Classifying XML Documents by using Genre Features

4th International Workshop on Text-based Information Retrieval
in conjunction with DEXA 2007
Regensburg, Germany
3-7 September 2007

Malcolm Clark & Stuart Watt
School of Computing
The Robert Gordon University

Outline

1. The Problems
2. Introduction to genre
3. Traditional genre
4. Common uses
5. XML & genre
6. Background: text classification
7. The structure of genre
8. Concepts and features
9. Which features?
10. Experiment
11. Future work
12. Conclusions

The problems

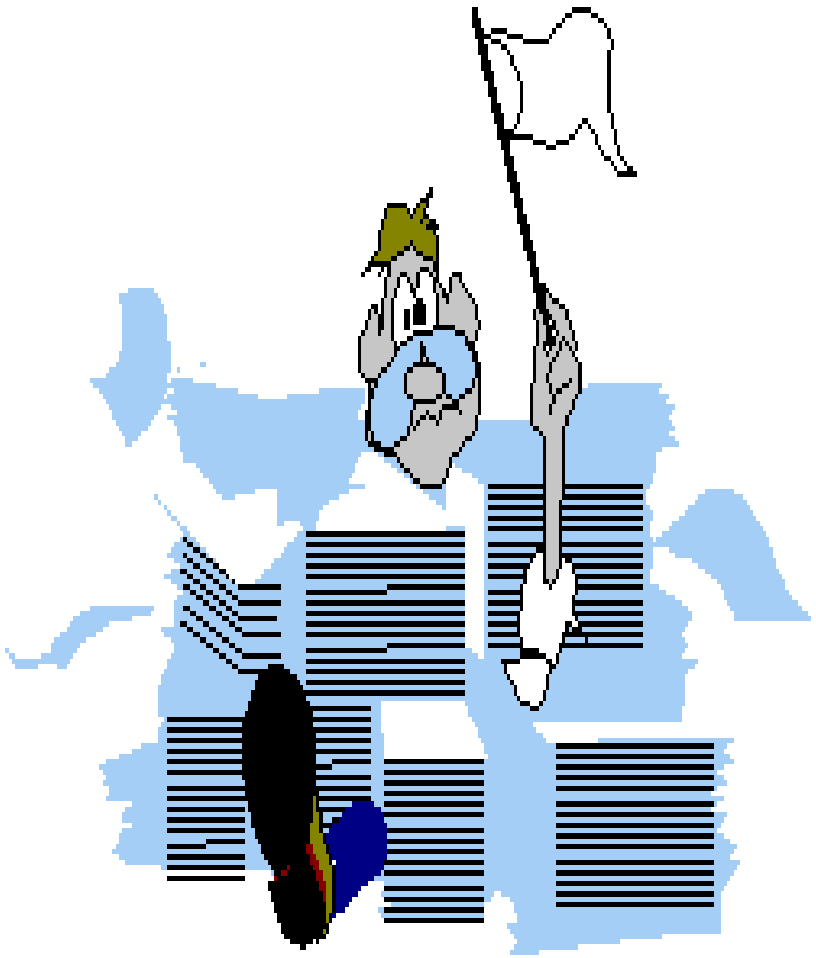
•Retrieve relevant information to users' needs. Ideally the relevant topic and genre

•The text types could be: academic and scientific articles, biographies, news articles, F-A-Q and so on.

•Text also structured by social consensus and in different contexts such as email (body) and discussion room conversations.

•Research into structured text retrieval (i.e. XML) is fast paced Lalmas et al. (2004), Lalmas and Ruthven (1997) and Wilkinson (1994).

•This paper and our research regards structured-text retrieval…



Introduction to genre

1. Traditionally a **genre**, (French: "kind" or "sort") is a loose set of criteria.
2. Genre normally used to identify types of literature or music.
3. Genre in the context of this research is the ‘purpose’ and ‘form’ of the document.

Traditional genre

Generally genre divided into two main schools of thought:

•The North American School: socio-historical, rhetorically-oriented concept, with the emphasis placed on how texts function in social and interactional contexts.

•Sydney School: based on an applied linguistic approach, with the focus on formal textual features.

