Web Genre Classification via Hierarchical Multi-label Classification

Vedrana Vidulin

Gjorgji Madjarov

Ivica Dimitrovski

Dragi Kocev

Contact: *vedrana.vidulin@irb.hr*

MAESTRA

LEARNING FROM MASSIVE, INCOMPLETELY ANNOTATED, AND STRUCTURED DATA

Jožef Stefan Institute,
Ljubljana, Slovenia

Ruđer Bošković Institute,
Zagreb, Croatia

Faculty of Computer Science and
Engineering, Skopje, Macedonia

Introduction

- **Web genre** represents form and function of a web page.
  - e.g. “Scientific” paper about the topic of text mining.
- Web pages may contain parts each of different genre: **multi-label classification**
- Web genres form **hierarchy**
  - e.g., “Prose fiction” and “Poetry” are both subgenres of genre “Fiction”
- State-of-the-art approaches mostly use single-label classification, while only few use multi-label classification but without exploiting hierarchical structure of web genres.
- Santini (2011) showed that flattening genres from different levels of hierarchy reduces classifier’s predictive performance.

Why Web Genre Hierarchy is not explored?

- Major obstacles were lack of:
  - Comprehensive genre taxonomy – a group of web genre experts could not agree about a single taxonomy (Rehm *et al.*, 2008)
  - Web-page-based corpora labelled with such a taxonomy,
  - Machine learning methods that are able to fully exploit the complexity of such data.
- Proposed solution:
  - Bypass manual construction of web genre hierarchy using **data-driven hierarchy construction** instead

Data

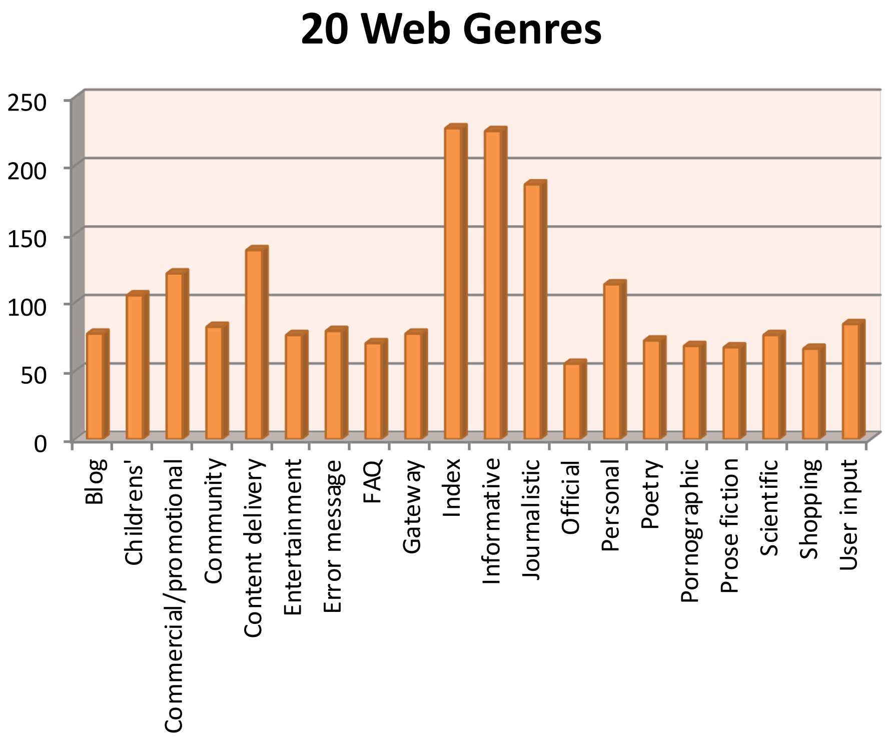
- Extracted from 20-Genre Collection **multi-label** corpus (Vidulin *et al.*, 2009)
- Corpus manually annotated by three independent annotators

2,491 features belong to four groups

Examples: 1,539 web pages

Surface features:	**Structural features:**	**Presentation features:**	**Context-URL features:**
Genre-specific words, Function words, Punctuation marks, Text statistics	Part-of-speech tags, Part-of-speech trigrams, Sentence types	Token type, HTML tags	Https, URL depth, Document type, Top-level domain, national domain, content words, etc.
Feature values expressed as ratios to eliminate the influence of page length.			Binary feature values representing presence of a property in web page URL

Genre Labels



From 1,539 web pages:

- 1,059 labeled with one genre
- 438 with two
- 39 with three
- 3 with four

1.34 labels per web page

Research Questions

- Which data-driven hierarchy construction method yields hierarchy of genres with best performance?
- Does constructing a hierarchy improves the predictive performance?
- Does constructing a data-driven hierarchy yields satisfactory results when compared with expert-constructed hierarchy?

