Use Case 1 in Media & Communications – Research Challenges

News Aggregation Service

KW Partner: FU Berlin
IB Member: Neofonie GmbH

1. General Description of Business Use Case

The business case deals with the provision of an aggregrated news service that is able to provide business clients with accurate search, thematic clustering, classification of news stories, and e-mail notification of stories of interest.

There is a business interest in following news in specific categories or on specific subjects such as particular companies or developments in the business area. To do this, human resources need to be allocated which proves costly 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. These sources are not just the main news feeds and media outlets but also press releases, announcements on websites and other "alternative" sources.

Current approaches operate by analyzing simple metadata such as tags in the source documents or the text of links found to those documents, which is dependant upon correctly interpreting ambiguous and inexact natural language. As a result, there remains a significant effort for the human user in sorting and selecting relevant stories.

2. Proposed Semantic Web-based Solution

The news aggregration service http://www.newsexpress.de is intended to demonstrate a new class of news aggregration system which can be used to provide a targeted view of news to users through the insertion of its Web interface into client Web sites, whether in an Intranet or on the World Wide Web.

The current solution is a hybrid, semi-automatic approach consisting of two phases:
(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 first step is aimed at converting the semi-structured source data (HTML or XML) into the internal format of the aggregration service. The second step classifies the story at different granularities within certain themes and pre-defined classifications using a mix of element analysis (e.g. HTML META tags) and content analysis (NLP on the story text).

Hence, given the lack of clear semantics in current Web sources, there is no explicit use of Semantic Web solutions at the current time. However it is recognised that as information providers would move to semantic technologies, the returns on automated solutions would increase. The XSLT template would be replaced by mappings that could be increasingly dynamically determined, and the existence of ontologies could simplify greatly the task of story classification.

3. Identified Research Challenges

3.1 Content Annotation

3.1.1 Problem Statement

Adding semantic data to sources is still not well enough understood by users. The supporting technology doesn't make it simple enough to add ontology-based metadata to information.

3.1.2 Knowledge Processing Task and Component

The **annotation manager** is the component that is tasked with facilitating the association of metadata to resources. The approach is seen as being semi-automatic, i.e. the extraction of characteristics that can be automatically determined combined with the prompting of the user to provide values for those characteristics that can not be automatically determined.

3.1.3 Requirements Analysis

First of all, this is an UI issue with the various tools currently in use to produce content. Alternatives, such as plug-ins for legacy systems, should add this functionality intuitively into the content production process. This may also help to automatically capture information about the resource that later can only be manually provided (e.g. a source, or date of creation).
As it is, an annotation tool is more likely to be an additional task performed later on created resources. Such a tool should extract as much information automatically as is possible, and facilitate the user in the manual aspect of the task (e.g. making intelligent suggestions based on textual analysis or the annotation choices in similar resources).
Guidelines would also be useful in terms of appropriate granularity. Simplicity in usage (hiding ontology details) must be balanced with flexibility in selecting what annotation scheme will be used (which properties and taking which values). Annotations must be rich enough for their future processing, but not too complex for system performance.

3.2 Ontology Mapping

3.2.1 Problem Statement

Given the distributed and heterogeneous character of the Web, it is expected that resource annotations will not all follow a single annotation scheme, nor should they have to. However, if the news aggregration service is to be able to make sense of the collected stories it must be able to rely on some core description format.

3.2.2 Knowledge Processing Task and Component

The task of **data translation** is in charge of translating instances of heterogeneous information sources storing their data in different formats. In this case, instances use different ontologies to provide descriptions or markup for a piece of content. These ontologies may be referencing equivalent concepts but may also provide descriptions at different levels of conceptualisation and granularity. The wrapper component that realises this task accepts input descriptions using different ontologies and returns semantically equivalent descriptions expressed in a common single ontology.

3.2.3 Requirements Analysis

The wrapper must be able to provide best mappings for the entire set of possible input ontologies. Given that it may not be possible to specify all possible inputs, it must be able to make intelligent decisions on properties and values it does not recognise (e.g. by assessing semantic similarity to known properties and values). Increased dynamicism by the determination of mappings is a key requirement for system scalability and fault tolerance (when a source changes its annotation scheme) on the Web.
Guidelines are required for the core ontology used by the service, which must find a balance between expressability (in order to be able to model all information from all sources) and simplicity (the service only needs to model all that it requires to support all of its functionality and less complexity should mean better performance).
Given the desire to improve upon existing non-semantic approaches, e.g. by bringing relevant stories to users in a timely manner, on-the-fly concept mapping must demonstrate satisfactory performance capabilities as to not bottleneck the system.

