Use Case 1 in Media and Communications - Research

News Aggregation Service

IB Member: Neofonie GmbH
KW Partner: FU Berlin

1 Overview

Challenge
To integrate and accurately classify news articles, dynamic web pages, press releases and feeds into a news service
Solution
At the moment a semi-automated integration takes place: (1) Template creation using neofonie search:webstract, (2) Automatic processing of the pages and feeds.
Why a Semantic solution
Semantic technology includes ontology mapping which is becoming increasingly important. As more information providers turn to semantic technology, the returns on automatic solutions increase. However to remain leaders in the market support must still be maintained for legacy systems. Integrating semantic technology in current solutions is therefore a must.
Key Business Benefits
Greater returns on diminishing labour costs. The possibility of concentrating development on new applications instead of on integration.

This business case deals with the provision of an aggregated news service which is able to provide accurate search, thematic clustering, classification of news stories, and e-mail notification of stories of user interest.

There is a business interest in following news in specific categories, including economics, science and IT, or on specific subjects such as particular companies, or developments in the business area. To do this, human resources need to be allocated which is of course a costly business in terms of both time and money. Presently a human will need to manually sort through a large number of different sources to find the news of interest to the business. This involves selecting not just news stories from the main news feeds and media outlets but also press releases, announcements on websites and other 'alternative' sources.

Keys components

Existing Software
Aggregration system
Natural Language Processing

Research and development
Semantic annotation of documents
Ontology mapping
Semantic extraction of themes
Fault tolerance
Security and trust

Technology locks
Artificial Intelligence issues

While existing developments in providing aggregation services as feeds or Web portals aim to bring information more quickly to the user or users more quickly to the desired information, there remains a significant effort in terms of sorting and selecting the relevant stories. The systems operate by analysing simple metadata tags in the source documents (which are largely not standardized, variable and inexact) and the natural language syntax which is an ambiguous and inexact science.

We believe that the combination of use of semantics in the source documents and a semantically-aware aggregation system can be a major step in reducing the significant effort that still exists in finding the news stories of interest among the information swell existing on the Internet.

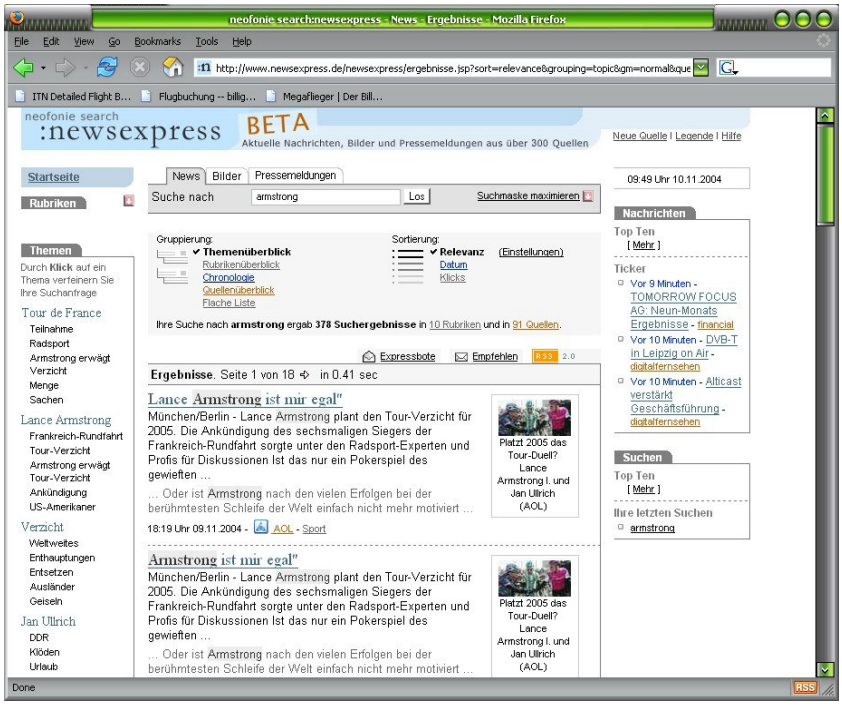

Figure 1 – The NewsExpress website

2 Current Practices and Technologies

2.1 Current business practises

The news aggregation service http://www.newsexpress.de from neofonie GmbH[1] (see Figure 4.1) is representative of a new class of news aggregation system which can be used to provide a targeted view of the news to users through insertion of its Web-based user interface into client Web sites, whether in an Intranet or Internet environment.