The Choice of Hierarchy Construction Method

- Experimental setup:
  - Balanced *k*-means clustering with *k* of 2, 3 and 4
  - Model constructed using CLUS system for predictive clustering → predictive clustering trees (PCT) for hierarchical multi-label classification
  - 3-fold cross-validation
  - 8 example-based and 8 label-based evaluation measures

	HammingLoss	*Accuracy*	*Precision*	*Recall*	*Fmeasure*	*SubsetAccuracy*	*MicroPrecision*	*MicroRecall*	*MicroF1*	*MacroPrecision*	*MacroRecall*	*MacroF1*	*OneError*	*Coverage*	*RankingLoss*	*AvgPrecision*
HMC - manual hiear.	0.094	**0.276**	**0.327**	**0.341**	**0.334**	0.172	0.31	**0.33**	0.32	**0.424**	**0.296**	**0.297**	0.643	5.561	0.238	0.47
HMC - BkM (k=4)	**0.081**	0.261	0.31	0.3	0.305	**0.177**	**0.368**	0.291	**0.325**	0.368	0.262	0.284	**0.635**	**5.435**	**0.232**	**0.475**
HMC - BkM (k=3)	0.09	0.223	0.273	0.272	0.273	0.131	0.301	0.263	0.281	0.328	0.212	0.211	0.677	5.878	0.254	0.44
HMC - BkM (k=2)	0.084	0.206	0.247	0.247	0.247	0.127	0.328	0.24	0.277	0.361	0.205	0.227	0.682	5.956	0.259	0.433
MLC	0.111	0.136	0.172	0.165	0.168	0.073	0.165	0.163	0.164	0.063	0.1	0.065	0.83	7.955	0.36	0.317

Multi-branch hierarchy is more suitable for the domain

Impact of Hierarchy on Predictive Performance

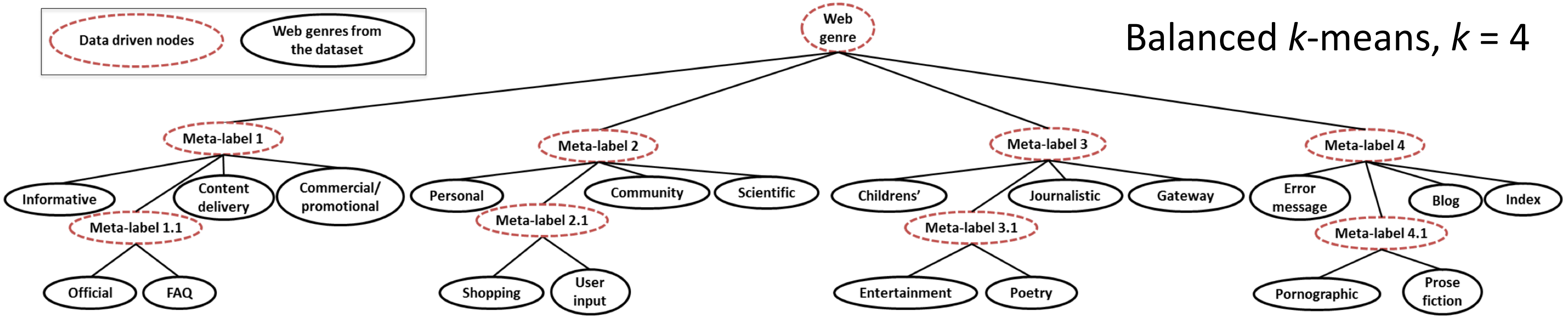
- Experimental setup:
  - CLUS system for predictive clustering was used to construct multi-label PCTs (MLC) and hierarchical multi-label PCTs (HMC).

A hierarchy of genre labels improves the performance over the flat genre labels: the improvement in performance is across all of the evaluation measures

