Lexical Profiling of Existing Web Directories to Support Fine-grained Topic-Focused Web Crawling

Mark Greenwood
Goran Nenadic

School of Computer Science,
The University of Manchester

Outline

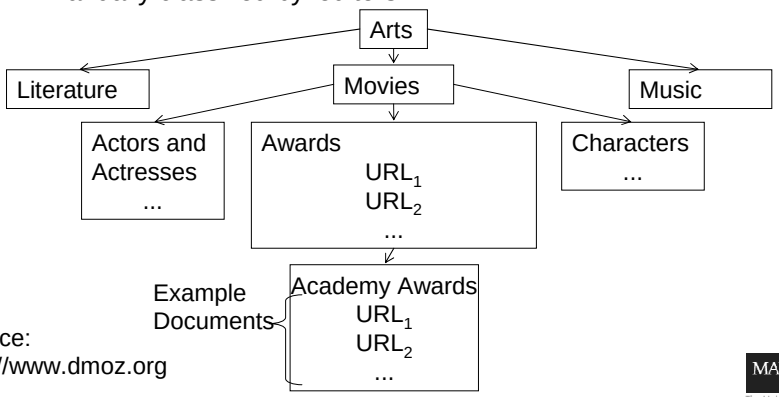
- Motivation
- Using Web Directories
- Lexical Profiling
- Classification
- Experiments
- Results
- Conclusion

Motivation

- The Web
  - Large document collection
  - Encyclopaedic
- Topic specific corpus creation
  - Creating indexes of documents relevant to a specific topic from the Web
  - Useful for search within a domain or as part of larger IR system
- Term-level feature based model
- ‘Background knowledge’ inclusive model
- Small training sets

Using Web Directories

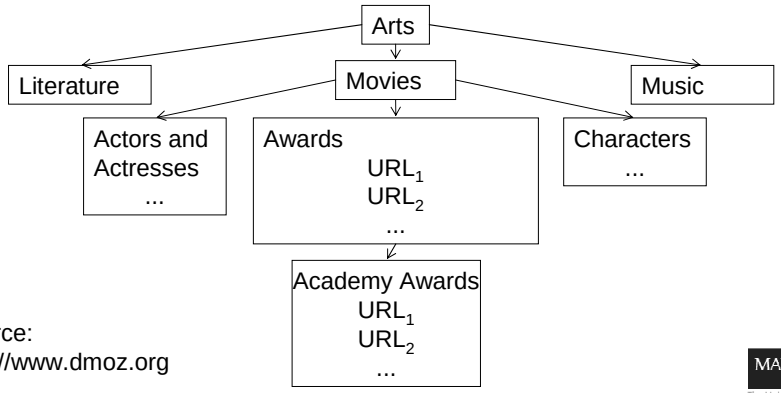
- Web Directories used for example documents
  - Pages listed under subject headings
  - Headings organised taxonomically
  - Manually classified by ‘editors’



Source:
http://www.dmoz.org

Using Web Directories

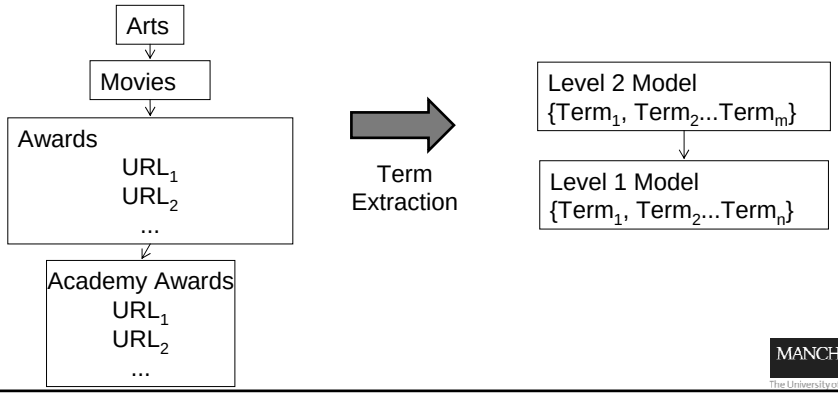
- Existing methods
  - Concatenate levels below the target topic to form large training sets
  - More general topics



Source:
http://www.dmoz.org

Using Web Directories

- Term-level features + Lexical Profiling
  - Create model of finer-grained topics
  - Include background knowledge



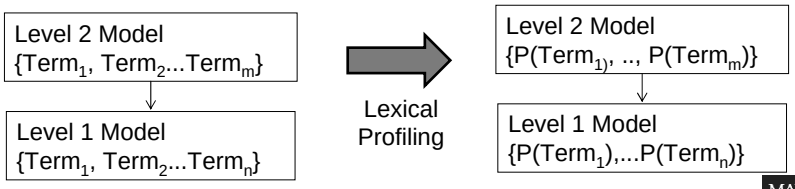
Term Extraction

- Automatic Term Recognition
- TerMine (http://www.nactem.ac.uk/termine)
  - Web Service
  - Based on C-Value *Termhood* measure
  - Statistical and Linguistic pattern analysis

Lexical Profiling of Terms

- Expansion of terms
  - Left-linear combinations of word-level substrings
  - Richer model, allows for generalisation

Term	Lexical profile (P(Term))
microarray core facility	{microarray, core, facility, microarray core, core facility, microarray core facility }



