Vers la composition de workflows pour le filtrage sémantique de nouvelles

Christophe Desclaux
Université Nice Sophia Antipolis
christophe@zouig.org

Simon Urli
I3S, CNRS
Université Nice Sophia Antipolis
urli@i3s.unice.fr

Mireille Blay-Fornarino
I3S, CNRS
Université Nice Sophia Antipolis
blay@polytech.unice.fr

Catherine Faron Zucker
I3S, CNRS
Université Nice Sophia Antipolis
faron@polytech.unice.fr

Abstract

//[je ne sais pas comment écrire abstract en francais avec LaTex]:CH

Le web se révèle aujourd'hui un merveilleux support de diffusion d'informations. Cependant, tandis que les sources se multiplient (flux rss, services web, ..), la quantité d'informations croît et il est difficile de les filtrer en fonction des centres d'intérêts des utilisateurs. Actuellement de nombreux outils qui exploitent les ontologies ou les thésaurus sont mis au point. Ils permettent d'annoter les informations, d'en déduire des critères et d'ensuite obtenir uniquement les informations pertinentes. La composition de ces outils constitue des workflows qui devraient encore s'enrichir grâce à l'apparition de nouvelles ontologies ciblées sur différents domaines et outils de lecture. Cependant la construction de telles chaînes logicielles n'est pas à la portée de tous.
Dans cet article nous montrons comment de tels workflows ont été construits et présentons nos perspectives en matière de construction automatique de ces workflows en fonction des besoins utilisateur. Ce travail s'appuie sur le projet ANR EMergence Yourcast qui vise à automatiser la diffusion des informations sur de grands écrans, et pour lequel la pertinence des informations diffusées est donc particulièrement pertinent.

Abstract

Internet is becoming today a wonderful medium for disseminating information. While the sources are multiplying (RSS, Web Services, ...), the amount of information is growing and it becomes difficult to filter them according to user interests. Many tools are currently developed that exploits ontologies or thesauri to annotate informations. They enable to query these annotations according to criteria to retrieve only the relevant information. The composition of these tools constitute workflows that should be enriched by the emergence of new ontologies modeling different domains and text analysis tools. However the composition of this tools-chain is not for everyone.
In this paper we show how these workflows are built and present our approach for automatically building workflows based on user needs. This work is supported by the ANR Emergence Yourcast project dedicated to automate the dissemination of information on large screens, and for which the relevance of the information published is particularly relevant.

1 Mots clé

Semantic Web, Product Line, Workflow

//[je ne sais pas comment formaliser les mots clé avec ce template]:CH

2 Introduction

Le web se révèle aujourd'hui un merveilleux support de diffusion d'informations. Tandis que les sources se multiplient (flux rss, services web, ..), la quantité des informations croît et il devient difficile de les filtrer en fonction des centres d'intérêts des utilisateurs [2]. Des outils qui exploitent des ontologies ou des thésaurus ont été mis au point qui permettent d'annoter les informations, d'en déduire des critères et d'ensuite obtenir uniquement les informations pertinentes en formulant des requêtes sémantiques sur les annotations.

Il devient aujourd'hui possible de construire à la fois des workflows mettant en jeu ces différents outils pour annoter les flux d'informations pour sélectionner les informations. Cependant la construction de ces workflows reste technique malgré les nouveaux supports logiciels tels que les mashup [4] [5]. En effet elle se base sur de nombreux critères qui doivent être pris en compte pour assembler les services et le fait que ces services soient hétéroclites ne permet pas une génération aisée des applications. De plus de nouvelles ontologies, sources, systèmes d'annotations apparaissent régulièrement tandis que le web se démocratise [1] [3]. Dans ce contexte, nous abordons dans cet article la production automatique de ces workflows à partir d'un ensemble de caractéristiques proposées à l'utilisateur.

Nous présentons en section 3 le contexte de notre travail que s'inscrit dans le cadre du projet ANR EMergence Yourcast qui vise à automatiser la diffusion des informations sur de grands écrans. Nous montrons au travers d'un cas d'étude les différents choix qui se posent à l'utilisateur et décrivons dans la section 4 les workflows mis en place pour répondre à ce cas particulier. Fort de cette expérience, nous proposons en section 5 de produire de tels workflows en utilisant un développement dirigé par les modèles et les feature models.

3 Contexte et besoins exprimés dans le projet YourCast

Dans le cadre du projet YourCast, nous visons à diffuser sur de grands écrans des informations en provenance de différentes sources en particulier celles issues du web. Or de tels systèmes exigent une adhérence forte aux attentes des utilisateurs et l'adéquation des informations avec les centres d'intérêts des personnes est essentielle à l'acceptation de tels systèmes.

