Classifying XML Documents by Using Genre Features

Malcolm Clark, Stuart Watt
The Robert Gordon University
{mc,sw}@comp.rgu.ac.uk

Abstract

The categorization of documents is traditionally topic-based. This paper presents a complementary analysis of research and experiments on genre to show that encouraging results can be obtained by using genre structure (form) features. We conducted an experiment to assess the effectiveness of using extensible mark-up language (XML) tag information, and part-of-speech (P-O-S) features, for the classification of genres, testing the hypothesis that 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 information. An experiment was carried out on a subsection of the initiative for the evaluation of XML (INEX) 1.4 collection. The features were extracted and documents were classified using machine learning algorithms, which yielded encouraging results for logistic regression and neural networks. We propose that utilizing these features and training a classifier may benefit retrieval for most world wide web (WWW) technologies such as XML and extensible hypertext markup language) XHTML.

1. Introduction

XML is becoming a dominant format for managing structured text of all types, and is in widespread use in many fields, varying from digital libraries and electronic publishing to the so called ‘semantic web’. As the quantity and size of XML document collections continue to expand, for example, as web and digital libraries, the need for information retrieval (IR) systems which exploit structured text, as opposed to traditional bag-of-words (B-O-W) based IR systems, is also increasing.

A genre in this context is the set of structures, layout and style of writing (or as Dewdney et al. [1] states “conventions”) which show the user the documents’ purpose and form through its structure regardless of the topical nature of the writing. Since XML retains genre information, such documents should be explored for genre categorization.

The background to the current situation is given in Section 2 of this paper, explaining how the need for research in this field has arisen. Section 3 presents the data used in the experiment, the features being analyzed, the classes of classification, the genres and the results. This leads on to Section 4, which is dedicated to a discussion of related works. The penultimate section of this paper discusses the potential areas of interest for future research and the final part, Section 6, summarizes our conclusions.

2. Background: text classification and genre

Text classification, in the form of information filtering, keyword assignment, or labeling, is often used to reduce the information overload associated with using a large collection; this helps the user to focus on the actual documents required. Genre classification is defined by Meyer zu Eissen et al. [2] as discrimination of documents by their form, style, or targeted audience. A user may wish to find tutorials, reviews or academic papers, as these are increasingly being made available in XML form. The form or structural information contained within the interface of specific XML or hidden in the tag information needs to be exploited. Genre and text classification work by grouping collections of documents into smaller groups which are usually pre-labeled, but genre utilizes the form (and sometimes style, function etc) of the documents whilst text classification traditionally discriminates by topic using such things as keywords. In this context, we and other authors view genre as orthogonal to topic [3].

Our contention is that a search engine using the structural information implicit in most documents, and especially XML documents, could assist, we believe, the user to retrieve the correct type of information, with high performance and effective filtering of the results. Genre is not used nearly enough, for example, if the search query “Tim Berners-Lee” is entered into a search engine, the results returned contain articles, reviews, etc. The retrieval of such a wide range of information may lead to ‘information overload’.

In these cases, using standard approaches would fail because the standard interpretation of user needs does not take into account the need to discriminate by genre, but only by topic. We argue that search results could be improved by using structural information to filter and reduce the cognitive load of the user.

XML and XHTML provide new opportunities for capturing and recording genre by enabling explicit representation of the structure as well as the P-O-S information within documents. Since the structure is maintained, it allows the genre to show its purpose, for example, an ‘article’ genre in the INEX collection (Figure 1):

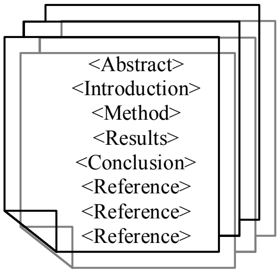

Figure 1. Academic article structure

Other examples are scientific papers which may contain <en> equation tags, and cumulative index pages which maintain a specific structure throughout. There is a need for other features which can aid the extraction and classification of information from XML sources; we thus propose a suitable set.

2.1. Genre text classification – other work

Most research into the classification of documents is based around the topical aspect of a document; research into genre therefore needs to diversify. Many authors, such as Rauber et al. [4], Boese et al. [5, 6], Finn et al. [3, 7], Santini [8-10] have recognized the value of genre in conventional libraries, web and information searches.

The authors mentioned above all see genre as an excellent distinguishing feature to be exploited in digital collections. Campbell & Toms [11] contend that the “attributes” of a document’s genre enable it to be specifically identified and show that genre attributes play a significant role in identifying documents. They conducted experiments using both form and function, exposing users with backgrounds in information technology and the academic world to digital and hard copies of documents.