Its current implementation is a hybrid solution, made up of

(1) the manual creation by a source expert of a XSLT template for each news source
(2) the automatic processing of that news source through a thematic clustering algorithm (NLP) and classification with category mappings.

The template is able to convert the semi-structured source data (HTML or XML) into the internal format of the aggregation service and to classify the news story at different granularities within certain themes and (pre-defined) classifications using a mix of element analysis (e.g. value of HTML META tags) and content analysis (NLP on the story text).

As a result heterogeneous sources are integrated and made available through a single interface. Views of the integrated news data is available based on the traditional chronological model (i.e. latest news first) or on the basis of user selection (i.e. number of clicks). Additionally searches can be made based on classification or theme (the clustering of subjects which occur often together in a story).

[1] http://www.neofonie.com

2.2 System requirements Analysis

The aim of the further development of newsexpress is to increase the accuracy and automatisation of the service. In seeking to achieve these aims, the following issues stand out:

(1) The use of a consistent and clearly understood vocabulary in metadata by the source documents to enable a correct classification
(2) The mapping between different vocabularies and a core system vocabulary to ensure semantic integration over the distributed and hence heterogeneous character of the Web
(3) The extraction of semantics from the source documents in order to remove the need for a fully manual metadata authoring by source document authors (which discourages the creation of metadata in the first place)
(4) Further development in Natural Language Processing and particularly the use of controlled vocabularies in relating NLP results consistently to specific subjects
(5) Addition of domain knowledge to searches to ensure more intelligent and accurate results. This could be seen as an extension of the thematic clustering which is already a step towards modelling relations between concepts within a particular topic.
(6) The need for security from the point of view of source providers (that their content will not be manipulated or inappropriately used) and trust from the point of view of the aggregation service (that news is accurate and not falsified)
(7) The need for fault tolerance in the integration process. A syntactic process breaks when the source syntax is changed, until the template is manually altered.
(8) Duplicate recognition. By modelling the subject of news stories and comparing models, stories which repeat the same issues can be filtered out while stories on the same topic offering a different slant on the story can be used to complement results.
(9) Performance. As semantic technologies are introduced, the tasks of knowledge extraction, concept mapping and semantic-based search must demonstrate sufficient performance capabilities to not bottleneck the system.

It is also recognised that adding semantic data to sources is still not well enough understood by users. The supporting technology doesn't make it simple enough to add metadata (also ontology based) to information. This is a UI issue with the various tools that are currently being used to produce content.

2.3 Review of the current systems

Current systems can be categorised in two broad categories
(1) Feed services, e.g. RSS-based
(2) Web portals, e.g. Google News

Feed services such as those based on the use of RSS or Atom (e.g. Syndic8 http://www.syndic8.com/) offer a flat integration of news from different sources (i.e. all stories exist on the same level). There is a heterogeneity of RSS versions and even their use by authors which has caused integration problems. The content offered by RSS is generally very basic (e.g. title, author, link to full story) and not suitable for any intelligent searching or organising.

Portals such as Google News (http://news.google.de) offer a large body of information that can be processed (i.e. sheer system power) and the organisation of that information through text processing, user clicks, etc.

newsexpress seeks to offer a cleaner solution in focusing on a more elegant semantic approach to news story organisation, being faster to publish breaking news stories and providing stronger means to find and follow news stories of interest (subject classification, thematic clustering, different views on the same stories). It is also implemented to support further development in the use of semantic technologies.