Expertise Mining: Domain Specific Concept Extraction for Expert Search

Georgeta Bordea

gb@deri.org
Expertise Mining

- **Extraction of expertise topics**
  - Domain specific term extraction using taxons
  - Ranking aggregation
  - Semantic grounding on the Linked Data Cloud

- **Expert finding**
  - Relevance (TFIRF, probabilistic language modelling)
  - Advanced expertise measures based on a domain taxonomy (area coverage, application in different domains)

- **Expert profiling**
Domain Specific Expertise Topic Extraction

- **Saffron baseline**
  - Candidates - syntactic patterns and a web filter
    \[
    s(w) = |w| \times f(w) + \alpha \times E(w), \alpha = 3.5
    \]

- **Taxon extraction**
  - Candidates - content bearing words with high distribution
    \[
    s(t) = \sum_{k \in K} PMI(t, k)
    \]

- **Domain semantic relatedness (DSR) – post ranking**
  \[
  DSR(w) = \sum_{t \in T} PMI(t, w)
  \]
Term Extraction Baselines

- **NCValue**

\[ C_{value} = \log |w| \times f(w) - \text{Avgf}(e) \]

\[ NC_{value} = 0.8C_{value} + 0.2\sum f_a(c) \times \text{weight}(c) \]

- **TermExtractor (Velardi)**
  - Domain pertinence – contrastive corpora (Gutenberg, OANC)
  - Domain consensus
  - Lexical cohesion

\[ s(w) = \alpha \times DR + \beta \times DC + \gamma \times LC \]
Expertise Topic Extraction

Datasets

- **Krapivin**
  - Computer science
  - 2304 documents with author assigned keyphrases and ACM index terms

- **Genia**
  - Biomedical domain
  - 1,999 Medline abstracts with manually annotated terms

- **FAO (Food and Agriculture Organisation)**
  - 780 documents with Agrovoc terms assigned by professionals
DSR Evaluation on the Krapivin Dataset (F-score top 30)
DSR Evaluation on the Genia Dataset (F-score top 30)

![Graph showing performance comparison between Baseline, DSR, NCValue, and Velardi.](image-url)
DSR Evaluation on the FAO Dataset (F-score top 30)
Expert Finding Datasets

- **Scientific publications**
  - Semantic Web Conference Corpus (3290 full publications and abstracts)
  - ACL Anthology Reference Corpus (10921 full publications)

- **Workshop committee members**
  - Semantic Web Corpus: 102 workshops, 126 expertise topics, 1003 experts, 2.7 topics/expert
  - ACL Corpus: 248 workshops, 286 expertise topics, 2009 experts, 3.3 topics/expert

- **UVT Expert Collection**
  - 1168 experts, 1491 expertise areas, expertise profiles provided by the employees themselves
Ongoing Work

- Controlled vocabulary indexing for expert finding
  - Joint work with Hugues

- Building hierarchies of concepts for expert finding

- Concept consolidation
  - Joint work with Hasti