PressIndex: a Semantic Web Press Clipping Application

Florence Amardeilh, Olivier Carloni, Laurence Noël

PressIndex: a Semantic Web Press Clipping Application

Florence Amardeilh[1 2], Olivier Carloni[1 3], Laurence Noël[1 4]

[1] Mondeca,
3 cité Nollez,
75018 Paris, France
{fistname.lastname}@mondeca.com

[3]LIRMM, CNRS, University
Montpellier 2, 161 rue Ada,
34392 Montpellier cedex 5, France
olivier.carloni@lirmm.fr

[2]LaliCC, University Paris 4,
28 rue serpente,
75006 Paris, France
florence.amardeilh@paris4.sorbonne.fr

[4] LEDEN, University Paris Nord 8,
4 rue de la croix Faron,
93210 La Plaine Saint-Denis, France

Abstract. PressIndex is a project focusing on integrating semantic web technologies to provide new services for media monitoring and press clipping activities, especially in the domain of Competitive Intelligence. Ontology modeling, natural language processing tools, rule reasoning engines along with interactive user interfaces supported by the underlying semantics of the application are used to build tailored knowledge bases and semantic systems directly used to the knowledge discovery process.

Keywords: Semantic Annotation, Ontology Population, Rule Reasoning and Inference, Dynamic Publishing.

1 Introduction

Press clipping applications are concerned with knowledge management issues such as storage, management and analysis of relevant data distributed in different locations. To deal with the extraordinary growth in the volume information available, the companies usually rely on customized report analysis made by specialized press coverage solutions. PressIndex is one of them, offering value-added solutions for optimized media watch. It covers a broad range of printed and online European titles, including Nationals, Regionals, Technical, Trade and Consumer Press. In total, more than 9,500 media sources (print, television, radio, websites, blogs) are regularly covered by PressIndex's consultants who manually index the cited companies in order to produce tailor-made reports and accurate press coverage.

In order to improve and to quicken its press coverage delivery to its clients, PressIndex wants to enhance its press clipping process thanks to the Semantic Web technologies. It looks forward combining Text Mining and Machine learning

techniques along with Semantic Web standards to further automate its media watch and content aggregation. The new application is based on the following components:

- Information gathering adapted to the monitored medium (scan, read, audio, video)
- Knowledge extraction and indexing using Natural Language Processing tools, to extract strategic economic, stock-market and corporate information
- Aggregation and capitalization of the extracted data in a knowledge base constrained by a domain ontology, using a mediation layer to synthesize the collected information and control its consistency
- Distribution to clients in multiple tailored formats (company product or market summaries; business or financial event tracking in an industry…)

In the following section we present the overall PressIndex architecture and the PressIndex ontology. In section 3, we describe in further details the four components composing this solution. First set of results are the aim of section 4. Section 5 will provide conclusion and directions for future work.

2 The PressIndex Solution

One of the main bottlenecks of the existing application is content aggregation. The same event may appear several times, in different press issues, and within a different context: firstly as a rumor, then as an announcement and finally as an established fact. Thus, apart from aggregating the various mentions of the same event into a single one in the knowledge base, PressIndex also has to monitor the event historic: controlling if a rumor contradicts a precedent rumor, if an announcement confirms a rumor, etc.