Estimating Page Relevance

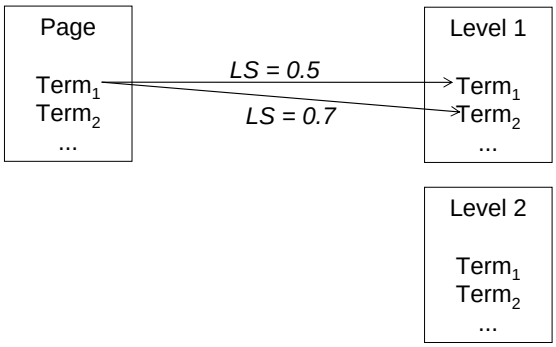
- Based on Lexical Similarity between terms

$$LS(t_1,t_2)=\frac{|P(h_1)\cap P(h_2)|}{|P(h_1)|+|P(h_2)|}+\frac{|P(t_1)\cap P(t_2)|}{|P(t_1)|+|P(t_2)|}$$

- Weight given to terms with
  - Longer nested constituents
  - Common heads
- Pages crawled compared to each level of model

$$PageScore(Page,Level_n)=\frac{\sum\max\{LS(p,Level_n)\}}{|Page|}$$

Estimating Page Relevance - Example



PageScore(page,$Level_1$) = 0.4
PageScore(page,$Level_2$) = 0.7

Page assigned Level 2

Classification - Links

- Overall score

$$LinkScore = LinkContextLevel \times LinkContentLevel \times PageLevel$$

- Link Context
  - Words either side of anchor text
  - Left linear combinations of words treated as term candidates

$$LinkContextScore(P(context), Level_n) = \max\{\max\{LS(p, Level_n)\})\}$$

- Link Content
  - Anchor text

$$LinkContentScore(text, Level_n) = \frac{\sum LS(text, Level_n)}{|Level_n|}$$

Link Classification - Example

"National Center for Genome Resources The Tree of Life Home Page - Useful for tracking down taxonomic"

- Term Representations:
  - Content = "Tree of Life Home page"
    - Treated as term candidate
    - Level assigned for which average LS is highest
  - Context = {"national", "center", "for", "Genome", "Resources", "national center", "center for", "for Genome", etc.}
    - Individual word combinations treated as term candidates
    - Maximum lexical similarity in current level found for each term candidate
    - Largest score used to represent the score
    - Level assigned for which score is highest

Experiments

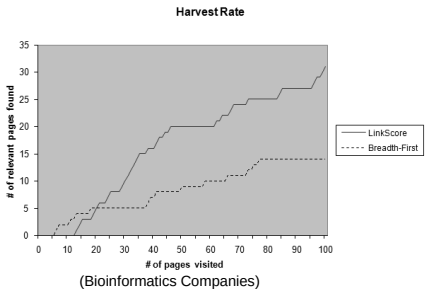
- Topics chosen
  - Science/Biology/Bioinformatics/Companies
  - Society/Issues/Science & Technology/Biotechnology/Genetics/GM Food
- 100 pages gathered
- Link Prioritisation compared to Breadth-First

Taxonomy Level	Documents	Terms
Companies	4	49
Bioinformatics	5	103
Biology	2	84
Science	0	0

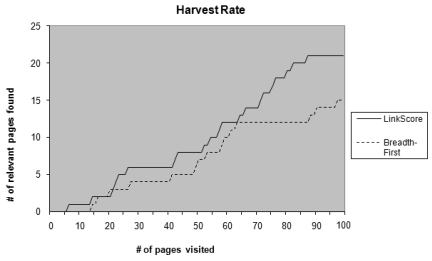
Taxonomy Level	Documents	Terms
GM Foods	4	34
Genetics	4	9
BioTechnology	4	62
Science and Technology	4	136
Issues	0	0
Society	0	0

MANCHESTER 1824
The University of Manchester

Results



(Bioinformatics Companies)

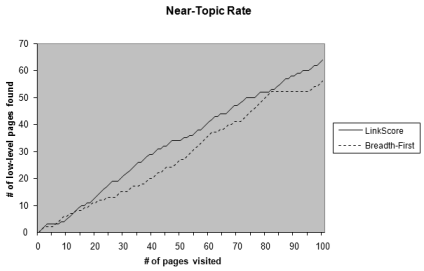


(GM Foods)

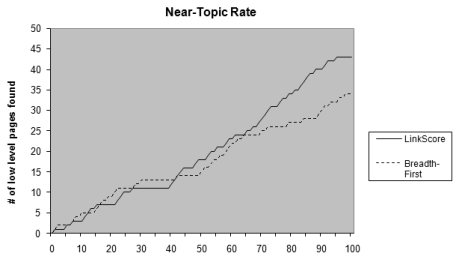
- Harvest rate improved
  - Bioinformatics – double the relevant documents
  - GM Foods – 40% more relevant documents
  - Using ‘background knowledge’ from taxonomy can improve efficiency

MANCHESTER 1824
The University of Manchester

Results



(Bioinformatics Companies)



(GM Foods)

- Near-topic rate improved
  - Number of pages classified in level 1 or 2
  - Prioritised crawl has improved ability to stay close to desired topic

Conclusion

- Profiling existing corpora
  - Can provide fine-grained models for Topic-focused Web Crawling
  - Show improvement in harvest rate
  - Can provide 'background knowledge' to aid keeping the crawl on track
- Future Work
  - Incorporate *Termhood* scores, adding more weight if more 'significant' terms match
  - Genetic algorithm weighting of linkScore – Are there 'more telling' features of links?