CIS Centre for Information and Language Processing
ALBERTA INGENUITY CENTRE FOR MACHINE LEARNING
Andrea Stubbe, Christoph Ringlstetter,Randy Goebel

Elements of a Learning Interface for Genre Driven Search

Genre and Search

What is Genre?

Genre refers to...

... similarity of texts in terms of function (purpose) and form.

Classifiers can recognize genres automatically

- *Using Machine Learning techniques*
- *Exploiting expert knowledge*

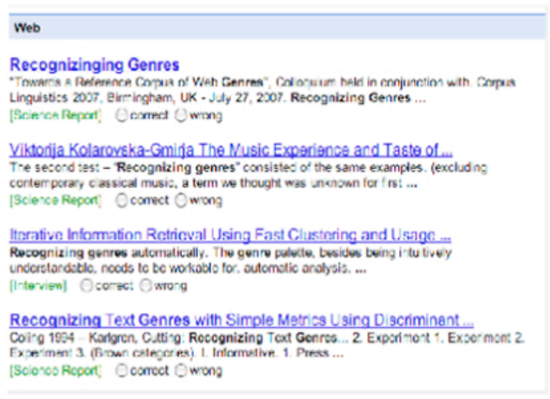
Genre can be used to specify search queries

- *Poems on roses vs. "How to grow roses"*
- *Considered helpful in user studies (Rosso, M. zu Eissen/Stein)*
- *Enrichment of search results improves estimation of relevance (Jose & Joho)*

Genre Hierarchy with 32 Classes

A. Journalism	**C. Information**	D.3 Protocol
A.1 Commentary	C.1 Science Report	**E. Directory**
A.2 Review	C.2 Explanation	E.1 Person
A.3 Portrait	C.3 Recipe	E.2 Catalog
A.4 Marginal Note	C.4 FAQ	E.3 Ressources
A.5 Interview	C.5 Lexicon, Word List	E.4 Timeline
A.6 News	C.6 Bilingual Dictionary	**F. Communcation**
A.7 Feature Story	C.7 Presentation	F.1 Mail, Talk
A.8 Reportage	C.8 Statistics	F.2 Forum, Guestbook
B. Literature	C.9 Code	F.3 Blog
B.1 Poem	**D.Documentation**	F.4 Form
B.2 Prose	D.1 Law	**G. Nothing**
B.3 Drama	D.2 Official Report	G.1 Nothing

Enhanced Search Interface



Use this to incrementally improve classifiers

- *Especially useful for small training corpora*
- *Problem: Introduction of noise*

User Behaviour & Noise

Taxonomie of user behaviour

- *fully cooperative: retrieves and rates everything*
- *cooperative: retrieves only some promising pages and rates all of them*
- *semicooperative: retreives promising pages, but only rates some of them*
- *uncooperative: no explicit information, but observable behaviour (lingering time)*

Sources of noise and information loss

Cooperativeness	Noise	Information Loss
	-	-
	-	only feedback for docs labeled as genre
	snippet recognition errors	only feedback for docs recognized as genre
	relevant topic/wrong genre or vv., exogenous events, snippet rec. errors	only implicit feedback for some of these docs

Page retrieval

- *User typically retrieves 2 pages per result set*
- *Rational user retrieves only relevant pages*
- *Feedback only for positively labeled data, only improves precision*

Not enough labeled data

- *User guesses genre based on snippet*
- *Feedback for unlabeled data improves recall*

Snippet genre recognition factor

- *measure for how well the genre is recognized from the snippet*
- *needed to estimate introduction of noise*

Experiments

- *5 snippets for the 8 journalistic genres recall 45.5%, precision 54.2%*
- *20 snippets for Blog, FAQ and Catalog, 10 for News recall 84.32%, precision 83.02%*

Automatic Classification & Adaption

Classifiers

- *one hand-crafted classifier per genre*
- *similar to decision trees*
- *identify genre specific features*
- *avoid statistically derived features*

