IDP v.0.3 vision

Max Talanov

June 11, 2008

1 Outline

Contents

2 Overview

Purpose of this document is to create some overall picture of 0.3 version of IDP project.

3 Main features

3.1 Semantic network (SN)

Main enhancement of 0.3 version is migration from text mining to concept mining, that is mainly use of semantic network for data representation of processed documents and heuristics for annotation and normalization.

3.1.1 Options

There are several options to use open-source Java projects:

KAON is an ontology management infrastructure targeted for business applications,
(+) good GUI for ontology, TextToOnto, LGPL;
(-) querying and reasoning is still experimental, RDF only, last update 2003

TextToOnto The aim of TextToOnto is to support developers in the ontology construction process by applying text mining techniques. For this purpose it builds on KAON.

KAON2 - is an infrastructure for managing OWL-DL, SWRL, and F-Logic ontologies.
(+) reasoning implemented, Text2Onto;
(-) commercial for non educational purposes

Text2Onto A Framework for Ontology Learning and Data-driven Change Discovery.
(+) LGPL, OWL

NeOn State of the art application for using ontologies for large-scale semantic applications in the distributed organizations. Particularly,improve the capability to handle multiple networked ontologies that exist in a particular context, which are created collaboratively, and might be highly dynamic and constantly evolving.
(+) integration with Text2Onto,querying;
(-) commercial, no Reasoning implemented

pellet is an open source, OWL DL reasoner in Java that is developed, and commercially supported, by Clark & Parsia LLC. OWL is an international, web standard produced by the W3C.
(+) free, integration with Protege, Jena; - tabular calculus.

Protege is a free, open source ontology editor and knowledge-base framework.
(+) free, OWL.

Mulgara is a scalable RDF database written entirely in Java. Mulgara is an Open Source fork of Kowari.
(+) free

Jena is a Java framework for building Semantic Web applications. It provides a programmatic environment for RDF, RDFS and OWL, SPARQL and includes a rule-based inference engine.
(+) SPARQL, OWL, free.

conceptNet + minorthird conceptNet is integrated with montylingua for NLP, it is possible to integrate semantic network of conceptNet with minorthird to extract concepts from texts.
(+) free, minorthird is integrated with idp;
(-) not OWL.

We have to have some research to do to decide what is proper base for further development. At the moment it seems to me that TextToOnto/Text2Onto + pellet/Jena + Protege is best for our purposes, because they are free and modern.

3.2 Reasoning, querying and analysis

This should be implemented on base of SN. Several reasoning servers, *pellet*, *KAON2* and *Jena* provide option for reasoning and querying. This should be base for busyness rules description and analysis as well as reports.

3.3 Subst2 (normalization based on SN)

There should be several steps

- Add multiple rules use for normalization, add probability consideration of different rules.
- Make normalization able to learn based on positive examples.
- Make normalization rules description based on OWL(SN).
- Make use of SN described heuristics, ex.: for date ranges description.
- Generalize learned rules to heuristics(optional).

3.4 Non text documents

We have to start processing different formats of documents:

- DOC
- OXML
- ODF
- PDF
- HTML pages, not only local
- XML
- RSS channel

For Microsoft formats it is possible to use POI, for PDF PDFBox.

3.5 JBoss

It could be good option to use application server of client server messaging and multiple users service, as well as web-services use.

3.6 Third parity application updates

We should move to new versions of applications:

- minorthird 20080414
- JGAP 3.3.3

4 Milestones

Some vision of sequence of steps to Beta.

1. Various document formats processing.
2. Analise of SN platforms servers -> *estimates to implement document integration.*
3. Analise of SN reasoning servers -> *estimates to implement analytical environment.*
4. Implementation of SN document representation.
5. Concept mining implementation.
6. Reasoning integration.
7. Subst2.
8. JBoss migration analysis and implementation.
9. Multiproject enhancement.
10. minorthird and JGAP new versions migration.

This is very rough sequence and could not be considered as guidance or a project plan.