Common uses for genre in a research context

Commonly used in three areas or research (Meyer zu Eissen and Stein 2004):

- Literary theories of genre (including kinds of literature)
- Genre and the WWW
- **Automatic genre identification**

XML and Genre

XML (eXtensible Mark-Up language) retains the structure:

The genre indicates

The document's

form and substance

(Dewdney et al.2001)

<abstract><abstract/>
<introduction><introduction/>
<background><background/>
<experiment>
<method><method/>
<classifiers_selected></classifiers_selected>
<classification_setup></classification_setup>
<results></results>
<experiment/>
<related_work></related_work>
<discussion></discussion>
<future_work></future_work>
<conclusions></conclusions>
<acknowledgements></acknowledgements>
<references>
<reference></reference>
<reference></reference>
<reference></reference>
<reference></reference>
</references>

Academic Article

Background text categorization

- Aka text classification (TC) or topic spotting
- The categorization of documents is normally implemented by labeling and classification.
- Task is to automatically sort a set of documents into categories, classes or topics from predefined categories (Sebastiani 2005).
- Purpose of TC: make a set of documents easier to manage without resorting to manual sorting.

The structure of genre

Genre classification: genre classification is defined as discrimination of documents by their form, style, functionality etc.

Meyer zu Eissen and Stein (2004)

Other concepts persistent in genre: content, function or interface.

Genres incorporate these concepts.

These concepts ideal for WWW IR support.

Concepts ideal for structured documents especially XML.

Concepts and features

Concept	Small Selection of Feature Examples
Style	Readability and part-of-speech (POS) aka grammar statistics.
Form	Text statistics, whitespace, and formatting tag analysis.
Content	Terms, words in HTML title tag and URL, numeric types, closed-world sets, punctuation.
Functionality useful for web content)	Number of links in a web page; number of e-mail links.

Which features belong to which set?

An iambic pentameter may be analyzed as a style of writing or, indeed, content or form because it consistently contains an unstressed syllable followed by a stressed syllable

For example:

Shall I / com PARE/ thee TO / a SUM / mer's DAY?
Thou ART / more LOVE / ly AND / more TEM / per ATE
(Shakespeare Sonnet 18)

A document may consistently contain a high frequency of punctuation, possibly indicating that a poem can be judged as form or content.

