

Vers la composition de workflows pour le filtrage sémantique d'informations

Christophe Desclaux
Université Nice Sophia-Antipolis
christophe@zouig.org

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

Simon Urli
I3S, CNRS
Université Nice Sophia-Antipolis
urli@i3s.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. 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, WebServices, ...), the amount of information is growing and it's difficult to filter them according to our interests. Currently many tools that exploits ontologies or thesauri are developed. They help us to annotate informations and to derive criteria and then get only the relevant information. The composition of these tools constitute workflows that should be enriched by the emergence of new ontologies focused on 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 have built and present our outlook for the automated build of workflows based on user needs. This work is supported by the ANR Emergence Yourcast project designed to automate the dissemination of information on large screens, and for which the relevance of the information published is particularly relevant.

1 Présentation et importance du filtrage des informations dans le web de données

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 [3]. Des outils qui exploitent

les ontologies ou les thésaurus dans le web sémantique [8] ont été mis au point qui permettent d'annoter les informations, d'en déduire des critères et d'ensuite obtenir uniquement les informations pertinentes.

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 [6] [9]. 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].

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 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.

2 Système de diffusion des informations sur grands écrans et filtrage

//[Clairement je cherche le titre...]: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 d'information disponibles grâce à l'utilisation de flux RSS [7]. 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 aillant déjà subi une étape de filtrage qui a permis de les classer selon leurs thématique générale (technologies, international, médical...).

Sélections des informations 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 <http://news.google.fr> vous pouvez accéder à des nouvelles liées à l'économie ou bien les news locales. Cependant vous ne pouvez pas récupérer les news économiques liées à la ville de Marseille ou plus largement à la région PACA. Dans notre exploitation des flux RSS nous devons donc pouvoir fournir un filtrage multi-critères permettant un tri fin des informations.

Des critères utilisateurs Pour capturer ces critères il existe différents systèmes actuellement tel que le service <http://www.google.fr/reader/> ou bien <http://rsslounge.aditu.de>

cependant ceux-ci ne proposent pas de regroupement des flux et de filtrage multi-critère de ceux-ci. En effet, nous avons 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 que nous avons construit et 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'annotation diversifié 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 annotations basées sur la récupération d'entités nommées présentes sur un grand nombre de bases de données RDF (Resource Description Framework [11]). Nous utilisons alors des systèmes d'extraction d'information et d'annotation sémantique qui permettent d'ajouter des liaisons vers les entités nommées spécifiques.

3 Mise en œuvre

Dans le cadre de l'étude présentée précédemment nous avons donc mis au point deux workflows, dont nous présentons à présent l'architecture brièvement. Comme présenté sur le diagramme suivant nous avons ainsi le workflow d'enrichissement des informations (coloré en bleu ainsi que le workflow de filtrage des informations.

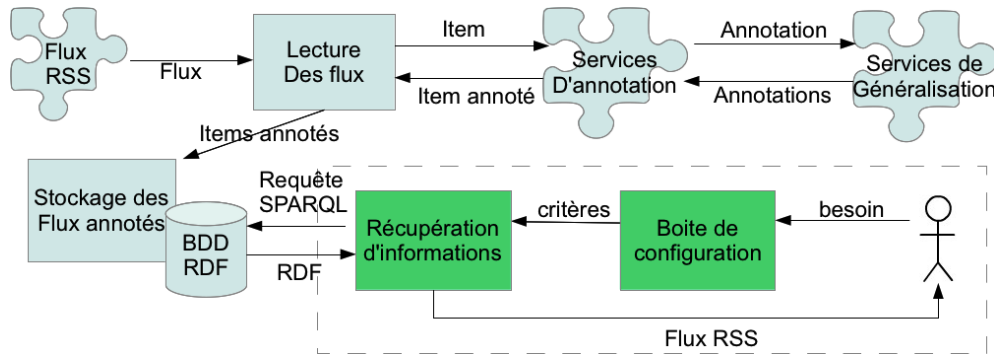


Figure 1: Workflow de traitements....

3.1 Workflow d'enrichissement des informations

Nous allons nous attarder particulièrement sur ce workflow car il fait appel à de nombreux services d'annotation et nécessite une modularité dans l'organisation de ceux-ci.

3.1.1 Lecture de flux RSS

L'application 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 va contenir un lot de news les plus récentes publiées chez le fournisseur du flux. Le flux peut être mis à jour et nous devons l'interroger périodiquement pour récupérer les dernières informations publiées.