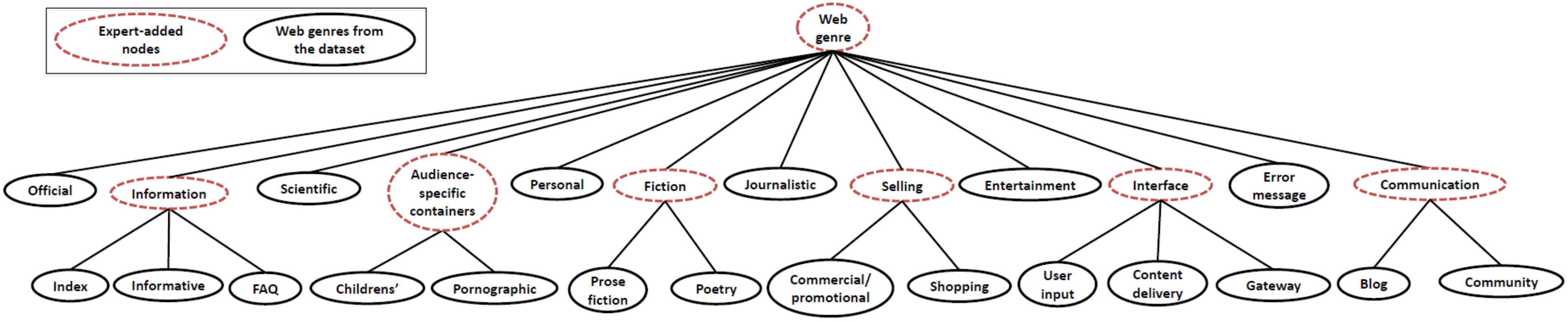
	HammingLoss	*Accuracy*	*Precision*	*Recall*	*Fmeasure*	*SubsetAccuracy*	*MicroPrecision*	*MicroRecall*	*MicroF1*	*MacroPrecision*	*MacroRecall*	*MacroF1*	*OneError*	*Coverage*	*RankingLoss*	*AvgPrecision*
HMC - manual hierar.	0.094	**0.276**	**0.327**	**0.341**	**0.334**	0.172	0.31	**0.33**	0.32	**0.424**	**0.296**	**0.297**	0.643	5.561	0.238	0.47
HMC - BkM (k=4)	**0.081**	0.261	0.31	0.3	0.305	**0.177**	**0.368**	0.291	**0.325**	0.368	0.262	0.284	**0.635**	**5.435**	**0.232**	**0.475**
HMC - BkM (k=3)	0.09	0.223	0.273	0.272	0.273	0.131	0.301	0.263	0.281	0.328	0.212	0.211	0.677	5.878	0.254	0.44
HMC - BkM (k=2)	0.084	0.206	0.247	0.247	0.247	0.127	0.328	0.24	0.277	0.361	0.205	0.227	0.682	5.956	0.259	0.433
MLC	0.111	0.136	0.172	0.165	0.168	0.073	0.165	0.163	0.164	0.063	0.1	0.065	0.83	7.955	0.36	0.317

Data-driven vs. Expert-driven Hierarchy

No grouping of genres in the expert hierarchy that can be noted in the data-driven hierarchy: there is a semantic gap between the meaning of the genres and how these meaning is well represented in the data.



Balanced *k*-means, *k* = 4



Data-driven vs. Expert-driven Hierarchy

Models constructed using data-driven and expert-driven hierarchies have relatively similar predictive performances: each of the models is better than the other according to 8 evaluation measures.

	HammingLoss	*Accuracy*	*Precision*	*Recall*	*Fmeasure*	*SubsetAccuracy*	*MicroPrecision*	*MicroRecall*	*MicroF1*	*MacroPrecision*	*MacroRecall*	*MacroF1*	*OneError*	*Coverage*	*RankingLoss*	*AvgPrecision*
HMC - manual hiear.	0.094	**0.276**	**0.327**	**0.341**	**0.334**	0.172	0.31	**0.33**	0.32	**0.424**	**0.296**	**0.297**	0.643	5.561	0.238	0.47
HMC - BkM (k=4)	**0.081**	0.261	0.31	0.3	0.305	**0.177**	**0.368**	0.291	**0.325**	0.368	0.262	0.284	**0.635**	**5.435**	**0.232**	**0.475**
HMC - BkM (k=3)	0.09	0.223	0.273	0.272	0.273	0.131	0.301	0.263	0.281	0.328	0.212	0.211	0.677	5.878	0.254	0.44
HMC - BkM (k=2)	0.084	0.206	0.247	0.247	0.247	0.127	0.328	0.24	0.277	0.361	0.205	0.227	0.682	5.956	0.259	0.433
MLC	0.111	0.136	0.172	0.165	0.168	0.073	0.165	0.163	0.164	0.063	0.1	0.065	0.83	7.955	0.36	0.317

Features Related to Data-driven and Expert-driven Hierarchy

- Different scenarios exploit different attributes from the dataset

- Data-driven:
  - appearance of the word FAQ in the URL of the web page
    - content related attributes
      - part-of-speech trigrams
- Expert-driven:
  - content related features on the top levels
    - HTML tags information on the lower levels

Conclusions

- The results reveal that using a hierarchy of web genres considerably improves the predictive performance of the classifiers.
- The data-driven hierarchy yields similar performance as the expert-driven with the difference that it was obtained automatically and fast.
- This means for even larger domains (both in terms of number of examples and number of web genre labels) it would be much simpler and cheaper to use data-driven hierarchies.

Further work

- We plan to develop hierarchies of web genres structured as directed acyclic graphs, which seems more natural in modelling relations between genres.
- It could also be useful to adapt the hierarchy construction algorithm to break down existing genres into sub-genres.
- We experimented with single PCTs and plan to test ensembles of PCTs.