Our approach used the concept of form only…

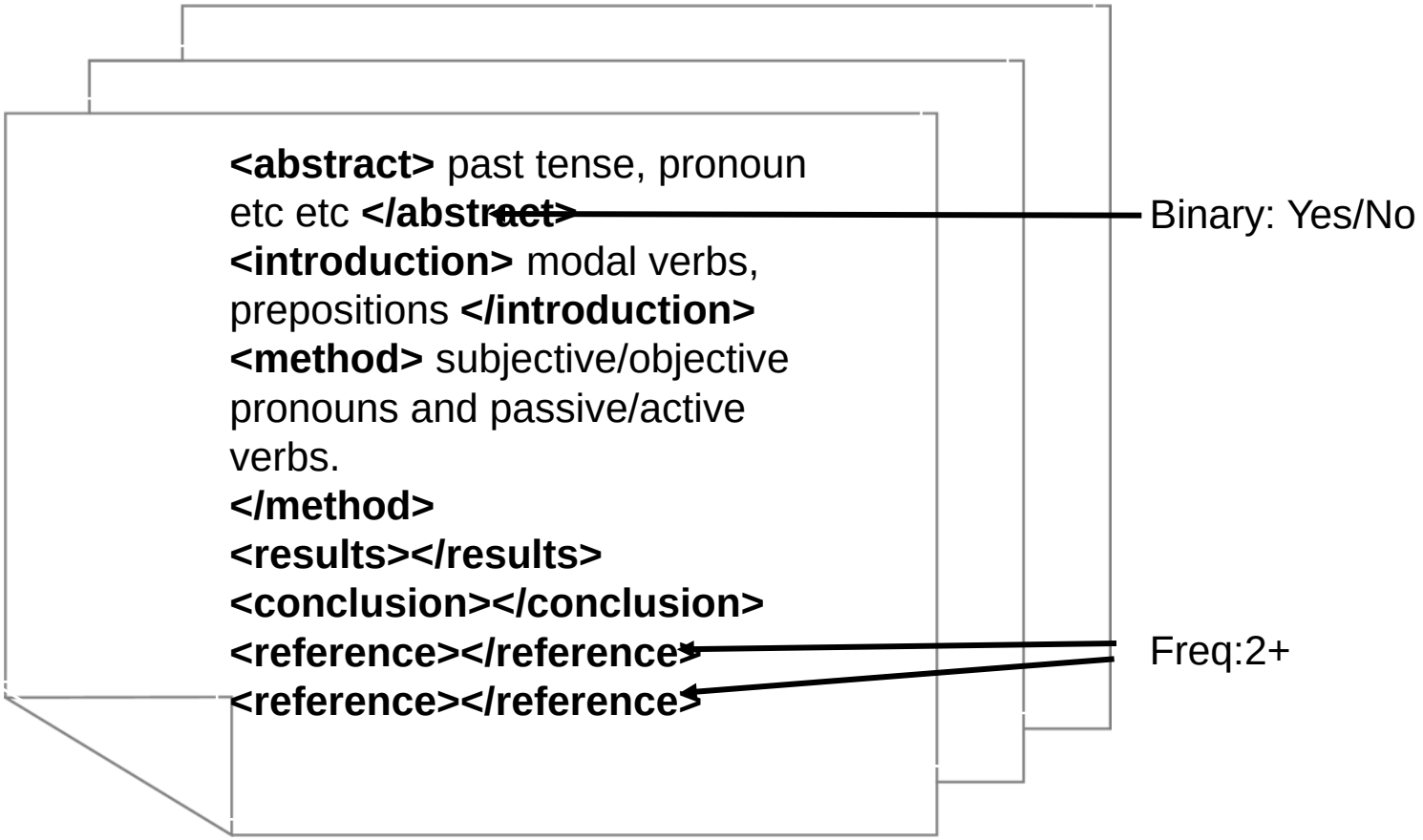
Experiment

Experiment

Hypothesis: if a focus on genre can lead to high precision on normal textual documents, then good results can be achieved using XML tag information in addition to P-O-S (grammatical structure) information

Hypothesis and task

Task: to exploit the document features, train a classifier using nominal and numeric attributes. For example:



Corpus

- INEX 1.4 collection
- 1093 documents (10% of total)
- IEEE Computer Society's publications from 1995-2002
- Pre-labeled by the INEX 1.4 corpus administrator
- INEX labeled each document by title, i.e. theme article, biography, cumulative index, etc.
- There is NO genre label

Features

This study focused on 28 features (11 grammar,17 XML).

Two sets:

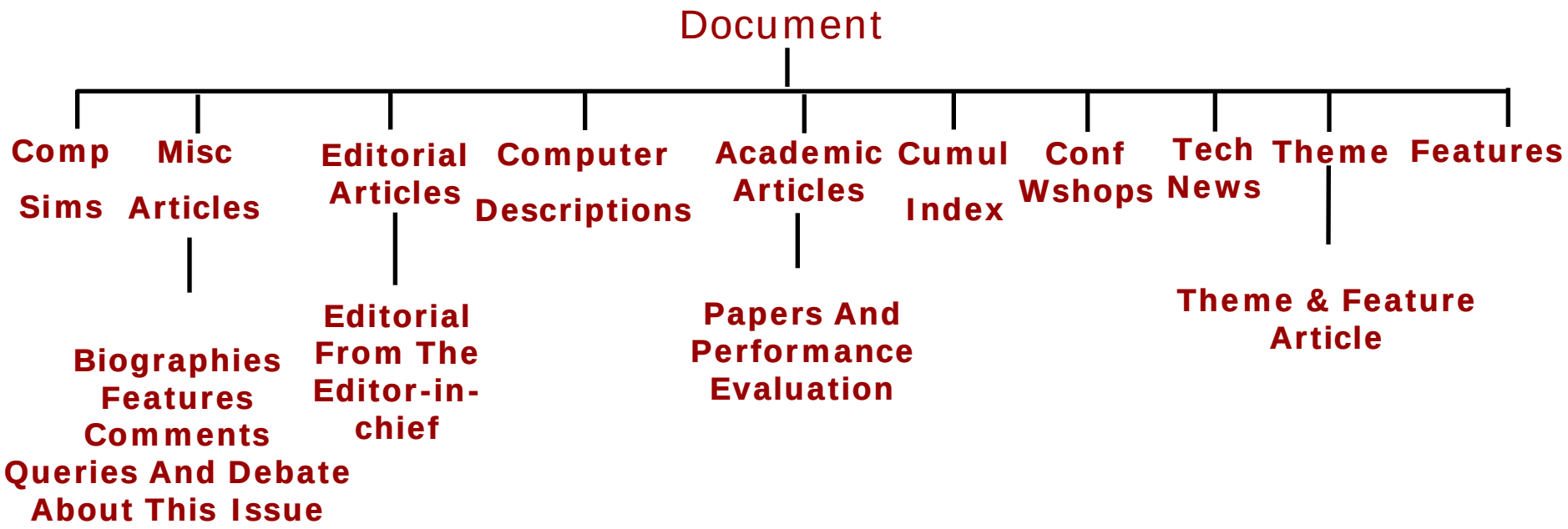
P-O-S (grammar): modal verbs, prepositions, tense, subjective/objective pronouns and passive/active verbs.