3.1.2 Annotation

Sur chaque news nous devons utiliser les services d'annotation pour récupérer les entités nommées intéressantes le concernant. Pour cela nous faisons appel à divers services d'annotation qui nous renvoient les informations. Nous avons aussi besoin de faire appel à des services de généralisation qui vont travailler en aval des services d'annotation pour ajouter des notions à une annotation.

De manière concrète nous utilisons les services d'annotation suivants:

- **OpenCalais** nous permet de récupérer les entités des textes correspondant à des lieux français ainsi que des noms de personnalités en langue anglaise. Nous lui indiquons en entrée le texte à annoter et récupérons une liste de mots correspondant à nos critères.

- **WikiMeta** fonctionne de la même manière qu'OpenCalais mais permet de récupérer directement des entités nommées pointant vers la base de connaissance DBpedia, il peut donc être utilisé seul en tant que service d'annotation principal

Et les services de généralisations suivants :

- **DBpedia** 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** <http://rdf.insee.fr/geo/> La base de données RDF de l'INSEE nous permet de récupérer les ensembles géographiques liées 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 renvoi en sortie des URI vers les entités géographiques liées à la news étudiée. Nous utilisons pour cette interrogation la requête SPARQL suivante qui permet d'avoir les informations désirées:

```
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'))  }}
}
```

3.2 Workflow de filtrage des informations

Le workflow de filtrage permet de créer des critères utilisateurs afin de récupérer les données intéressantes sur la base de données. Chaque critère est défini par le couple (TypeDeRelation, EntitéNommée) et va permettre un choix précis des éléments à récupérer. Nous fournissons une application complète permettant de choisir les critères à appliquer sur la base [5].

4 Perspective : 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 des différents services.

4.1 Objectifs

La construction automatique de workflows d'annotation permet de répondre à différentes problématiques. 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.

En outre, à 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 service 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.

4.2 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 pris comme une ligne de produit indépendante à partir de laquelle l'utilisateur choisira le produit qui lui convient. Nous représentons ces lignes de produits par des modèles de features (FM).

//[Réf FM / SPL - Expliquer choix de la SPL ?]:SI

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

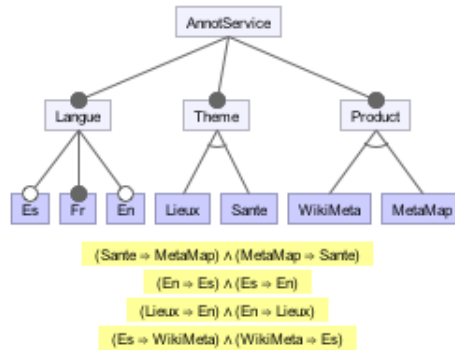


Figure 2: Workflow de traitements....

4.3 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'annotations sont spécialisés pour certains types d'informations, comme le service MetaMap [10] et ne traitent

que certaines langues. Une relation exprime à la fois qu'un produit (ie. 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 politiques va être annoté par un ou plusieurs *Service 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.

5 Conclusion

References

- [1] M Acher. *Managing Multiple Feature Models : Foundations , Language and Applications*. PhD thesis, Nice-Sophia Antipolis, 2011.
- [2] Jean-Noël Anderruthy. *Web 2.0: (r)évolutions et nouveaux services d'Internet*. 2007.
- [3] Mokrane Bouzeghoub, Dimitre Kostadinov, and Others. Personnalisation de l'information: aperçu de l'état de l'art et définition d'un modèle flexible de profils. *Memorias de CORIA*, page 201, 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] Desclaux Christophe. ZeOntologyNewsExtractor.
- [6] Ingbert R Floyd, M Cameron Jones, Dinesh Rathi, Michael B Twidale, and Information Science. Web Mash-ups and Patchwork Prototyping: User-driven technological innovation with Web 2.0 and Open Source Software. *Source*, pages 1–10, 2007.
- [7] Orla Lassila and Swick Ralph. Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation 22 February 1999. Technical report.
- [8] TB Lee, J Hendler, and O Lassila. The semantic web. *Scientific American*, 2001.
- [9] N. Milanovic and M. Malek. Current solutions for Web service composition. *IEEE Internet Computing*, 8(6):51–59, November 2004.
- [10] National Library of Medicine. MetaMap.
- [11] D Winer. RSS 2.0 Specification (RSS 2.0 at Harvard Law), 2005.