KP-Lab
KNOWLEDGE PRACTICES LABORATORY

Implementation of Text Mining Services in the KP-Lab Project

11/12/08

Marek Schmidt

KP-Lab
KNOWLEDGE PRACTICES LABORATORY

Knowledge Practices

- Knowledge Practice
  - An innovative process, routine, or procedure of working with knowledge. Knowledge practices represent socially constituted, rather than merely individual activities.
- Trialogical Learning
  - Learners are collaboratively develop, transform, or create shared objects in a systematic fashion.
  - Concentrates on the interaction through developing these common, concrete objects

KP-Lab
KNOWLEDGE PRACTICES LABORATORY

Concept Mapping Knowledge Practice

- Concept Map
  - Diagram showing the relationships among concepts.
  - Graphical tools for organizing and representing knowledge.
  - Used to stimulate the generation of ideas, to aid creativity.
- Usage Scenario
  - Students are given research materials on a given topic, asked to collaboratively create a concept map.
  - Learning is stimulated by discussions of/process of creation/the concept map itself

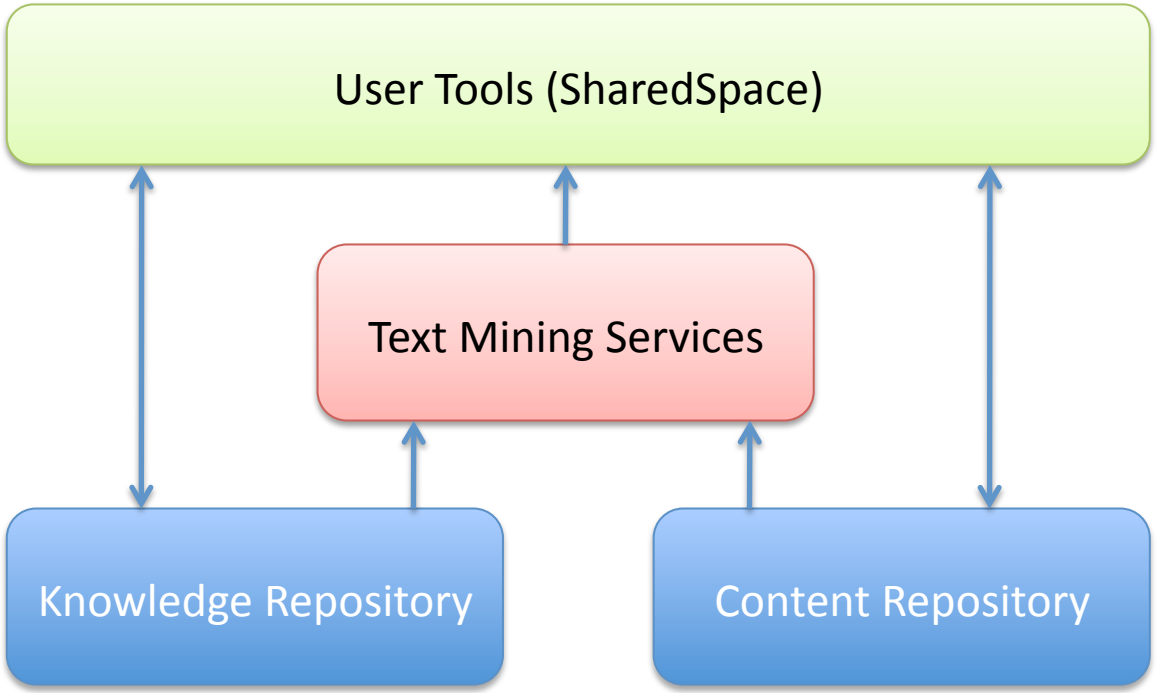
KP-Lab
KNOWLEDGE PRACTICES LABORATORY

Application of Text Mining Services

- Service shall provide suggestions for
  - New concepts
  - New related/similar concepts
- Application of Ontology Learning methods

KP-Lab
KNOWLEDGE PRACTICES LABORATORY

General Architecture



KP-Lab
KNOWLEDGE PRACTICES LABORATORY

Text Mining Services

- Classification Service
- Clustering Service
- Ontology Learning Service
  - findConceptCandidates ()
  - findRelationCandidates (concepts)

KP-Lab
KNOWLEDGE PRACTICES LABORATORY

Ontology Learning Services

1. Automatic Term Recognition
2. Syntactic Patterns
3. Statistical Methods for Similarity

KP-Lab
KNOWLEDGE PRACTICES LABORATORY

Automatic Term Recognition

1. Minipar Parser
   - Produces POS tags and a dependency tree
2. Term candidates extraction from the dependency tree
   - Set of patterns (nouns, nouns with modifiers)
   - All sub-terms are extracted
3. Scoring of term candidates
   - Experimented with several scoring functions
   - Termhood (TfIdf, Weirdness, LR test), Unithood (C/NC-Value)
   - Background frequencies from general corpus (Gigaword)

KP-Lab
KNOWLEDGE PRACTICES LABORATORY

Syntactic Patterns

1. General idea
   - Map semantic relations as a set of syntactic patterns (like Hearst Patterns)
   - Create set of patterns from seed patterns by computing paraphrases
2. Result
   - Does not work very well in general setting
   - Keep patterns for is-a and general S-V-O, S-V-P-P

KP-Lab
KNOWLEDGE PRACTICES LABORATORY

Statistical Methods for Similarity

1. Extract Co-occurrences on different levels
   - Different levels produce different types of relatedness
   - Syntactic – term to syntactic features (modifiers, verb, subject, object, …)
   - Sentence, Document – term sentence / term document matrix
2. Compute similarity (Dekang Lin)

$$sim(A,B) = \frac{\log P(common(A,B))}{\log P(description(A,B))}$$

KP-Lab
KNOWLEDGE PRACTICES LABORATORY

Lin Similarity of Words

$$sim(t_1, t_2) = \frac{2 \times I(F(t_1) \cap F(t_2))}{I(F(t_1)) + I(F(t_2))}$$

$$I(S) = -\sum_{f \in S} \log P(f)$$

$$P_{MLE}(f) = \frac{\left|\{t \mid f \in F(t)\}\right|}{\left|\{t\}\right|}$$

KP-Lab
KNOWLEDGE PRACTICES LABORATORY

Similar Background Terms

- We compute similarity also on background terms
  - To help define the meaning of some new domain-specific term

KP-Lab
KNOWLEDGE PRACTICES LABORATORY

Extracting slipped terms

- Problem
  - User asks for related terms to term not extracted during pre-processing
- Solution
  1. Fulltext index on sentences
  2. Match all verbatim occurrences of the term
  3. Extract features from these occurences

KP-Lab
KNOWLEDGE PRACTICES LABORATORY

Implementation

- Extraction Core
  - Python, Minipar, sqlite, Xapian
  - Java, GATE, text2onto
- Web Service front-end
  - JBoss Seam
  - Aperture (Nepomuk, Content Extraction)

KP-Lab
KNOWLEDGE PRACTICES LABORATORY

Evaluation

- Automatic Term Recognition
  - retrieval of user-annotated keywords (LT4el, Genia)
- Related Terms
  - Problematic
  - User level concept mapping tools not ready yet

KP-Lab
KNOWLEDGE PRACTICES LABORATORY

References

- Dekang Lin
  - An Information-Theoretic Definition of Similarity (1998)