondée sur un consensus et permettant de lier le contenu exploitable par la machine avec sa signification pour les humains.

Il est important de noter que le terme sémantique est employé dans un sens différent de celui utilisé dans le langage naturel. C'est-à-dire, le mot sémantique signifie ici "interprétable par les machines". Les machines devront être capables d'utiliser des ressources provenant de diverses sources [4].

1.6 RDF/RDFS/SPARQL

1. RDF

   RDF (Resource Description Framework) est un modèle conceptuel, associé à une

syntaxe, un cadre pour la description de ressources dont le but est de permettre à une communauté d'utilisateurs de partager les mêmes méta données pour des ressources partagées. Il a été conçu initialement par le W3C pour permettre de structurer l'information accessible sur le web et de l'indexer efficacement [10].

C'est une façon de représenter le monde sous la forme de déclarations simples, cette représentation est composée de 3 éléments :

- Un Sujet (ressource),
- Un Prédicat (propriété),
- Un Objet (valeur).

On s'en sert généralement pour la description de métadonnées externes à la ressource décrite. Puisqu'il ne s'agit pas d'un vocabulaire de métadonnées, on s'en sert alors avec des vocabulaires issus d'autres normes [11].

2. RDFS

RDFS (RDF Schema) permet de définir des vocabulaires RDF, principalement :

- Des classes,
- Des relations de sous-classe,
- Le typage des prédicats : domaine, co-domaine,
- Une relation de sous-propriété.

RDFS offre les moyens de définir un modèle (ou bien encore un schéma) de méta données qui permet de :

- donner du sens aux propriétés associées à une ressource ;
- formuler des contraintes sur les valeurs associées à une propriété afin de lui assurer aussi une signification.

Prenons l'exemple de [10], si l'on a une propriété qui représente un auteur, on peut exiger que les valeurs de cette propriété soient une référence à une personne (et non pas une voiture). On peut aussi vouloir restreindre quelles sont les propriétés s'appliquant à une ressource. Cela n'a probablement aucun sens d'autoriser une propriété "date de naissance" à être appliquée à un morceau de musique [10].

Les objets de ce langage (des graphes étiquetés) sont munis d'une sémantique formelle en théorie des modèles, ce qui permet de définir une relation de subsomption

entre les documents RDFS [12].

3. SPARQL

   SPARQL est un standard du W3C, un langage d'interrogation ontologique et un langage de requêtes adapté à la structure spécifique des graphes RDF. Il représente une amélioration de RDQL [13] (qui est le langage le plus utilisé pour interroger les ontologies implantées par RDF et il utilise une syntaxe similaire au SQL). SPARQL ajoute de nouvelles fonctionnalités pour construire les graphes de sortie.

1.7 Conclusion

Ces dernières années, le Web a rapidement évolué, car les données échangées sont devenues très hétérogènes. Les architectures orientées services, et en particulier les services Web, permettent l'accessibilité, la découverte et l'utilisation universelle de n'importe quelle application logicielle sur le Web en utilisant des normes ouvertes.
Le web sémantique propose des solutions formalisées pour améliorer la recherche sémantique des ressources. De nouvelles spécifications sont apparues pour améliorer cette recherche. L'un des objectifs du W3C est de permettre la recherche de ressources à la fois par des humains et par des machines. Le web sémantique s'est finalement orienté vers une solution d'indexation basée sur les ontologies [4].
Dans ce qui suit, nous parlerons de l'indexation d'une manière générale et de son rôle dans la recherche d'informations, on parlera également de ses techniques et de quelques travaux précédents sur l'indexation.

2

État de l'art sur l'indexation sémantique

2.1 Introduction

La recherche d'information (RI) suscite depuis fort longtemps l'attention de la communauté scientifique. La mise en œuvre de solutions capables d'améliorer la performance a toujours été primordiale. Des systèmes de recherche d'information (SRI) ont été conçus, leur objectif est de fournir aux utilisateurs les documents pertinents par rapport aux besoins qu'ils expriment. Les SRI utilisent des listes inversées qui rassemblent différents termes choisis pour représenter les contenus des documents et les liens vers ces documents. En complément, à chaque couple (terme, document) est associé un poids qui représente l'importance du terme dans un document. Cette conception est ainsi obtenue par un processus nommé indexation qui selon l'AFNOR est [14] : *l'opération qui consiste à décrire et à caractériser un document à l'aide de représentations des*

concepts contenus dans ce document, c'est-à-dire transcrire en langage documentaire les concepts après les avoir extraits du document par une analyse. [...] La finalité de l'indexation est de permettre une recherche efficace des informations présentes dans un fonds de documents et d'indiquer, sous une forme concise, la teneur d'un document. Lorsqu'une requête est soumise au système, les termes qu'elle contient sont mis en correspondance avec les termes d'indexation extraits des documents pour en déduire les documents à restituer à l'utilisateur. La phase d'indexation est donc une phase primordiale dans le processus de recherche. Dans la littérature, diverses méthodes et stratégies ont été proposées pour l'indexation [15].

2.2 Techniques d'indexation

L'indexation peut être faite d'une manière manuelle, automatique, de façon assistée (semi automatique) ou bien par concepts (indexation conceptuelle) [13].

L'indexation manuelle

Ce type d'indexation repose habituellement sur un jugement de signification plus ou moins intuitif, toujours lié à l'indexeur. Le travail à réaliser pour la mise au point d'une indexation est assez important : connaissance du contenu de l'information, choix des concepts à représenter et traduction de ces concepts en descripteurs. De plus, les mêmes notions peuvent être exprimées de manières très diverses. En raison de ces divers problèmes, des méthodes d'indexation automatique sont donc apparues [16].