Although the experiments by Campbell & Toms [11] are excellent examples of genre work, researchers have recognized the difficulties inherent in genre classification tasks; Rauber et al. [4] believe that there are many document genre similarities with overlaps occurring throughout. We identify completely with this problem of similarity as some genres in this test collection and the new INEX Wikipedia collection do overlap and look similar.

Collins et al. [12] criticize organizations for not exploiting their collections to their full potential, but also point out that, until as recently as 1995, the only techniques in existence involved text analysis and statistics which were employed to detect semantic information automatically. Collins et al. do perhaps have a point, not with regard to traditional text archives, but in the context of XML archives. The XML archives are growing and are simply not used to their maximum potential as regards genre classification. To illustrate this point, all the papers mentioned were written after experiments had been performed on traditional text collections.

Numerous shallow parsing experiments exist for XML retrieval [5, 6, 13, 14], but not for XML genre research. Many papers have, however, been written on web genre research techniques, for example [5, 13] which are applicable for most digital collections.

2.2. The structure of genre

In current research, a number of approaches are being employed to judge the genre of documents. There are several underlying concepts that are persistent in genre definitions: the style, form, and content of the document [6]. Web genres incorporate the style, form, and content of the document which are orthogonal and not related to the topic or classification of many genres. These three concepts are more relevant than any existing definitions found within WWW IR circles. The conceptual features of style, format and content can be used with other digital document formats, especially XML.

Campbell & Toms [11] suggest that the conceptual features consist of a grouping of unique facets or levels, i.e. function, form and interface. Function is representative of the meanings of the words contained within the documents, form is the layout or appearance, and, finally, interface is the way in which the document is read or used.

By looking at these conceptual features, many pieces of genre classification works can be seen to fit with the concepts they describe (Table. 1).

Table 1. Concept examples

Concept	Small Selection of Feature Examples
Style	Readability and part-of-speech (POS) statistics [6, 3, 15]
Form	Text statistics, whitespace, and formatting tag analysis. [6, 16-18]
Content	Terms, words in HTML title tag and URL, number types, closed-world sets, punctuation. [6, 17]
Functionality	Number of links in a web page; number of e-mail links [17, 18]

Taking style, form, content and functionality into consideration, there are hundreds of features that can be measured and the normal practice is to group them as feature sets.

Content can be analyzed by looking at punctuation, such as ';' or '?', closed world sets, such as 'Dr', 'Mr.', 'Mrs.' or emoticons '☹'; or word frequencies,

Style context systems include categorizing document genre by punctuation frequencies or readability. For example, the document might use colons and semi-colons for elongated sentences, use long and/or conjunctive adverbs such as 'nevertheless' and 'otherwise' or 'indeed', and be written in complete sentences.

Form looks at features that could include text statistics and analysis of the tag structure which is contained directly in the document, unlike the implicit content mechanisms of document type definition's (DTD), schemas and namespaces. Using tag frequencies for features is, of course, not a totally new concept and has been discussed and utilized before [16, 19].

There is some debate regarding the sets to which some features belong, for example:

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

This overlap is useful because it enables the classification to be tested against each feature set, for example, style versus form. Some documents do provide visual markers that allow the reader to conceptualize the format. Documents contain distinguishable features which allow the reader to identify the purpose and genre of a document.

3. Experiment–genre classification

Our hypothesis was as follows: that 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 (or grammatical) information. By using the structures shown in the corpus genres, exploiting the features, and training a classifier on binary (nominal) and numeric attributes we should be able to differentiate between genres. For example, a binary result of 'yes' for an abstract tag and with a numeric greater than 0 for reference tags will indicate an academic article.

3.1. Method

We chose the INEX 1.4 collection because the genres represented in this collection are good examples of typical genres. We used a total of 1093 documents (randomly chosen) which comprise a small subsection (just over 10%) of the whole INEX corpus (the IEEE Computer Society's publications from 1995-2002). The documents had already been pre-labeled by the INEX 1.4 corpus administrator and are true to the original formats used in the magazines. INEX labeled each document by title, i.e. theme article, biography, cumulative index, etc. There is no genre label or tag to specify the particular genre type, only the knowledge of how the IEEE magazines are structured. Since the documents consist of old IEEE journals and magazines, the types of genres were already established and understood (collection and/or DTD available from author on request).

3.1.1. Indexing. After the initial indexing to study the content, we indexed the collection by form and P-O-S features (3.1.2). In order to be as thorough as possible, we chose to record not only frequency counts, but simple binary values to see if we could find any simple heuristics to define each genre, for example, an <abs> tag would indicate an abstract, a <ref> tag a reference so both of these could be considered as indicative of an academic article. The full lists of features are available on request.