All classifiers and features are made public:
http://www.cis.uni-muenchen.de/~andrea/genre/

Features

- *form/appearance: line length, document structure*
- *vocabulary/word lists: genre-specific, pos./neg. adjectives, names, emoticons, informal language*
- *part of speech*
- *patterns: ordered dates*
- *combinations: "..." + pronouns + names = agents*

Preparations

- *convert classifiers into DNF*
- *give explicit upper and lower bounds*

$((A > 3) \& (B < 8)) \,||\, ((B > 3) \& (B < 6) \& (A > 5))$

$((A > 3) \& (A < INF) \& (B > -INF) \& (B < 8)) \,||$
$((A > 5) \& (A < INF) \& (B > 3) \& (B < 6))$

Adaption (false negative)

```
for each (disjunction d of classifier) {
    c_d = sum of required changes to recognize doc
}
if ( min(c_d) < max_allowed_change) {
    C_temp = classifier including changes for d
    pos = new correctly recognized docs
    neg = new falsely recognized docs
}
if (pos > neg) {
    C = C_temp
}
```

Experiments with Corpus Data

Experimental Setup

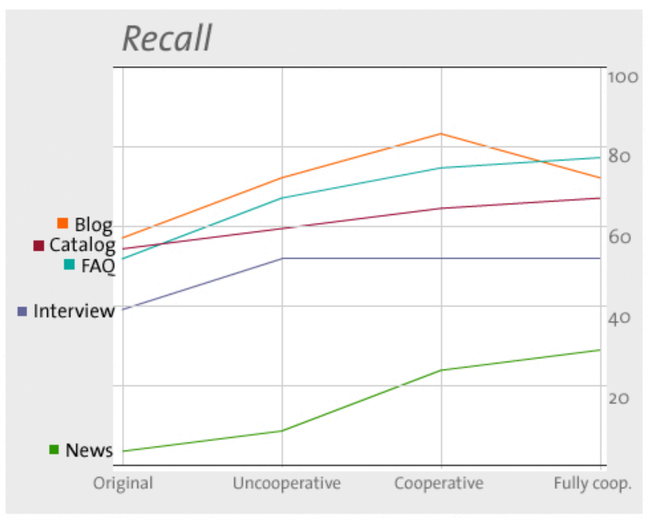
- *randomly generate result sets of 20 pages, with binary labeled (is/is not genre G)*
- *fully cooperative user gives feedback for all pages*
- *other users retrieve two pages per result set*
- *prefer pages of desired genre G (labeled or guessed from snippet)*
- *use initial classifiers trained on 20 pages per genre*
- *recall 60.5%, precision 65.4%*

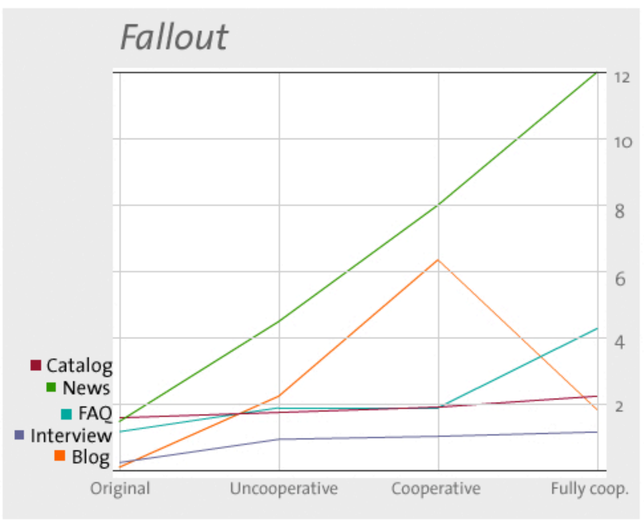
Setting 1

- *FAQ, Blog and Catalog (160 docs each)*
- *980 random*

Setting 2

- *Interview, News (400 docs each)*
- *1000 random pages*





LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
Google
iCORE
UNIVERSITY OF ALBERTA