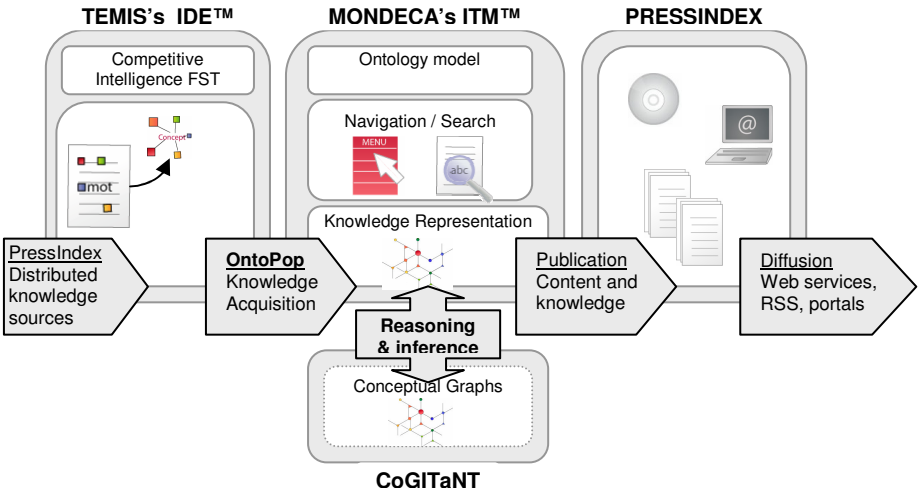

Fig. 1. The PressIndex architecture

To achieve the goals of the new press clipping application, PressIndex signed a partnership with Mondeca and Temis in order to construct a modularized and flexible architecture (see **Fig. 1**) based on their state-of-the-art tools, respectively the Intelligent Topic Manager (ITM), a knowledge management tool based on semantic web technologies [1] and the Insight Discoverer Extractor (IDE), a natural language processing tool to extract pertinent data from textual corpora [2]. Moreover, the PressIndex application will also integrates CoGITaNT[1], a GPL library dedicated to develop applications based on conceptual graphs for doing rule inference [3].

The PressIndex application also highly relies on its competitive intelligence ontology, the PressIndex Ontology. The ontology was manually designed to fulfill the needs of the upcoming press clipping application. It was implemented in OWL language. Then it was imported in both ITM to configure the domain-oriented knowledge base and in CoGITaNT to be able to run the reasoning and inference rules (after transforming it into the conceptual graph format).

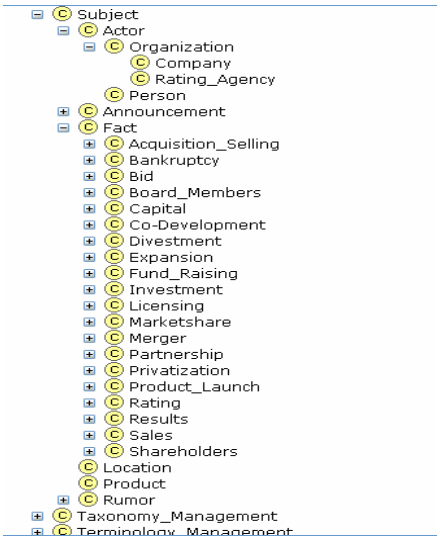

Fig. 2. Part of the PressIndex Ontology

The PressIndex Ontology (see **Fig. 2**) represents the Actors (the Persons, Organizations, Companies and Rating Agencies), the Products, the Locations and the Facts. The Fact concept deals with competitive intelligence's events:

- describing a particular company such as Bankruptcy, Capital, Fund-Raising, Privatization, Results,…
- or connecting a company with actors or products such as Acquisition-Selling, Board Members, Co-Development, Investment, Merger, Partnership, Product Launch, Shareholders. Those are also modeled as relations in the ontology.

The Facts have shared properties such as the list of companies involved, the location where the event took place, the date when it occurs, the sentence containing the extracted fact, the source reference when available. The PressIndex Ontology also distinguishes the facts from their rumors and announcements. Relations between the rumors and announcements are modeled to be used by the inference engine to point the user with inconsistent extractions such as contradictory rumors or confirmed rumors, inconsistent announcements or a rumor refutation by an announcement.

In the following section, we present the new process for clipping the various press resources. We will illustrate our presentation with some examples taken from the project.

[1] http://cogitant.sourceforge.net

3 The Press Clipping Process

3.1 Filtering the resources

The actual PressIndex application deals with around 200,000 documents every day. The press coverage mainly concerns distributed European knowledge sources, in various domains, on different supports (e.g. printed newspapers, online websites, radio and television resources, etc.) and in different formats (the prints are scanned, compressed and OCR; the websites are automatically crawled; some companies provide their contents as a digital stream directly connected to PressIndex).