3.1.2. Genre features analyzed. This study focused on 28 specially selected features which were all classified together, but could be categorized into two sets:

- **P-O-S**, for example, modal verbs, prepositions, tense, subjective/objective pronouns and passive/active verbs.

- **XML tag** information for example, frequency of URL tags, title, image, abstract and so on.

We found that many of the features selected by the authors listed in 2.2 were contained within the old INEX documents collection and aided the identification of documents by genre. It is important to note that some authors do not offer a complete list of form and P-O-S features, so that only a limited comparison of the features is possible. Some small clues are offered, however, such as tenses in the text or average image tags [6].

3.1.3. Genre types and corpus structure. Our previous experiments contained only a small list of exemplar genres [16]. This paper contains ten top-level genre types (see class/genre column in Table. 2), but some are identical structures and as such are merged (Figure. 2).

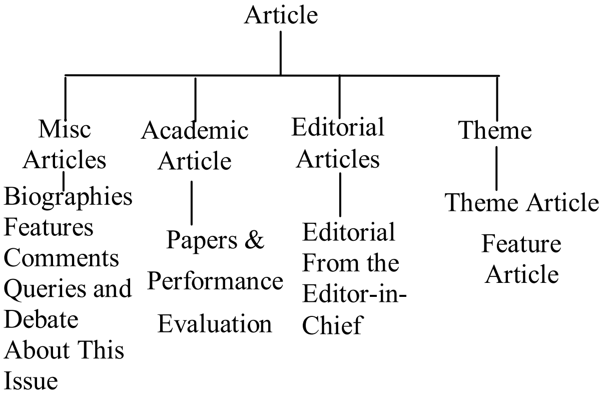

Figure 2. Similar genres merged

3.2. Classifiers selected

We experimented with all the classifiers available to us on the WEKA platform, including SMO which is an implementation of a support vector machine (SVM). The most effective were additive logistic regression (LogitBoost) with a decision stump learner and neural network (MultiLayerPerceptron) algorithms on the data set.

Feature selection was not used extensively in this experiment because of the quantity of features extracted. Feature weighting assists the user to achieve higher classification accuracy rates and to filter out ineffective features. For example, Bayes Net weights all the features equally whilst LogitBoost applies feature weighting. LogitBoost was setup with the Decision Stump regression learner, which carries out classifications based on entropy. LogitBoost implements boosting, i.e. additive logistic regression through a step-wise optimization.

3.3. Classification – setup

We used ten-fold stratified cross-validation which divides the training and test data into a 90/10 proportional split. Basically, each of the ten groups contained the same number of instances for each class. Each classifier was tested ten times in a round robin fashion. During each test, iteration 10% (one group) was utilized for testing while the remainder was used for training. This process is advantageous since it minimizes the risk of each classification algorithm being misrepresented by good or bad results.

It became apparent during the initial classification experiments that many articles were structurally identical. For example, theme and features had to be merged together as this factor had a bearing on the results.

3.4. Results

The measures are calculated as follows[1]:

- Precision (P) = total relevant documents retrieved / total documents retrieved.
- Recall (R) =total relevant documents retrieved / total relevant documents.
- F-measure (F) = 2 precision x / (precision + recall) which is the harmonic mean of precision and recall.
- True Positive (TP) = number of documents of the test set that have been incorrectly classified.

3.4.1. Features. Interestingly, the most effective features ranked for the dataset were tag counts per document, average word length, frequency of reference tags, average paragraph size, images, size of document in kb, and table tag counts.

3.4.2. LogitBoost result. The averaged results for LogitBoost using only P-O-S are encouraging, but still quite poor. The full results for P-O-S are not shown due to space issues, but by using LogitBoost, 973 (89.021%) out of 1093 documents were classified correctly with 120 misclassifications. The averaged results for LogitBoost using P-O-S and form together are also encouraging (Table. 2 below). By using LogitBoost, 1066 (97.5%) of 1093 documents were classified correctly with only 27 misclassifications.

[1] A thorough overview of machine learning evaluation measurements is detailed in Sebastiani's paper [20]

One possible reason that LogitBoost performed so well could be that boosting weights features according to how well they discriminate between genres.

Once more, due to space issues, the MLP results are not shown here in full, but are summarized. The results for the Neural Network classifier (MLP) are very similar to the LogitBoost statistics: 1062 correct, 31 incorrect with average 97.1% accuracy. The worst results were obtained by using Decision Tree (J48), with 892 correct, 201 incorrect and 81.6% accuracy.

