Continuous Multi-Source Information Gathering and Classification

Ralf Steinberger, Bruno Pouliquen, Stefan Scheer and António Ribeiro
European Commission
Joint Research Centre
Ispra, Italy
www.jrc.cec.eu.int/langtech

Motivation

• Looking for documents on ‘terrorism’ on the Internet

• How do you find what you need?
  – Use your favourite search engine…
Motivation

• Looking for documents on ‘terrorism’ on the Internet

• How do you find what you need?
  – Use your favourite search engine…
  – Too much information
  – Too many information sources

• The usual information retrieval tools are not enough…
  – We need intelligent tools to rank and sort documents

Outline

• Motivation
• Domain
• OSILIA System
  – Overview
  – Description of Parts
• Evaluation
• Conclusions
• Current and Future Work
Domain

‘Fight against Internet Fraud’

• Attacks on Computers
  – e.g. viruses, Trojan horses

• Attacks on Internet
  – e.g. worms, denial-of-service attacks, hacking of web sites

• Computer-based Fraud
  – e.g. cracking, credit card stripping, attacks on financial encryption

• Crime using Internet
  – e.g. paedophilia, terrorism, instructions for bomb-making

OSILIA System

• Gathers information from multiple sources

• Can be tuned to particular subject domains
  – select sources
  – define a set of search words

• Cleans irrelevant information from documents

• Filters relevant documents
  – a document is downloaded if it matches a query

• Classifies documents
Overview

Crawler

- **Information Sources**
  - 20 news and meta-news sites
- **Languages**
  - Two: English and German
- **Source parameter files**
  - starting web site address
  - search depth
  - sections not to be crawled
- **Language parameter files**
  - Boolean search term query
  - weights per word
  - threshold for document storage
Cleaner

Solicitor admits wills fraud

Cleaner

Solicitor admits wills fraud
Filter

- Apply the subject domain query **again**
  - documents downloaded due to search words **outside** the article

Classifier

- **Lemmatiser**
  - to improve the keyword identification

- **Keyword Assignment**
  - log-likelihood test
  - chooses relevant words for storage in the database
  - searches with keywords retrieve relevant documents

- **Simple classification technique**
  - weights **per** word
  - **threshold** for assignment to a class
  - ranks documents
Search Interface

- Meta-data extracted
  - Dublin core ‘elements’
    - Title
    - Date
    - Source
    - …
  - Other ‘elements’
    - Language
    - Subject
    - Geographical references
    - …
- Queries on some ‘elements’

Evaluation

- Compared documents gathered

<table>
<thead>
<tr>
<th>Subject Type</th>
<th>Number of documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attacks on Computers</td>
<td>500</td>
</tr>
<tr>
<td>Attacks on Internet</td>
<td>327</td>
</tr>
<tr>
<td>Computer-based Fraud</td>
<td>459</td>
</tr>
<tr>
<td>Crime using Internet</td>
<td>130</td>
</tr>
<tr>
<td>Total</td>
<td>1416</td>
</tr>
</tbody>
</table>
Evaluation

• Evaluator
  – User with background in journalism
  – Precision: 95% of documents covered the subject domain
    • 5% error: ‘An internet report on child abuse in local authority homes’
  – Recall: all major issues also found in the automatic gathering
  – Only about 20% of the articles brought previously unknown facts
  – Not much information than what is found in specialist publications

Conclusions

• A system which
  – Gathers information from multiple sources
  – Can be tuned to particular subject domains
  – Cleans irrelevant information from documents
  – Filters relevant documents
  – Classifies documents

• 95% of documents covered the subject domain
  – Only about 20% of the articles brought previously unknown facts
Current and Future Work

• Filtering
  – identification of duplicates

• Assignment of Eurovoc thesaurus descriptors
  – a set of descriptors translated in 11 languages
  – allows cross-language document similarity calculation
  – users identify a document of interest
  – get a list of similar documents

Current and Future Work

• Named-entity extraction
  – users interested in names of people, places and organisations
  – this information can be made searchable

• Language identification
• More sophisticated relevance ranking techniques
• An e-mail and sms notification facility