All those resources are filtered daily to only keep the relevant articles. At the end of this filtering step, still 50,000 to 100,000 resources per day are stored in the PressIndex's content management system.

3.2 Acquiring knowledge

After filtering the resources, they are automatically parsed to instantiate the PressIndex knowledge base with the relevant knowledge discovered in the documents. That knowledge acquisition process is achieved by the OntoPop platform [4] which integrates the information extractor IDE™ with the ITM™ ontology repository. The IDE™ is a finite-state transducer, which uses natural language processing methods to parse a textual resource for tagging it, producing a conceptual tree as shown in **Fig. 3**.

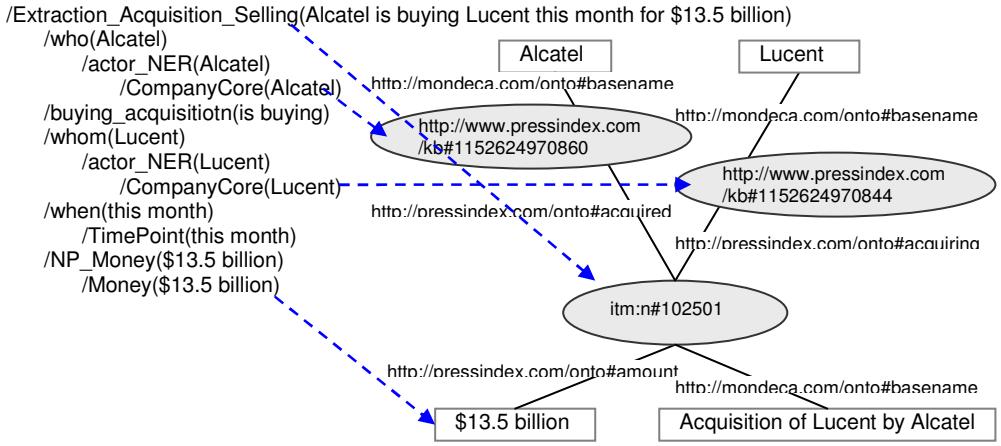

Fig. 3. Example of mapping an extract of a conceptual tree (left) to a RDF graph (right)

Then, OntoPop maps the produced tags with the ontology classes and relations defined in ITM to populate the ontology and/or to create RDF semantic annotations [5]. The right part of **Fig. 3** provides an example of a RDF graph generated from the conceptual tree on the left. As you can see, some tags are easy to map, such as the

Companies, but others can raise in-depth issues such as the date conversion into a knowledge base format to be further exploited by search functionalities for instance.

3.3 Aggregating, controlling & inferring knowledge

Before creating the instances in the ITM knowledge base (KB), OntoPop checks if the information already exists in the KB to avoid duplicates. If a corresponding instance is found, the new information is aggregated to it. Otherwise a new instance is created. OntoPop also controls the conformity to the ontology model (classes, cardinalities, domain and range properties, etc.). Simultaneously, an alert system listens to every creation of instances of a customized set of classes and relations. Consequently, each time an instance of one of those classes or relations is created, a copy of that instance and its entire relational context (all instances in relation with the former one) is sent to a CoGITaNT inference engine.

The CoGITaNT API [3] is a free library designed for the development of conceptual graph-based (CG) inference engines. The Conceptual Graph formalism (CG) (more precisely the SG-family [7]) has been chosen to provide inferences and validation services on the ITM knowledge for the following reasons:

1. Although the ITM knowledge may be serialized in several languages as RDF or OWL, the internal knowledge representation is based on Topic Maps formalism which (as RDF) might be defined as graphs (from graph theory). Thus, the translation from TM (or RDF) to CG is easier to define (as done in [6] for ITM Topic Maps, and in [8] for RDF).
2. Graph theory provides efficient algorithms to compute CG inferences.
3. Graph-based knowledge representation formalisms are intuitively understandable for the end user.
4. CG has formal semantics in first order logic (thus, unambiguous semantics) and provides querying and inference mechanisms equivalent to logical deduction. This is important to be sure of what exactly the knowledge means and how it is controlled.

