Automatic Subject Heading Assignment for Online Government Publications Using A Semi-supervised Machine Learning Approach

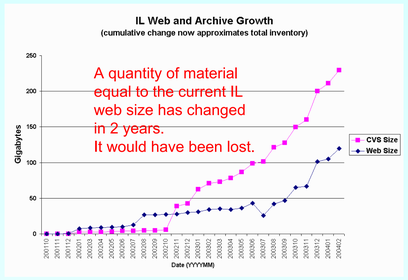
Xiao Hu Larry S. Jackson Sai Deng Jing Zhang
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

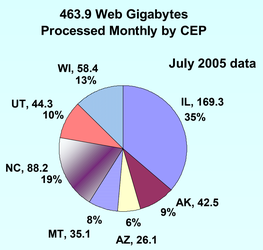
The Government Document Problem Space

Preserving Electronic Publications (PEP)
- http://www.isrl.uiuc.edu/pep
- Archive government documents published on the web
- Use share-/free-ware modules, open standards, and other cost-containment measures

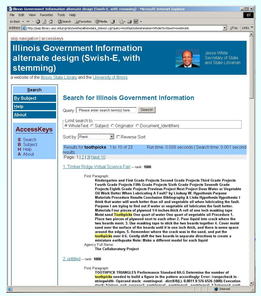
Capturing Electronic Publications (CEP)
- Support multi-state deployment of our web archiving facility
- Harvest government information for IL, NC, MT, AZ, AK, UT, WI

Electronic Documents Initiative (EDI)
- Provide permanent retention and web access to "official" State publications existing in electronic form





Illinois Government Information (IGI) search engine
- Provide users with full access to online government information
- Support searching by subject, website, originator, etc.



http://findit.lis.uiuc.edu
- ✓ Fully accessible
- ✓ Uses metadata byproducts of CEP web archival, delivered as EDI-style surrogates
- ✓ All open-source (CEP & Swish-E)
- ✓ Far outperformed the existing IGI search engine (of the Find-It! series)

Problems:
- Lack of author provided metadata
  -- Text mining techniques to automatically extract metadata
- Organize documents with a hierarchy of subject headings
  -- "Topic Tree" : adapted from GILS tag set
  http://www.finditillinois.org/metadata/subjtree.pdf
  -- **Semi-supervised algorithm to generate subject heading**

Automatically Assigning Subject Headings

Semi-supervised Learning
- Manually labeling subject headings is very expensive
- Unlabeled examples are abundant
- Use unlabeled examples help assign subject headings

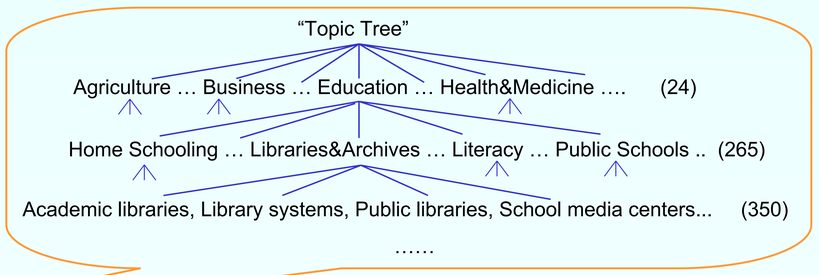
Expectation – Maximization (EM) Algorithm
- Assume unlabeled examples subsume same probability distributions as labeled examples

```
classifier = new Classifier (a small set of labeled examples);
while (! converge) {
    E-step: classifier.labeling (unlabeled examples);
    M-step: classifier.estimate (all examples);
}
```

Data Statistics				
	# of doc.	# of headings	# of unique terms	# of features
Training data	194	604	4,597	1,500 (by Info. Gain)
Testing data	145	410	--	

Classification Models	
Semi-supervised	Supervised
Expectation-Maximization (EM)	Naïve Bayes classifier
On a multinomial mixture model using logarithmical term frequency	

Evaluation Measures			
	Precision (P)	Recall (R)	F1 = 2PR / (P + R)
Macro-average	average the measures over all categories		
Micro-average	average the measures over all documents		



Other Techniques

Unsupervised Leading Methods
- SimpleKMeans:
  32% precision on the whole Illinois collection of 422,152 documents
- Pros: no need for labeled training examples
- Cons: limited precision

Collection-level default subject headings
- assign same subject headings to all documents in one website
- Pros: efficient
- Cons: coarse granularity of assigned subject headings
  Needs user studies to validate

Conclusions

- An example of applying a semi-supervised text categorization approach in a real-word practice
  - Assignment improvements are observed in experiments
  - More labeled training data may be needed to better demonstrate merits of the EM algorithm
- Experiments provide a reference to other projects working with online government information
- Working towards reducing the cost of subject heading assignment

Future Work

- Compare this approach with others
  - Effectiveness & Efficiency
- Perform a formal user needs assessment
- Test thoroughly on scalability
- Deploy in Illinois' IGI search engine

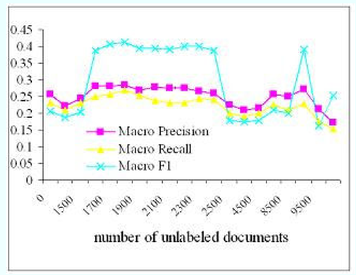


Figure 1: Macro averaging performance. The number 0 on the x-axis corresponds to the Naïve Bayesian classifier which doesn't need unlabeled documents.

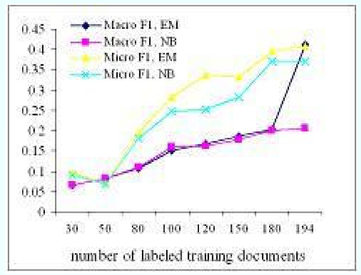


Figure 2. Learning curves of the EM and NB classifiers. In the EM classifier, the numbers of unlabeled documents are set as 10 times more than corresponding labelled ones.

Contact: {xiaohu, lsjackso, zhang20} @uiuc.edu , sai.deng@wichita.edu
Sai is currently working at Wichita State University Library, Wichita, KS