Vers la construction 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

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 nécessaire 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 importante.

Mots-clés : Web sémantique, Ligne de produits logiciels, Workflow, Modèle de variabilité

Abstract

Internet is becoming today a wonderful medium to broadcast informations. While sources are multiplying (RSS, Web Services, ...), the amount of informations is growing and it becomes necessary 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 informations. 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 these tools-chains is not accessible 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 broadcasting of informations on large screens, and for which the relevance of informations published is particularly important.

Keywords: Semantic Web, Software Product Line, Workflow, Feature Model

1 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 essentiel de les filtrer en fonction des centres d'intérêts des utilisateurs [3]. 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 est aujourd'hui possible de construire des workflows mettant en jeu ces différents outils pour d'une part annoter les flux d'informations et pour d'autre part sélectionner les informations pertinentes en fonction de la cible de diffusion.

Cependant la construction de ces workflows reste technique malgré les nouveaux supports logiciels tels que les mashups [6] [7]. 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 [2] [4], ce qui occasionne des mises à jour régulières de ces workflows.

Dans ce contexte, nous décrivons dans cet article la construction de ces workflows, puis discutons leur production automatique à partir d'un ensemble de caractéristiques sélectionnées l'utilisateur au travers d'une approche basée sur les lignes de produits logiciels.

Ce travail s'inscrit dans le cadre du projet ANR Emergence YourCast qui vise à automatiser la diffusion des informations sur de grands écrans. Nous présentons ce contexte en section 1. Nous montrons dans la section 3 les choix d'architectures que nous avons faits pour mettre en place les workflows adaptés au projet. Forts de cette expérience, nous proposons en section 4 de produire de tels workflows en utilisant un développement dirigé par les modèles et des modèles de variabilité (*feature models*)[8]. La section 5 conclut cet article.

2 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.

Or de *nombreuses sources d'information* sont aujourd'hui disponibles sous la forme de flux RSS. Ceux-ci sont généralement classés selon leurs thématiques générales (technologies, international, médical, ...). Une première étude nous a conduit à nous intéresser à une vingtaine de flux RSS sur divers sujets.

La *sélection des informations* dans ces sources multiples repose alors sur la mise en place de critères et leur composition. Ainsi quelques critères de sélection sont 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 par exemple récupérer les nouvelles économiques concernant la région PACA.

Pour capturer ces critères il existe actuellement différents systèmes tels que le service *Google Reader*[2] ou bien celui *rssLounge*[3]. Cependant ceux-ci ne proposent pas de regroupement des flux ni de filtrage multi-critères de ceux-ci.

Les exigences du projet YourCast sont celles d'un système intuitif qui permet de capturer simplement les besoins utilisateur à la fois en terme de choix des sources d'informations et de critères de sélection.

Pour cela, nous avons choisi d'enrichir des flux d'information en nous basant sur l'extraction d'entités nommées [5] présentes dans un grand nombre de bases de données RDF. Nous utilisons

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

pour cela des systèmes d'extraction d'information et d'annotation sémantique qui permettent d'ajouter des liaisons vers les entités nommées spécifiques.

3 Annotation et filtrage sémantiques de flux RSS

Afin de permettre une sélection plus adéquate des informations en fonction de leur pertinence sémantique nous avons mis au point deux types de workflows, dont la figure 1 présente l'architecture générale : un workflow d'annotation sémantique des nouvelles et un workflow de filtrage sémantique des nouvelles (encadré par des tirets). La division en deux workflows est essentielle car elle permet de faire travailler nos workflows de manière asynchrone.