For our concern, the ITM inference engine developed using CoGITaNT has the following characteristics. After a topic map created in ITM was transformed in a CG and added in the engine, the CoGITaNT engine KB has to be kept:

- Normalized: all individuals with same name are merged in a single one.
- Saturated with regard to the rule set: all knowledge deduced from the asserted part and the rules has to be explicitly and permanently represented in the engine base (for more details on rules and their applications, see [7]).
- Valid with regard to the constraints: the whole KB (asserted and inferred part) has to satisfy the constraint set.

After adding new asserted knowledge, if a constraint is broken because of the existence of an inconsistent pattern in the asserted or inferred part of the base, all inferences are cancelled and the added knowledge is rejected. The impact in ITM is a new property for the checked instances with value either to ‘valid’ or ‘inconsistent’, so that the user can easily correct the invalid instances. He can also visualize the inferred knowledge through a graphical interface.

At this time, we can define constraints as banned patterns (i.e. patterns not to be found in the KB) like a constraint of maximum number of relations for an instance or a constraint of acyclicity for a relation (no cycle creation), see **Fig. 4**. Soon, the engine will be able to manage mandatory knowledge constraints in order, for example, to control the respect of minimum cardinalities.

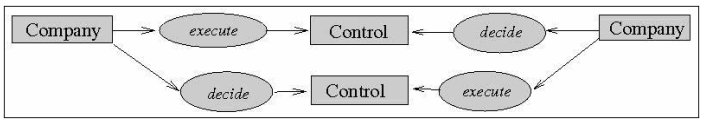

Fig. 4. Acyclicity constraint on relation "Control"

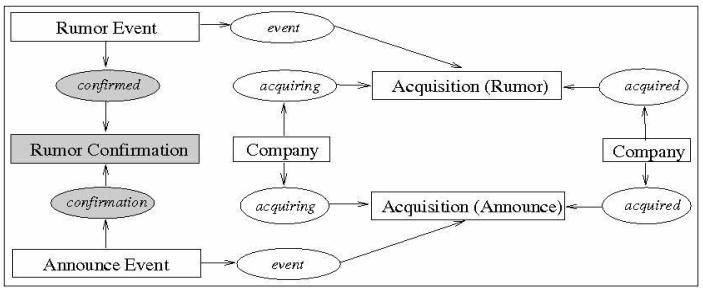

Fig. 5. Confirmation rule

Engine inference capabilities are used to discover new knowledge from the one extracted by the indexation process. For example, as shown in **Fig. 5**, if a 'company acquisition rumor' relation between two company instances was extracted and added to the engine and if an 'announcement' of the same acquisition is extracted from another document, a rule is set in CoGITaNT to deduce that the 'acquisition announcement' 'confirms' the 'acquisition rumor'. Then, we can query all 'confirmed' acquisitions for economic survey purposes.

3.4 Searching, Organizing and Publishing the knowledge content

Once the data have been filtered and controlled, the ontology model is used to provide enhanced search services. Search results can then be selected, sorted, organized and exported or saved, thus allowing the creation of new aggregated content into the knowledge base. Besides, the content stored in this knowledge base is intended to be published on a publicly accessible website. Semantic information is thus exploited to provide services at two levels: during the content aggregation process in the ITM-PressIndex application and within the published content on the Press Index website.

3.4.1 Semantic-driven search

Two different approaches are used to offer enhanced search services making use of the semantic information available. The first approach is based on class templates,

which can be defined from the ITM basic ontology. Each class template defines what type of attributes an instance of this class can have and what type of associations it can establish with other instances of the knowledge base. The multicriteria search interface used in the ITM-PressIndex application enables to select a class and to display only the types of attributes and semantic relations that are relevant search criteria for the instances of the considered class
The second approach relies on the PressIndex Ontology and will be used on the Press Index website. The principle is that of a faceted browser: it uses the ontology classification to offer different types of search filters ("location", "actor", and states of fact, i.e.: "announcement", "rumour", "fact"), which are the different search "facets" displayed to the user : the returned results are those fulfilling all filter restrictions, and the filter options are automatically re-evaluated according to taxonomic dependencies.