//[Je prefere qu'on garde les sous titre ici cependant Catherine souhaite les enlever : à Mireille de trancher (ca évite selon moi l'effet bloc de texte chiant à lire)]:CH

Il existe aujourd'hui de nombreuses sources d'information disponibles grâce à l'utilisation de flux RSS. Le choix des sources peut être simplement lié au travail ou à l'emplacement géographique de l'écran d'information. Dans notre cas, nous avons choisi d'agréger le plus de sources d'informations possible pour couvrir tous les champs d'application des écrans d'accueil. Nous faisons donc appel dans notre application à une vingtaine de flux RSS sur de vastes sujets. Ceux-ci sont essentiellement des flux ayant déjà subi une étape de filtrage qui a permis de les classer selon leurs thématiques générales (technologies, international, médical, ...).

Beaucoup de sources d'informations sont agrégeables. Des critères de sélection sont alors pré-établis par les fournisseurs de contenus. Par exemple sur le site de news de Google[1] l'internaute peut accéder à des nouvelles liées à l'économie ou bien des news locales. Cependant il ne peut pas récupérer les nouvelles économiques liées par exemple à la ville de Marseille ou plus largement à la région PACA. Dans notre exploitation des flux RSS nous souhaitons fournir un filtrage multi-critères permettant un tri fin des nouvelles.

Pour capturer ces critères il existe actuellement différents systèmes tels que le service *Google*

[1] `http://news.google.fr`

Reader[2] ou bien celui *rssLounge*[3]. Cependant ceux-ci ne proposent pas de regroupement des flux ni de filtrage multi-critère de ceux-ci. Or le besoin exprimé dans le projet YourCast est celui d'un système intuitif qui permette de capturer simplement les exigences utilisateur. Nous avons pour cela développé un système d'aide au choix qui après captation en langage naturel des besoins les retranscrit sous forme d'entités nommées qui sont typées par notre système en fonction des éléments que nous avons déjà pu instancier dans la base de connaissances.

Le système est basé sur une annotation des informations la plus vaste possible. En effet le système doit pouvoir annoter des informations provenant de domaines totalement différents. Nous avons choisis d'utiliser des annotions basées sur la récupération d'entités nommées présentes sur un grand nombre de bases de données RDF. Nous utilisons alors des systèmes d'extraction d'information et d'annotation sémantique qui permettent d'ajouter des liaisons vers les entités nommées spécifiques.

4 Annotation et filtrage sémantiques de flux RSS

Pour répondre aux besoins exprimés dans le cadre du projet YourCast nous avons mis au point deux workflows, dont la figure 1 présente l'architecture générale: un workflow d'enrichissement sémantique des nouvelles (en bleu) et un workflow de filtrage sémantique des nouvelles. La scicion en deux workflows est essentielle car elle permet de faire travailler nos workflows de manière asynchrone.

//[je présente peut etre la notion de workflow asynchrone de manière mal-adroite?]:CH

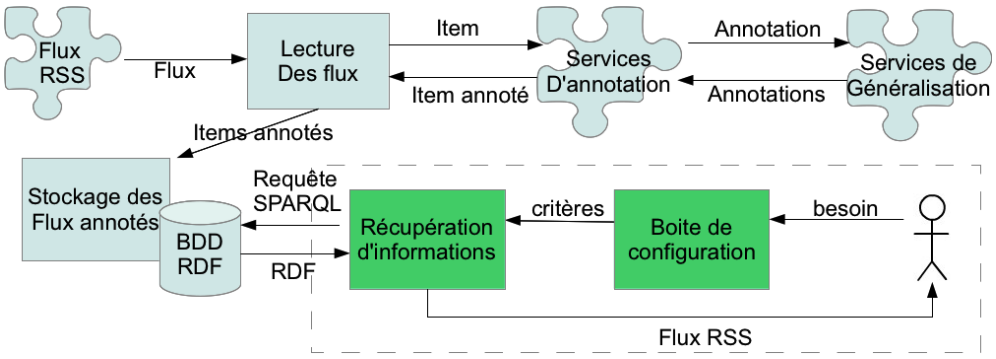

Figure 1: Workflows d'annotation et de filtrage sémantique de nouvelles

4.1 Workflow d'enrichissement sémantique des nouvelles

4.1.1 Lecture de flux RSS

Le workflow se base sur la récupération de flux RSS qui sont choisis par l'utilisateur et dont nous connaissons le thème et la langue, chaque flux RSS contient un lot de nouvelles les plus récentes publiées chez le fournisseur du flux. Le flux peut être mis à jour et doit être interrogé périodiquement pour récupérer les dernières nouvelles publiées.

