Composition de workflows pour le filtrage sémantique d'informations

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. Tandis que les sources se multiplient (flux rss, services web, ..), la quantité d'informations croît et il est difficile de les filtrer en fonction de nos centres d'intérêts. Actuellement de nombreux outils qui exploitent les ontologies ou les thésaurus sont mis au point. Il 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.

1 Introduction

//[j'aime pas vraiment le titre...]:MI

Le web se révèle aujourd'hui un merveilleux support de diffusion des informations. Tandis que les sources se multiplient (flux rss, services web, ..), la quantité des informations croît et il est difficile de les filtrer en fonction de nos centres d'intérêts[**?**]. Des outils qui exploitent les ontologies ou les 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.

//[enrichir ce qui précede avec des références en essayant si possible de faire ressentir les éléments de l'architecture.]:MI

Il devient aujourd'hui possible de construire à la fois des workflows mettant en jeux ces différents outils pour annoter les flux d'informations puis les sélectionner les informations. Cependant la construction de ces workflows reste technique malgré les nouveaux supports logiciels tels que les mashup [**?**].

//[expliquer un tout petit peu pourquoi c'est pas simpel]:MI

. De plus de nouvelles ontologies, sources, systèmes d'annontations apparaissent régulièrement tandis que le web se démocratise [**?**]

//[et surtout bien le dire]:MI

. Dans ce contexte, la production automatique de ces workflows à partir d'un ensemble de caractéristiques proposées à l'utilisateur apparaît comme d'une grande utilité.

Dans cet article, en section 2 nous présentons un cas d'étude qui a été mené dans le cadre projet ANR EMergence Yourcast qui vise à automatiser la diffusion des informations sur de

grands écrans. Nous montrons au travers de ce cas d'étude les différents choix qui se posent à l'utilisateur

//[MOntrer cela]:MI

et décrivons dans la section 3 les workflows mis en place pour répondre à ce cas particulier. Fort de cette expérience, nous proposons en 4 de produire de tels workflows en utilisant un développement dirigé par les modèles et les feature models pour produire automatiquement de tels workflows à partir de données utilisateur de haut niveau.

2 Système de diffusion des informations sur grands écrans et filtrage

//[Clairement je cherche...]:MI

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.

Des sources hétérogènes Or il existe aujourd'hui de nombreuses sources ..... le choix des sources peut être simplement lié au travail.... Dans notre cas, nous avons choisi ....

Sélections des informations Beaucoup de sources d'informations sont visibles aujourd'hui sur internet sous la forme de flux rss. Les critères de sélection sont alors pré-établis. Par exemple sur le site de news de google `http://news.google.fr` vous pouvez accéder à des nouvelles liées à l'économie ou bien les news locales mais si vous vous ne pouvez pas récupérer les news économiques liées à la ville de Marseille ou plus largement à la région PACA.

Dans notre cas nous avons choisi d'analyser une 20ene de flux rss sur de vastes sujets. Ceux-ci sont essentiellement des flux aillant déjà une étape de filtrage qui a permis de les classer selon leurs thématique générale (technologies, international, médical...).

Des critères utilisateurs Pour capturer ces critères il existe différents systèmes... [citer des exemples?]

Nous avions besoin d'un système intuitif qui permet de capturer simplement les exigences utilisateur. Nous avons choisis sur ce point 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.

Des processus d'annotations diversifiés 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 (Resource Description Framework [**?**, **?**, **?**, **?**]). Nous utilisons alors des systèmes d'extraction d'information et d'annotation sémantique qui permettent d'ajouter des liaisons vers les entitées nommés spécifiques.

3 Système de diffusion des informations sur grands écrans et filtrage

=======

Etant donné que nous aurions des bases annotées, comment supporter l'expression de criteres et leur utilisation pour filtrer ces bases.... qui conduit à orchestrer un ensemble de composants.

=¿ Challenges 2 Comment exprimer de tels critères pour que le filtrage des nouvelles soit opérationnel?

Sachant que il y a qualité de l'information, le critere.... qui conduit à orchestrer un ensemble de composants lié u précédent.

Donc .....

Il existe aujourd'hui de nombreuses sources d'informations, peut-on soumettre ces masses d'informations aux mêmes artefacts de filtrage et comment?

L'expression des critères a été étudiée .....

4 Mise en oeuvre

Dans le cadre de l'étude présentée prcédemment nous avons donc mis au point deux workflows, dont nous présentons à présent l'architecture brièvement.

//[Mettre une figure visualisant le workflow]:MI

4.1 Workflow d'enrichissement des informations

4.1.1 Lecture de flux RSS

Construction du diffuseur d'informations par un utilisateur final

4.1.2 Annotations

4.1.3 Mémorisation

//[je ne sens pas cette partie.. à voir]:MI

4.2 Workflow de filtrage des informations

4.2.1 Gestion des critères

4.2.2 Filtrage

5 Vers la contruction automatique de workflows

Notre objectif à terme est de construire une ligne de produits qui capturerait les différents sources et systèmes d'annotation disponibles, les qualifierait et permettrait à un utilisateur final de construire ses propres workflow en sélectionnant pour lui les sources et les systèmes d'annotations idoines conformément à ses besoins.

5.1 Feature Model de sources

5.2 Feature Model de services d'annotation

5.3 Metamodele de mise en relation des FMs

5.4 Vers la génération des codes

Des travaux relatifs à la construction de workflows scientifiques ont été menés...

Notons que la pertinence d'un ssytème sur la langue n'est pas seulement oui ou non mais est quantifiée ce point reste une piste ouverte.

6 Conclusion