RSS Merger

Fekade GETAHUN, Richard CHBEIR

LE2I Laboratory UMR-CNRS, University of de Bourgogne
Engineer's wing, 9 Savary St., 21078 Dijon Cedex France
{fekade-getahun.taddesse, richard.chbeir}@u-bourgogne.fr

1 Introduction

RSS is the latest XML based specification used mainly in the publishing industry to render news article, products, notice, etc. in standardized and transparent way. Merging or combing news collected from set of providers is very important to different users (e.g. journalists, researchers, students, merchants, etc). E-news readers use RSS aggregator to add new, change existing, remove existing feed address; download and read news without roaming from site to site. However, to the best of our knowledge, there is no RSS aggregator that identifies the similarity between news items and merges them (putting them together) in an intelligent way. Here, we propose our *RSS merger* that accepts the interest of a user (e.g. key words, favorite news feeds, etc.) and produces a feed containing news items reorganized as per the rules provided by the user. In this demo paper, we will demonstrate the architecture, different components and the ways users can use the system.

2 System Architecture

The desktop version of our RSS merger application has 6 main and interacting components as shown in Figure 1: `Parser`, `Concept Extractor`, `Relatedness`, `Rule editor`, `Merger` and `Output Generator`. The `Parser` component starts by getting feed address (link) stored in user profile. In addition, the parser checks the validity, well formedness of the feed and it extracts the news items embedded in channel. The `relatedness` component measures the semantic relatedness between texts, simple elements and items. The relatedness between texts (i.e. content of simple elements) is computed using Vector Space Model (Salton, 1983) after extracting key concepts embedded in each text using the `concept extractor` component. The relatedness between simple elements and items is identified by combining the relatedness between textual values (c.f. Getahun et al, 2009). The `concept extractor` component generates the concept set of a given text after removing the stop words using predefined dedicated dictionary. The component uses WordNet (Richardson et al, 1995) taxonomy as source of external knowledge base representing the semantic information. In (Getahun et al, 2009), we have provided action-oriented merging approach. The merging of items is dependent on the relationship value existing between the items. The `rule editor` component, shown in Figure 2, allows users to associate the relationship existing between objects with action list. To do that, we provide the user with a wizard having three steps. In Step 1, the user chooses/modifies the output type which can be RSS 2.0, XHTML or HTML. In Step 2, the user provides his/her perception of merging news items by associating a relationship with an action. In addition, users are privileged to provide custom rules by combining several actions. The ELSE rule, in Figure 2, is applied to merge

intersecting and disjoint news items using the rules provided in Step 3. If the user didn't provide an ELSE rule, the RSS merger applies the default merging rules associated to relations.

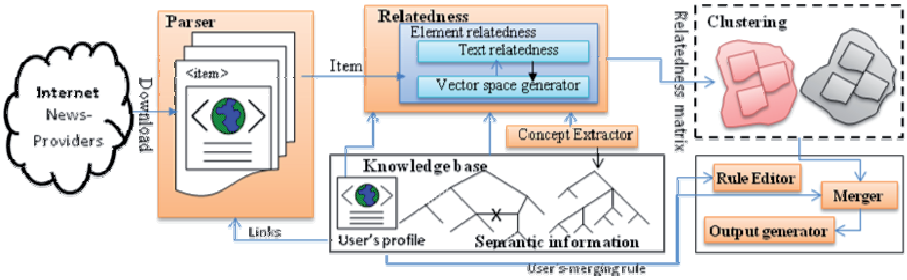

FIG. 1 – *RSS Merger architecture*

The final step (i.e. Step 3) is optional and used to merge sub-elements of an Item as in Step 2. The `merger` component performs the actual merging of the news items taking into consideration the rules provided by the user and stored as part of his profile. `Clustering` is pre-condition and facilitates merging process. The merger accepts a matrix containing pairwise relatedness between news items in the same cluster and merging rule so as to get the merged version. Finally, the `output generator` component produces a document containing the result of merger in the format specified by the user.

FIG. 2 – *Screen shot of RSS merging rule editor*

References

Getahun, F., J. Tekli, R. Chbeir, M. VIVIANI, and K. YETONGNON (2009). Semantic-based Merging of RSS items. WWW Journal 11280, Springer US.

Salton, G. and M. McGill (1983). Introduction to Modern Information Retrieval. McGraw-Hill.

Richardson, R and A.F. Smeaton (1995). Using Wordnet in a knowledge-based approach to information retrieval. Technical Report ca-0395, Dublin City University, Dublin, Ireland.