L'indexation automatique

Cette indexation utilise des méthodes logicielles pour extraire et établir une liste ordonnée (un index), de tous les mots et leurs occurrences apparaissant dans les documents, et qui correspondent le mieux au contenu informationnel d'un document [16].

En construisant un index avec les mots extraits des textes intégraux, l'indexation automatique élargit considérablement le champ de la recherche et autorise des requêtes

plus souples. Ce type d'indexation suppose un certain nombre d'opérations [13] :

- Le filtrage des mots (analyse morphologique).
- La lemmatisation (Stemming) (analyse lexicale).
- La modélisation vectorielle.
- La pondération.

L'indexation automatique a de nombreux avantages, elle permet par exemple [13] :

- De limiter les choix parfois subjectifs de l'indexeur.
- D'alléger le travail requis par une indexation manuelle.
- D'éviter les incohérences des interprétations différentes entre plusieurs indexeurs.

L'indexation semi-automatique

Les systèmes actuels tentent de remplacer l'homme pour une importante part de son expertise et de lui épargner de nombreuses tâches ; mais, ils ne le remplacent pas complètement. L'indexation automatique suppose une intervention totale du système, ce qui est loin d'être le cas, car l'intervention humaine est toujours nécessaire, d'où l'indexation semi-automatique [17].

L'indexation conceptuelle

L'indexation conceptuelle consiste en toute explicitation symbolique de connaissances contenues dans les documents ou bien à leur propos, en permettant la recherche et la manipulation. Ceci dépasse la simple explicitation des concepts contenus dans les documents, mais recouvre également tout ajout de connaissances pouvant servir d'une manière ou d'une autre, par exemple, la position géographique d'une caméra au moment où un plan audiovisuel a été tourné, ou l'âge du réalisateur [17] [13].

L'annotation

Les notions d'annotation et d'indexation semblent équivalentes, néanmoins nous relèverons les différences suivantes :

- Indexer, c'est avant tout décrire un document pour le retrouver ;

– Annoter, c'est décrire l'interprétation du document par un lecteur, en vue de n'importe quelle tâche d'exploitation future de ce document.

– On indexe pour rechercher plus tard, on annote pour donner des traces de son interprétation, pour documenter la tâche que l'on est en train d'accomplir. Ces traces pourront alors être destinées à soi-même, ou partagées [17].

Dans le cas général, annoter un document, c'est : attacher à l'une de ses parties une description qui correspond à l'usage que l'on souhaitera en faire plus tard [17].

2.3 Indexation sémantique et ontologies

L'indexation sémantique via des ontologies s'appuie sur les technologies du web sémantique. Dans ce type d'approches, la connaissance du domaine (terminologique en particulier) est représentée sous forme d'ontologies, c'est-à-dire en particulier de concepts, d'instances de ces concepts et de relations [15].

2.4 Quelques travaux connexes à l'indexation sémantique

Plusieurs travaux ont contribué à l'avancement de la recherche dans le domaine de l'indexation et de la recherche d'informations. Nous pouvons citer :

– Baziz et al. [18] proposent une indexation utilisant le " noyau sémantique " d'un document. Il s'agit d'un ensemble de concepts pondérés suivant leur représentativité dans les documents et liés entre eux par des mesures de similarité. Cette structure dépend de la mesure de similarité considérée. Baziz et al. ont remarqué que la pondération des concepts est un point crucial, autant que le choix de ces concepts, pour les performances du système. L'idée de noyau sémantique permet de rendre graphiquement de façon claire à l'utilisateur les concepts dans un document et leurs liens.

– Seco et al. [19] pensent qu'une ontologie seule suffit à trouver le contenu informationnel des nœuds. Leurs thèse est qu'il est possible de retirer de la structure

de cette ontologie un sens au nombre d'hyponymes qu'a un concept : plus un concept a de descendants, plus il est spécialisé par d'autres concepts, moins il est lui-même caractéristique. Pareillement, les feuilles de la taxonomie ont une valeur informationnelle maximale.

- Song et al. [20] proposent un modèle de RI basé sur des ontologies de domaine, définies avec OWL. Les différentes ontologies de domaines sont intégrées pour former une ontologie unique. Les termes définis dans l'ontologie sont alors utilisés d'une part comme métadonnée pour annoter les contenus du web et d'autre part comme termes d'indexation de la collection.
- Gilles Hubert et al. [15] proposent un modèle de données dans le cadre d'une indexation à base d'une ontologie de référence. Cette structure de données permet en outre une mise à jour dynamique et en temps réel des résultats de l'indexation lors de la mise à jour de la collection de documents. Cette structure assure ainsi la cohérence permanente entre l'index, le corpus et l'ontologie de référence. L'avantage principal de ce modèle est qu'il n'est plus nécessaire de reconstruire l'index car il est à jour à tout moment. Ainsi, cette structure permet de mettre en place une indexation sémantique dynamique.

2.5 Conclusion

Cette partie traite l'indexation, ses techniques et quelques travaux précédents. Nous avons commencé par donner les différentes manière de mettre en place une indexation, qui peut se faire automatiquement, manuellement, d'une manière assistée ou par concepts. Nous avons également introduit l'indexation sémantique peut se base sur les ontologies et fait un tour d'horizon des travaux les plus significatifs et relatifs à l'indexation sémantique par ontologies. Le prochain chapitre sera consacré à la description de la démarche d'indexation de différentes sources d'information, ainsi qu'à la description des services web d'indexation qui auront pour but d'exploiter les résultats du processus d'indexation réalisé préalablement automatiquement ou manuellement, selon le type de la source.