3.3 Ontology Development/Maintenance

3.3.1 Problem Statement

The news aggregration service classifies the stories it collects to allow a thematic clustering of similar stories or the placement of stories in a categorization scheme. In the non-semantic approach this relied on NLP and suffered from the ambiguities of natural language. A semantic approach clearly offers an advantage in this task as long as content is well annotated. The problem shifts to developing the ontology that models the classification of news stories, as news content can cover an extremely wide range of subjects. While a comprehensive ontology might allow very fine-grained search or high-relevance selection of stories for a user, the manual development and maintenance of such an ontology would bring significant cost disadvantages to a system implementation.

3.3.2 Knowledge Processing Task and Component

The **ontology management** task is focused on dealing with the construction and evolution of an ontology in a business use case. In this case, the ontology to be handled is expected to be continually evolving.

3.3.3 Requirements Analysis

While ontology development guidelines or ontology recommendations could feed into the production of an initial “upper level” categorization of news stories, the semantic value of the approach is realized in finer grained modelling of the domains of news stories, hence enabling intelligent search and retrieval for users. While ontology re-use might be possible, domains might also be modelled on the fly based on analysis of similar stories. The important difference here is not to produce a simple property value-allocation but to determine semantic relations between the concepts in the resource annotation, e.g. that “Jan Ullrich” is a “cyclist” or that “Tony Blair” “leads” “the British Labour Party”. Determining semantic equivalencies is also important, and good methodologies are required (e.g. the specification or determination of unique properties – if two concepts have the same value for an unique property then they are the same concept). Finally, given the difficulties in directly relating concepts automatically, levels of certainty could be allocated to those relations, which play a role in the semantic search and retrieval.

While system scalability will require determining concept relations in an automated fashion, the task must also detecting inconsistencies which may result from changes it determines and as a result being able to reject them or query the administrative user before applying them.

3.4 Search

3.4.1 Problem Statement

The search functionality of the system has relied on text matching and been complemented by thematic clustering which is offered as an alternative means to find relevant stories. With the provision of semantic annotations with the news stories backed by conceptual relations odeled in an ontology, search needs to be extended to take advantage of this knowledge in order to produce better precision and recall of stories.

3.4.2 Knowledge Processing Task and Component

The **query processor** is the component, which should be able to generate a semantic query, as determined from the user input at some UI, and pass it for interpretation at the **reasoner**. The reasoning component is able to use the ontological knowledge available to it to answer the query by making the appropriate logical inferences. Finally, a **results reconciliation** task is responsible for ensuring that there is no redundancy in the search results e.g. duplication or overlapping of information.

3.4.3 Requirements Analysis

Queries should be expressive enough to handle user requests for relevant news stories. While single term search is common among users, semantic relations can best be exploited by supporting the formulation of richer queries from the UI and, of course, being able to handle rich query structures in the reasoner. Given that the core ontology might be large, mechanisms are needed to perform efficient queries that, after all, likely only use a small fragment of the ontology.
The retrieval of matching stories must be able to determine between stories that complement one another and those whose content overlaps or duplicates others. This might be a case for semantic matching, i.e. comparing the models of two stories. Another aspect of complementary retrieval could be to select fragments of stories based on their annotation.

3.5 Security and Trust

3.4.1 Problem Statement

The dynamic integration of heterogeneous news sources by the aggregration service raises two issues,

- From the point of view of content providers, there is the need for security, i.e. that their content will not be manipulated or inappropriately used;
- From the point of view of the user, there is the need for trust, i.e. that the news they find is accurate and not fabricated or falsified.

3.4.2 Knowledge Processing Task and Component

Security may belong at the annotation manager. If we consider the target of the security concern to be the annotations, then the manager is tasked with storing those annotations in a secure manner. Concerns about content usage occur at the results reconciliation, if we consider the use of fragments of stories in the results list.

Trust could be considered at the point of content aggregration and factored into search results reconciliation. In the former case, the data translation task could take into account the consistency or reliability of the semantic model it generates for the given news story. In the latter, a query result could be ranked according to the system-determined trustworthiness of the source.

3.4.3 Requirements Analysis

In terms of security,

- The system should support the use of security measures in the storage of annotations produced at the source, as well as all data communication that takes place (e.g. during aggregration, including ontology mapping) and the local storage of aggregrated stories and their annotations.
- The annotation scheme could be extended to support the expression of rights usage for the story, determining for example that a given story can only be reproduced in full, or to support 'segmentation' of a story, defining in what way a story may be broken down into separate parts.

In terms of trust,

- Content aggregration could consider the semantic model it generates for a given story and based on prior models (for similar stories, in the most recent time) determine if the story can be trusted (e.g. if all recent stories have stated that "George Bush" is the President of the USA, a single story stating that the President is in fact "Mickey Mouse" is suspect).
- Ranking could be factored into the search results to prefer more trusted sources over less trusted. Trustworthiness of sources must be determined in some fashion, e.g. the allocation of digital certificates to trusted news sources by an accepted authority.