Table 2. Detailed accuracy by genre using all features (false positive (FP), precision (P), recall (R), Weighted harmonic mean of precision and recall (F1)

FP	P	R	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=26

Further deeper analysis revealed that, due to the standard deviation of 1.93, we could not be certain which classifier, LogitBoost or MLP, actually performed best.

Although the results are encouraging, it is important to identify a possible bias, such as the n sizes of some of the genre types: the 'theme' genre, for example, has an n=700 (this genre type has theme and feature articles) which could skew certain algorithms.

4. Related work and discussion

4.1. Related work

Kennedy et al. [17] used a neural net classifier in which the input units represent terms, the output units represent the chosen category(s) of interest, and the weights on the edges connecting units represent dependence relations. Finn et al. [3] used decision tree induction to break down the results into if/then categorization rules.

They found that P-O-S techniques are better than bag-of-words techniques in the context of domain transfer (the ability of a classifier to classify unseen documents correctly in a different domain). Finn et al. used other classifiers for initial testing such as OneR and k-Nearest Neighbor (k-NN). C4.5 (using a decision tree) performed consistently best for their experiments. The results [3] showed that in the single domain, hand-crafted shallow linguistic features perform best with an average score of 88%; P-O-S perform worst at an average of 85%.

Kessler et al. [21] and Boese [6] used logistic regression for the automatic detection of text genre.

4.2. Discussion

Much current research is being carried out on conventional IR document collections, such as the text retrieval conference (TREC) Enterprise collection (email), Blog and INEX Wikipedia collections which are arguably rich in genres. ACM and IEEE have published numerous papers detailing genre research experiments, and there has been a genre mini-track at HICSS (Hawaii international conference on system sciences) for many years [22].

Further work is essential, however, in the field of genre in IR research, using all collections (i.e. blogs) and not only XML, published using XHTML, etc but any which have useful structural characteristics. All the facets of creating a system, and implementing the various techniques and evaluation processes need to be subjected to long-term, detailed analysis. Many search engines, web or otherwise, will need to take genre clustering [8] into account to further enhance the classification of documents.

Although research [6] has shown that genre can be retrieved using features of form, content and style, sometimes using a mix of the three for the purpose of comparison, there are very few studies which focus solely on the form features of documents, i.e. [16, 1].

4.3. Discussion of genres

Extending the number of genres would be one of the most useful tests because some of the genre types may overlap significantly; this may reduce or enhance precision but could have a positive effect with regard to any results. Any benefits, of course, will only become apparent in the future if the goal is to classify the whole INEX 1.4 corpus: the greater the number of genres identified, the greater the validity of the work.

In future, documents may have to be classified which contain multi-genres: course induction booklets, for example, might contain timetables, course module descriptions and exam timetables, all of which constitute different genres and have to be taken into account.

5. Future work

Now that the new TREC blog and INEX Wikipedia collections are available, we plan to see whether the same genre patterns emerge in these bigger corpuses. The majority of the work will be carried out using XML collections, for example XML vocabulary XHTML, as this type of document is being used more extensively, i.e. the new TREC blog collection. Our method of exploiting the tag structure of the documents will be extended using web genres in the future.

Further research is also planned to find out whether Gibson's theory of 'affordances' [23] could be useful for genre, to discover how human beings perceive document genres, and which layout cues they perceive. Gibson's theory represents an effort to reassess the links between perception and meaning: he argues that instead of perceiving objects (i.e. texts) and then adding meaning later, there are visual combinations of 'invariant' structural properties of objects, such as text, which 'afford' or cue possibilities in relation to these documents [24].

6. Conclusions

The ever-increasing quantities of digitalized XML corpuses available have given rise to an urgent need for more research on the diverse methods of IR Structured text retrieval involving genre is under-utilized; research on genre, such as the classification of documents using form features alone, is still at an embryonic stage. This paper has highlighted some of the difficulties inherent in this field. Research on genre classification is normally carried out by means of parallel analysis with content or other feature sets. As far as this study was concerned, the most important question, which defined the margins of success or failure, was whether structured documents (XML) could be classified by genre by using the XML tags, also together with P-O-S. Our results have shown that this is possible. Our experiment also showed (obtaining encouraging results) that a set of documents can be classified using solely form features (format/layout, etc.), if the user has some knowledge of the attributes of the documents in their collection. Results showed that classification can also be improved by adding grammatical structure information (P-O-S) as a feature set.

A small note of caution: although our work has demonstrated that retrieval of genres maintained in XML documents is possible, the INEX 1.4 corpus represents scientific or technical documents and it is both right and necessary to question the representativeness of the corpus. This query can, of course, be answered by conducting further research on with other XML/XHTML collections.