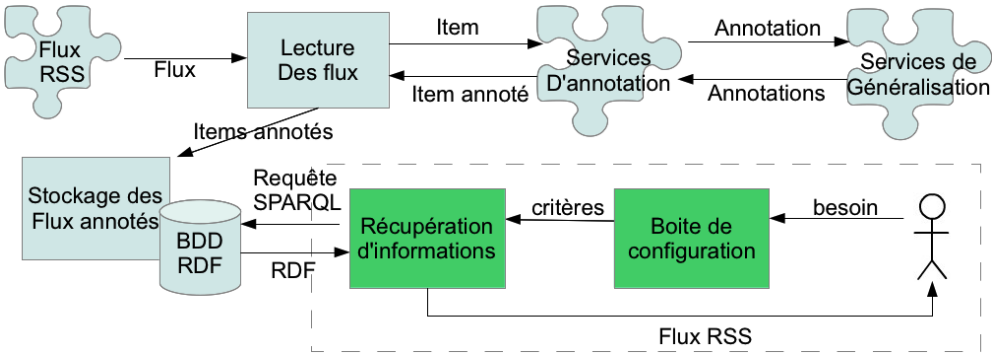

Figure 1: Workflows d'annotation et de filtrage sémantique de nouvelles

3.1 Workflow d'annotation sémantique des nouvelles

3.1.1 Lecture de flux RSS

Le workflow se base sur la récupération de flux RSS qui fournissent un lot de nouvelles. Le flux peut être mis à jour et doit être interrogé périodiquement pour récupérer les dernières nouvelles publiées.

3.1.2 Annotation sémantique de nouvelles

Des services enrichissent les nouvelles par des annotations et ces annotations elles-même peuvent être enrichies par l'utilisation de services de généralisation.

Ainsi, nous utilisons des *services d'annotation* pour qualifier chaque nouvelle par les entités nommées intéressantes la concernant. Les services d'annotation que nous avons utilisés sont :

- **OpenCalais**[4] permet d'identifier dans le texte les entités correspondant à des lieux français ainsi que des noms de personnalités en langue anglaise.
- **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 dû implémenter l'API Java se connectant à leurs services 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 sans utilisation d'un service de généralisation.

[4] http://www.opencalais.com
[5] http://www.wikimeta.com

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.

3.2 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 connexion à la base de données qui effectue une requête SPARQL multi-critères sur la base de donnée RDF pour récupérer les nouvelles à fournir à l'utilisateur. Enfin le système génère un flux RSS personnalisé pour répondre aux besoins utilisateur et le fournit au client. Ce flux RSS va pouvoir être mémorisé par le client pour récupérer en temps réel les nouvelles informations.

4 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, la construction automatique de workflows lui permettra, à partir d'un flux sélectionné, d'affiner la sélection des services d'annotations et généralisations qu'il souhaite utiliser afin d'obtenir une information enrichie par des annotations en fonction de ses besoins.

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.

4.1 Lignes de produits de services

Les choix que va faire l'utilisateur vont avoir un impact sur les trois concepts inhérents au workflow d'annotations représentés comme des pièces de puzzle dans la figure 1 : 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 indépendante, à partir de laquelle l'utilisateur choisira le service qui lui convient. Nous représentons ces lignes de produits par des modèles de variabilité (*feature models*) notés FM.

La figure 2 montre le FM *Service d'Annotation* qui permet de caractériser les services d'annotation existants (feature *Produit*), les thèmes (feature *Thème*) et les langues (Feature

[6] http://dbpedia.org
[7] `http://rdf.insee.fr/geo/`

Langue) supportés. La sélection d'une langue ou d'un thème influe directement sur le choix du service d'annotation par le jeu des contraintes internes au FM. On voit par exemple dans la figure 2 que la sélection de la feature *Santé* implique forcément la sélection du produit *MetaMap*.