[2] `http://www.google.fr/reader/`
[3] `http://rsslounge.aditu.de`

4.1.2 Annotation sémantique de nouvelles

Pour chaque nouvelle, nous utilisons des services d'annotation pour récupérer les entités nommées intéressantes la concernant:

- **OpenCalais**[4] permet d'identifier dans le texte les entités correspondant à des lieux français ainsi que des noms de personnalités en langue anglaise. gràce à l'API Java fourni par le projet OpenCalais, nous avons implémenté le code permettant de récupérer les annotations textuelles concernant le texte à annoter;
- **WikiMeta**[5] fonctionne sur le même principe qu'OpenCalais. Cependant le projet ne proposait pas au moment de nos recherche d'API permettant d'effectuer des requêtes et nous avons alors du implémenter l'API Java se connectant à leurs service web. Elle permet de récupérer directement des entités nommées pointant vers la base de connaissances DBpedia; il peut donc être utilisé seul en tant que service d'annotation principal.

Nous faisons appel à des services de généralisation qui travaillent en aval des services d'annotation pour ajouter des notions à une annotation :

- **DBpedia**[6] travaille sur les données provenant du service d'annotation WikiMeta et permet de les enrichir sémantiquement à l'aide de liens vers d'autres ressources liées à l'élément.
- **INSEE geo** [7] La base de données RDF de l'INSEE permet de récupérer les ensembles géographiques liés à une entité nommée géographique. Ce service d'annotation géographique s'utilise donc uniquement en aval d'un service d'annotation classique qui récupère des informations géographiques. Il renvoie des URI vers les entités géographiques liées à la nouvelle étudiée. Nous utilisons pour cela une requête SPARQL telle que celle ci-dessous qui permet de retrouver le nom et le type des subdivisions de la ville de Nice :

```
PREFIX geo: <http://rdf.insee.fr/geo/>
SELECT DISTINCT ?nom ?type WHERE {
 ?entite geo:nom ?nom
 ?entite rdf:type ?type
 ?ville geo:nom ?nomVille
     ?entite geo:subdivision* ?ville
     FILTER(regex(str(?nomVille), 'Nice', 'i'))
}
```

//[nous n'avons pas présenté la partie mémorisation, est-ce grave?]:CH

4.2 Workflow de filtrage des nouvelles

Le workflow de filtrage permet de créer des critères utilisateurs afin de récupérer les données intéressantes sur la base des annotations. Notre workflow va dans un premier temps transformer les besoins utilisateurs en critères sous forme d'une liste de couples (TypeElement,entiteDescriptive). Ces couples sont ensuite envoyés à l'élément de connection à la base de données qui effectue une requête SPARQL multicritères sur la base de donnée RDF pour

[4] http://www.opencalais.com
[5] http://www.wikimeta.com
[6] http://dbpedia.org
[7] `http://rdf.insee.fr/geo/`

récupérer les nouvelles à fournir à l'utilisateur. Enfin le système génère un flux RSS personnalisé pour répondre aux besoins utilisateur et le fourni au client. Ce flux RSS va pouvoir être mémorisé par le client pour récupérer en temps réel les nouvelles informations.

5 Vers la construction automatique de workflows

Notre objectif à terme est de construire une ligne de produits qui capturerait les différentes sources et services disponibles, les qualifierait et permettrait à un utilisateur final de construire ses propres workflows d'annotation en le guidant dans sa sélection de différents services.

Du point de vue de l'utilisateur final, l'utilisation d'une construction de workflows pourrait lui permettre, à partir d'un flux sélectionné, d'affiner la sélection des services d'annotations qu'il souhaite utiliser afin d'obtenir l'information la plus complète pour son utilisation.

A l'heure actuelle le workflow d'annotation interroge tous les services d'annotations disponibles, sans distinction de catégories, ce qui génère de nombreux appels de services dont la plupart sont inutiles. L'utilisation d'une construction automatique à partir des choix de l'utilisateur permettrait de restreindre les appels aux seuls services pertinents et donc d'accélérer les temps de traitements en réduisant la charge des services.

5.1 Lignes de produits de services

Les choix que va faire l'utilisateur vont avoir un impact sur trois concepts inhérents au workflow d'annotations : les *Flux d'information*, les *Services d'Annotation* et les *Services de Généralisation*. Chacun de ces concepts peut être vu comme une ligne de produits (services) indépendante, à partir de laquelle l'utilisateur choisira le service qui lui convient. Nous représentons ces lignes de produits par des modèles d'aspects ou *feature models* (FM).

La figure 2 montre ainsi l'aspect *Service d'Annotation* : les produits finaux correspondent aux feuilles de l'aspect *Produit*, les thèmes aux feuilles de l'aspect *Thème* et les langues à celles de l'aspect *Langue*. La sélection d'un aspect de langue ou de thème influe directement sur le choix d'un service final par le jeu des contraintes internes au FM.