3.4.2 Sorting and organizing search results : making use of semantic information to assist the user in the creation of new aggregated content

Search results of interest to the user can be selected and manipulated by using two different modules:

- The aim of the first module is to allow the user to sort quickly a selection of items (e.g. to publish a quick press review of a company's new partnerships sorted by date) and either save it in a folder in the knowledge base or export it in different formats for direct use. The sorting module uses class attributes (defined in the class templates) as sorting criteria. A mapping can be done so as to gather class attributes under a same sorting criteria (i.e. the class attributes "fact date", "announcement date" and "rumour date" are mapped with the "event date" sorting criteria). A default parameter can be set to define in which order the different sort criteria are to be applied but the user is also free to define this order by himself. .
- The organization module allows the user to re-arrange the items of his selection according to his own editorial criteria, through the creation of a publication structure. The different sections and sub-sections of the publication created by the user, all as well as the selected knowledge objects, can be moved and reorganized by drag-and-drop When the publication structure is exported in XML and HTML, all the information attached to the different knowledge objects is then retrieved and inserted into the publication structure.

3.4.3 Publishing the content of the knowledge base

For editorial purposes, all the information contained about a knowledge object or a publication unit can be exported in an XML format which is structured into different types of "set" elements (identification, description, categorization, related objects, metadata). These sets offer a first level of data organization to facilitate information display, all as well as search and navigation between published elements.

For knowledge exportation purposes, the information contained in the database can also be exported in XTM and OWL (plus RDF, soon to be available).

4 Experimentations

Up-to-date, the PressIndex application is still under development but it is planned to be released on the PressIndex portal by the end of this year. The architecture is mainly implemented, the final user interfaces for the clients are to be designed and further inference rules needs to be defined as well. Nevertheless, we already ran a sample of 10,000 documents: 5,000 in French and 5,000 in English.

As first results, the PressIndex was automatically populated with 25,440 instances of classes, 17,535 attributes were added to those instances and 1,844 relations links some of the instances. Moreover, the inference engine inferred 10,620 new instances. We still need to further evaluate the application, especially by calculating the precision and recall measures.

5 Conclusion

The added-value of the semantic web technologies will allow the new PressIndex's press clipping application to:

- Have an automatic information extraction process from the capitalization of that information to its diffusion;
- Transform a media press stream into a knowledge capitalization process with synthesis notes organized by companies, products, facts, etc.;
- Construct a new content from the information and knowledge capitalization;
- Create multilingual normalized information, independent from the language of the information sources.

References

1. Amardeilh F., Francart T: A Semantic Web Portal with HLT Capabilities, In Veille Stratégique Scientifique et Technologique (VSST04), Vol. 2, Toulouse (2004) 481-492
2. Grivel L., Guillemin-Lanne S., Lautier C. and al.: La Construction de Composants de Connaissance pour l'Extraction et le Filtrage de l'Information sur les Réseaux. In 3ème Congrès du Chapitre Français of International Society for Knowledge Organization, Paris (2001)
3. Genest D.: Cogitant. http://cogitant.sourceforge.net, (1997)
4. Amardeilh F., Laublet P., Minel J-L.: Document Annotation and Ontology Population from Linguistic Extractions. In proceedings of Knowledge Capture (KCAP05), Banff (2005)
5. Amardeilh F.: OntoPop or how to populate an ontology and annotate documents, In proceedings of the "Mastering the Semantic Gap from IE to Semantic Representation" Workshop at ESWC06, Budva, (2006)
6. Carloni O., Leclère M., Mugnier M-L.: Introducing Graph-based Reasoning into a Knowledge Management Tool: an Industrial Case Study, In proceedings of the "IEA/AIE 2006" (2006) 590-599
7. Baget J.-F., Mugnier M.-L.: Extensions of Simple Conceptual Graphs: the Complexity of Rules and Constraints, In proceedings of JAIRS, (2002) 16:425-465,
8. Baget J.-F.: RDF Entailment as a Graph Homomorphism, In proceedings of ISWC 05 (2005)