A Method for Collecting Unspecified Researchers' Homepages Utilizing Web-Specific Features

Yuxin Wang (SOKENDAI) Keizo Oyama (NI1, SOKENDAI)

What to do?
- create a high quality collection of unspecified researchers' homepages with high recall and high precision

Focus on:
- Unspecified researchers' homepages:
  - No name, organization or fields...
  - Homepages contain information on research activities
  - Publication list
  - Research topics, projects, achievements, etc.
  - Major, career, etc.
  - Affiliation, society, etc.
- Utilize both the content-based and web-specific features

Why we need collecting of researcher's homepages?
- Traditional search engine like Google:
  - Keyword(s) is necessary to specify the researcher
- DB of researchers like ReaD or Web of knowledge:
  - Collection is limited because
    - Not all the researchers registered their data
    - Not necessary to register their own homepages
- Collection of researcher's homepages — our approach:
  - Can be used as a complement of researcher's resources for compiling and maintaining of researcher's database
  - Can be used for domain specific search engine
  - Can be used to structure researcher related information services

Overall structure of the method
- Rough filtering of the possible web page:
  - to gather homepages with high recall as well as high precision by
    - defining keyword-lists
    - learning rules
    - building page group model utilizing Web-specific features
- Accurate evaluation of the web pages:
  - to calculate a likelihood score:
    - positive, negative, or possibly positive
  - Precise selection of the correct homepage
    - to find the best thresholds
    - between positive and possible positive
    - between possible positive and negative

Rough filtering of the possible web pages
- Corpus
- Candidate pages
- Keyword lists
- Rule set
- Generating rules with high-recall
- Rule set
- Generating rules with high-precision
- Precise selection
- Final pages

Using web-specific features
- Page group models
  - Out-link information
  - In-link information
  - All links
  - All in-links
  - Same domain
  - Out-links in same domain
  - Same site
  - Out-links in same site
  - Out-link in sup1d2d
  - Out-link in sup1d2d
  - Out-link in sup1d2d
  - Out-link in same site
  - All in-links

Experiment result
- Defined keyword-lists
- Performance comparison of different page group models

Conclusions and future work
- What we have done:
  - Investigated Ad hoc method for defining keyword-lists
  - Investigated statistical analysis for evaluating the keyword-lists
  - Built page group models for utilizing web-specific features
  - Tried to apply SVM for single page content-based model
- What we will do:
  - Modify present content-based keyword-lists with statistical method
  - Use decision tree to learn efficient rule set for collecting researchers' homepage from all data
  - Refine page group model
  - Find a method to utilize other web-specific features
  - Evaluate pages with a classifier using web-specific in addition to single page content features
  - Investigate a method for precise selection of the correct homepages