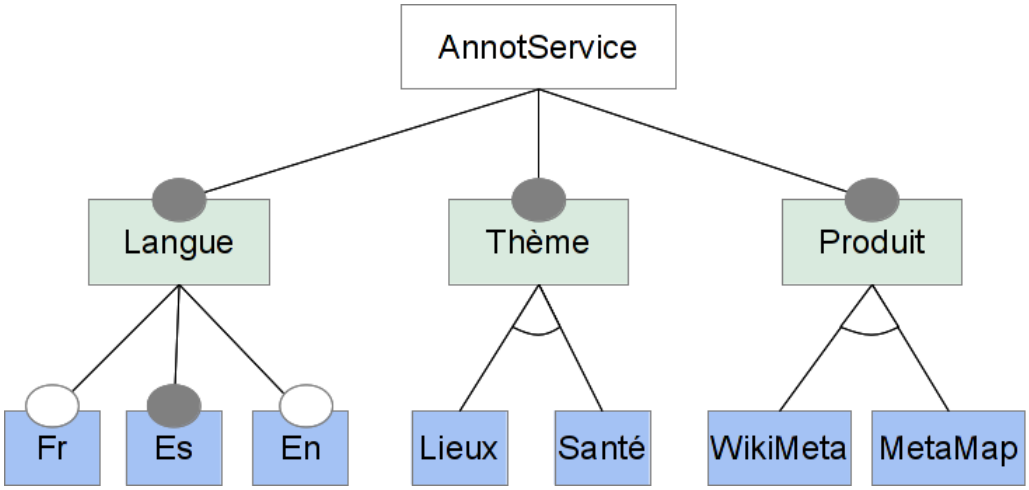

Figure 2: Workflow de traitements....

5.2 Modèle de mise en relation des LPL

Notre but est de fournir un outil permettant à un utilisateur final de sélectionner des services cohérents en fonction de ses choix. Cela implique que chaque choix que l'utilisateur fera dans un des FM doit avoir un impact sur les autres concepts du domaine. Nous exprimons donc des relations entre les différents concepts du domaine, relations qui s'expriment par le biais du *Thème* et de la *Langue*. En effet, les systèmes d'annotation sont spécialisés pour certains types d'informations, comme le service MetaMap [8] et ne traitent que certaines langues. Une relation exprime à la fois qu'un produit (i.e. un service) sera lié à un autre, mais aussi que sélectionner un produit influera sur la sélection d'un autre.

Par exemple, un *Flux* traitant de politique sera annoté par un ou plusieurs *Services d'annotation* pouvant annoter ce type d'information : les produits seront donc liés. Cependant ce *Flux* ne pourra pas être lié à un *Service d'annotation* parlant d'un thème différent ou traitant une langue différente : la sélection du flux influe directement sur la sélection des *Services d'annotation* à lier.

Pour cela nous définissons et appliquons des opérations de restrictions sur les relations entre les FMs, et ce en cascade. Pour reprendre notre exemple, la sélection d'un flux influe sur la sélection des annotations, en restreignant les services disponibles, ce qui aura également une influence sur les services de généralisation disponibles.

6 Conclusion

Dans cet article nous avons présenté le système d'annotation ZeOntologyNewsExtractor [9] dont nous présentons une preuve de concept sur `http://zone.zouig.org`. Nous avons explicité l'utilisation de l'ingénierie des connaissances pour enrichir sémantiquement des nouvelles. Puis nous avons présenté la nécessité de faire appel à l'ingénierie du logiciel et particulièrement la mise en place de workflows pour organiser les flux composant notre architecture. Enfin nous avons présenté une amélioration possible de notre approche utilisant la génération de lignes de produits logiciels.

//[il faut une ouverture que je ne trouve pas]:CH

References

[1] Jean-Noël Anderruthy. *Web 2.0: (r)évolutions et nouveaux services d'Internet*. 2007.

[2] Mokrane Bouzeghoub and Dimitre Kostadinov. Personnalisation de l'information: aperçu de l'état de l'art et définition d'un modèle flexible de profils. In *CORIA*, pages 201–218, 2005.

[3] Justus Bross, Matthias Quasthoff, Philipp Berger, Patrick Hennig, and Christoph Meinel. Mapping the Blogosphere with RSS-Feeds. In *2010 24th IEEE International Conference on Advanced Information Networking and Applications*, pages 453–460. IEEE, 2010.

[4] Ingbert R. Floyd, M. Cameron Jones, Dinesh Rathi, and Michael B. Twidale. Web mash-ups and patchwork prototyping: User-driven technological innovation with web 2.0 and open source software. In *HICSS*, page 86. IEEE Computer Society, 2007.

[5] N. Milanovic and M. Malek. Current solutions for Web service composition. *IEEE Internet Computing*, 8(6):51–59, November 2004.

//[il manque la biblio de simon sur les LDP]:CH

[8] `http://metamap.nlm.nih.gov/`
[9] `https://github.com/descl/ZONE`