XML tags…

XML Features

Form Feature	Type	Examples
Frequency of tags: Abstract, Table, List, Equation, Figure, Paragraph, References Average tags per document	Number of XML Tags in each document	<abs>,<tbl>, <en>,<fig>, <p>,<ref>, , <url>
Paragraph	Average paragraph length	1024
Number of URL links in article	URL	Too large to show example, however, articles with many links may indicate regular feature.

Collection of genres



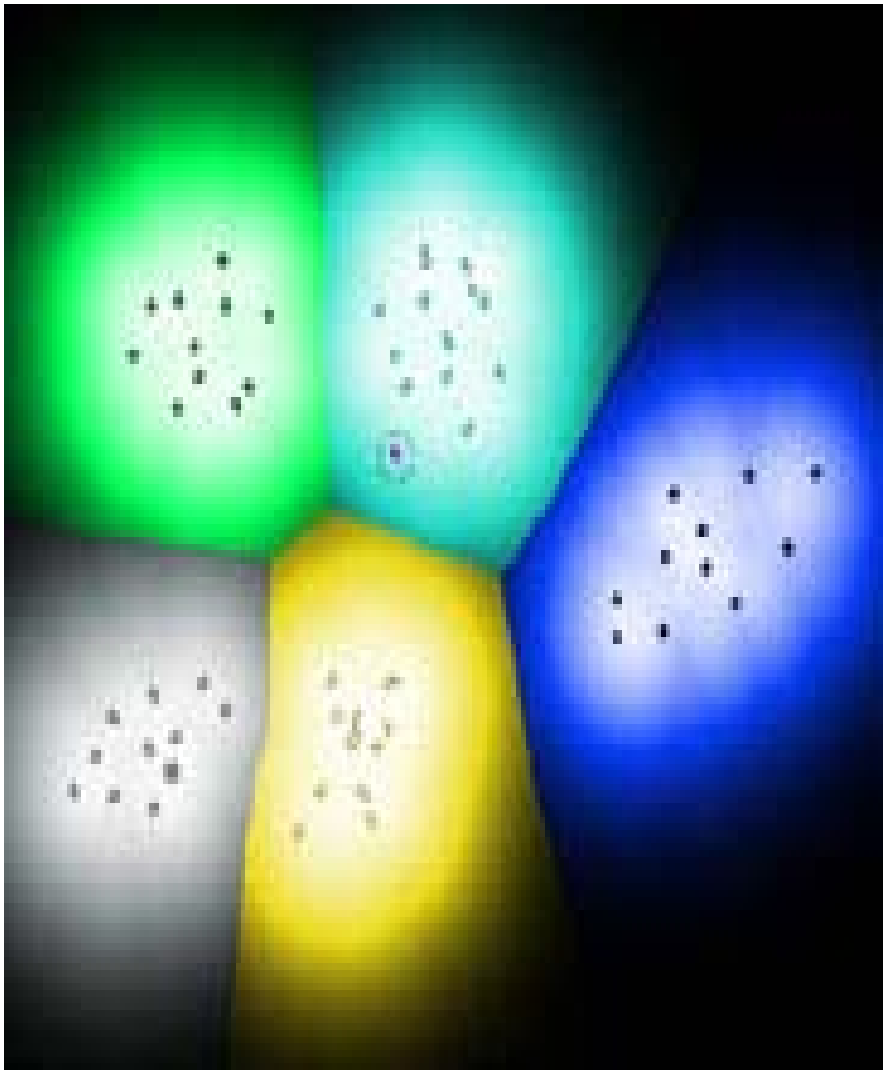
- Total of 10 ‘final’ genres.
- Many genres merged as similarities exist in structure.