Nevertheless, useful results were produced for XML genre classification using a relatively small, but still useful, corpus and many genres. We believe that the filtering of genre types when searching is highly beneficial to anyone seeking the ultimate IR goal – the retrieval of relevant results.

7. Acknowledgements

Malcolm sends huge thanks go to his wife Sam, mum Anne, also Ulises for his encouragement, help and discussion.

8. References

[1] Dewdney N, VanEss-Dykema C, MacMillan R. The form is the substance: classification of genres in text. Association for Computational Linguistics Morristown, NJ, USA; 2001. p. 1-8.

[2] Meyer zu Eissen S, Stein B. Genre classification of web pages. Proceedings of the 27th German Conference on Artificial Intelligence (KI-2004), ; 2004; Ulm, Germany: Springer; 2004. p. 256-69.

[3] Finn A, Kushmerick N, Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*. 2003;57(11):1506-18.

[4] Rauber A, Muller-Kogler A. Integrating automatic genre analysis into digital libraries. JCDL '01: Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries; 2001 June 24-28, 2001; Roanoke, Virginia: ACM Press; 2001.

[5] Boese ES, Howe A. Genre Classification of Web Documents. American Association for Artificial

Intelligence: Computer Science Department, Colorado State University; 2005.

[6] Boese ES, Howe AE. Effects of web document evolution on genre classification. CIKM'05. Bremen: ACM Press New York, NY, USA; 2005. p. 632-9.

[7] Finn A, Kushmerick N, Smyth B, Crestani F, Girolami M, van Rijsbergen CJ. Genre Classification and Domain Transfer for Information Filtering. 24th European Colloquium on Information Retrieval Research Proceedings of ECIR-02; 2002; Glasgow, Uk.: Springer Berlin / Heidelberg; 2002. p. 353.

[8] Santini M, Genres In Formation? An Exploratory Study of Web Pages using Cluster Analysis. *Proc CLUK*. 2005;5:8.

[9] Santini M, Some Issues in Automatic Genre Classification of Web Pages. *Proc of the JADT*. 2006.

[10] Santini M, Power R, Evans R. Implementing a Characterization of Genre for Automatic Genre Identification of Web Pages. Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions; 2006; Sydney, Australia: Association for Computational Linguistics; 2006. p. 699-706.

[11] Campbell DG, Toms EG. Genre as interface metaphor: exploiting form and function in digital environments. Proceedings of the 32nd Hawaii International Conference on System Sciences 1999; Hawaii: IEEE; 1999.

[12] Collins TD, Mulholland P, Watt SNK. Using genre to support active participation in learning communities. European Conference on Computer Supported Collaborative Learning (Euro-CSCL'2001); 2001 January 2001; Maastricht; 2001.

[13] Rehm G. Towards Automatic Web Genre Identification. Proceedings of the Hawai'i International Conference on System Sciences; 2002 January 7–10, 2002; Big Island, Hawaii: IEEE; 2002.

[14] Selamat A, Omatu S, Web page feature selection and classification using neural networks. *Information Sciences*. 2004;158:69-88.

[15] Finn A, Kushmerick N, Smyth B. Genre classification and domain transfer for information filtering. Springer; 2002. p. 353–62.

[16] Clark MJ. *Classifying XML Documents by Genre Vol.1.* [MSc]. Aberdeen: The Robert Gordon University; 2005.

[17] Kennedy A, Shepherd M. Automatic Identification of Home Pages on the Web. 2005.

[18] Shepherd M, Watters C. The Evolution of Cybergenres. In: Society IC, editor. HICSS '98: Proceedings of the Thirty-First Annual Hawaii International Conference on System Sciences; 1998; Hawaii: IEEE Computer Society; 1998. p. 97.

[19] Lim CS, Lee KJ, Kim GC, Multiple sets of features for automatic genre classification of web documents. *Information Processing & Management*. 2005;41(5):1263-76.

[20] Sebastiani F, Machine learning in automated text categorization. *ACM Computing Surveys*. 2002;34 (1):1-47.

[21] Kessler B, Numberg G, Schütze H. Automatic detection of text genre. Association for Computational Linguistics Morristown, NJ, USA; 1997. p. 32-8.

[22] Shepherd M, Polanyi L. Genre in digital documents minitrack. Proceedings of the 32nd Hawaii International Conference on System Sciences; 2000; Hawaii; 2000.

[23] Gibson JJ, *The ecological approach to visual perception*. 2nd ed. New Jersey: LEA; 1986.

[24] Watt SNK. Text categorisation and genre in information retrieval. In: Goker A, Davies J, Graham M, editors. Information retrieval: Searching in the 21st Century: John Wiley & Sons; In Press.