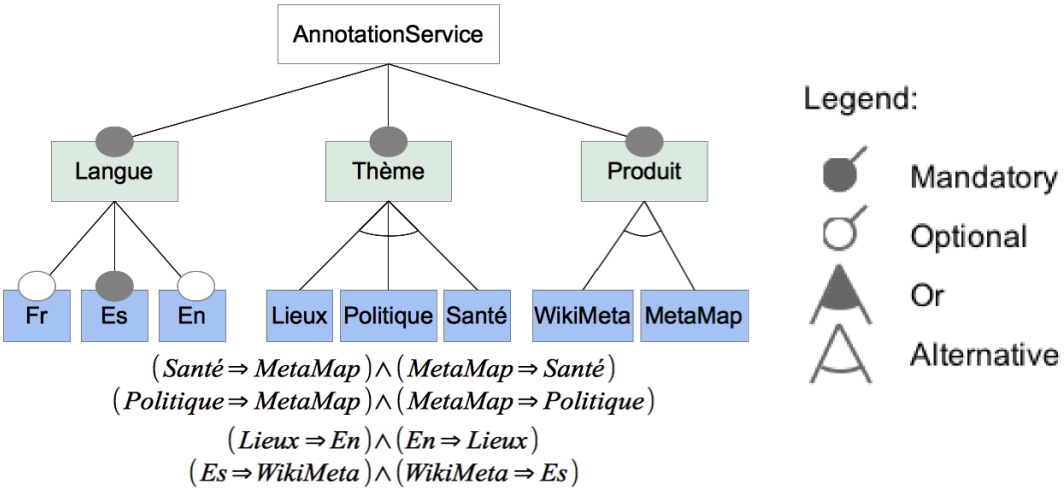

Figure 2: FM simplifié pour les services d'annotation

L'utilisation du FM via un DSL ou des travaux des langages naturels n'ont pas encore été abordés dans nos travaux. La construction du FM telle que présentée a été guidé par notre expérimentation. L'ajout de nouvelles sources devrait nous conduire à l'enrichir et éventuellement revoir sa structure.

4.2 Modèle de mise en relation des Lignes de Produits Logiciels

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 certains choix de l'utilisateur dans un des FM ait un impact sur les autres FM.

Nous exprimons donc des relations entre les différents FM représentants les concepts du domaine. Une relation indique 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. En effet, les systèmes d'annotation sont spécialisés pour certains types d'informations, comme le service MetaMap[8] spécialisé dans l'annotation de données médicales, et ne traitent que certaines langues.

Par exemple, un *Flux* traitant de politique pourra être annoté par un ou plusieurs *Services d'annotation* pouvant enrichir ce type d'information : les services, une fois sélectionnés, seront donc liés à ce flux, par exemple pour l'enrichir en données géographiques ou de santé. Cependant ce *Flux* ne pourra pas être lié à un *Service d'annotation* prenant en charge des langues différentes : la sélection du flux influe directement sur la sélection des *Services d'annotation* à lier, tout simplement en empêchant l'utilisateur de sélectionner un service qui n'est pas compatible selon le critères de *langue*.

Pour cela nous définissons et appliquons des opérations sur les FM [1] afin de restreindre les choix de l'utilisateur, 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.

[8] `http://metamap.nlm.nih.gov/`

Sur la base du modèle ainsi obtenu qui fait référence aux différentes configurations, nous envisageons de générer automatiquement les workflows en utilisant à la fois les assets associés aux FM et les transformations sur les modèles.

5 Conclusion

A l'heure où de plus en plus de services diffusent des flux d'informations sur le Web, il devient indispensable de posséder des outils afin de pouvoir filtrer ces informations. L'utilisation d'un système de filtrage sémantique par annotation des informations semble particulièrement intéressant en permettant à un utilisateur non-informaticien de définir de multiples critères de recherche et de filtrage.

Nous avons présentés dans cet article le système d'annotation ZeOntologyNewsExtractor[9] dont une preuve de concept est accessible 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 des lignes de produits logiciels afin de construire automatiquement nos workflows d'annotation.

L'intégration de ce travail au projet YourCast constitue une validation grande échelle du prototype. Ce projet a, en effet, pour objectif d'aider à la construction de systèmes de diffusions d'informations et il nous paraît intéressant d'être en mesure de proposer aux usagers un outil de filtrage sémantique d'information.

D'autre part des services tel que google ranking pourraient de trier les informations par importance et l'utilisation de processus de regroupement permettrait de lier des news similaires entre elles [9].

References

[1] Mathieu Acher. *Managing Multiple Feature Models : Foundations , Language and Applications*. PhD thesis, Université Nice Sophia-Antipolis, 2011.

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

[3] 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.

[4] 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.

[5] Eric Charton, Michel Gagnon, and Benoît Ozell. Automatic semantic web annotation of named entities. In *Canadian Conference on AI*, pages 74–85, 2011.

[6] 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.

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

[8] Technical Report, Kyo C Kang, Sholom G Cohen, James A Hess, William E Novak, and A Spencer Peterson. Feature-Oriented Domain Analysis (FODA) Feasibility Study. *Distribution*, (November), 1990.

[9] Fekade Getahun Tadesse. *Framework de gestion sémantique de flux d'actualités*. PhD thesis, Université de Bourgogne, 2011.

[9] `https://github.com/descl/ZONE`