Classifiers selected

We experimented with many of the classifiers on WEKA platform.

Most effective (ranked order):

- Additive logistic regression (LogitBoost) with a decision stump learner.
- Neural network (MultiLayerPerceptron).
- SMO (support vector machine)
- Naïve Bayes
- Decision Tree (J48)



Logitboost results

•Averaged results for LogitBoost using only P-O-S are encouraging, but still quite poor.

•Using only XML tags (form) LogitBoost, 973 (89.021%) out of 1093 documents were classified correctly with 120 misclassifications.

•Averaged results for LogitBoost using P-O-S and form together. 1066 (97.5%) of 1093 documents were classified correctly with only 27 misclassifications.

Results

False Positive	Precision	Recall	F1	Class/Genre
0.001	0.993	0.993	0.993	Misc.Articles n=150
0.003	0.857	0.818	0.837	Computer Prescriptions n=22
0.002	0.867	0.929	0.897	Computer Simulations n=14
0.004	0.789	0.833	0.811	Conferences/Workshops n=18
0	1	1	1	Cumulative Index n=5
0	1	1	1	Academic Article n=700
0.008	0.916	0.926	0.921	Theme n=94
0.001	0.969	0.969	0.969	Features n=32
0.003	0.897	0.813	0.852	Editorial Articles n=32
0.005	0.815	0.846	0.830	Technology News n=24

Other Results

- Results for the Neural Network classifier (MLP) are very similar to the LogitBoost statistics: 1062 correct, 31 incorrect with average 97.1% accuracy.

- Poorest results were obtained by using Decision Tree (J48), with 892 correct, 201 incorrect and 81.6% accuracy.

- Other full results and comparisons available on request.

Issues

•Not sure if MLP or LB is best due to 1.93% standard deviation.

•One possible reason that LogitBoost is best is that boosting weights features according to how well they discriminate between genres.

•**Possible bias:**

n sizes of some of the genre types: the ‘Theme’ genre has an n=700 which could skew certain algorithms.

•Work on classifying the whole 12,108 corpus has continued leading to similar results.

Future work

1. Creating or extending skimming/information extraction models, such as those found in DeJong's FRUMP Predictor/Substantiator using e-mails, Wikipedia and self crawled Yahoo collection.

2. Test eight genres of email's using the J.J. Gibson 'invariant' features which are used to make decisions during the categorisation/relevance process.

    a. Utilize eye-tracking experiments to obtain an accurate understanding of how humans view the invariant layout cues for example, white-space patterns or other formatting features which constitute genres and test how genres 'afford' their purpose.

3. Explore the genres which evolve in a collection which is created through human social concensus ie Yahoo Urticaria forum which consists of 75000 documents.

4. Yahoo discussion rooms are rich in natural language but the question is whether retrieval is improved by exploiting genre patterns and rules which normally emerge through social consensus in a community of practice and have a distinctive purpose and form.

Conclusions

- XML/XHTML is so widely used research needs to be increased especially into genre.

- We have shown albeit on a small scale that by using XML and P-O-S information it is possible to automatically differentiate between genres in structured text.
- **Caution:**
  - The INEX 1.4 corpus represents mostly scientific, technical and academic documents.
  - Question the representativeness of the corpus.
  - Try and answer this query by conducting further research with other XML/XHTML collections.

- Nevertheless we still obtained promising genre classification results which need to be built on in the future.

Thanks for listening!