Genre classification

By Daisy Abbott and Yunhyong Kim, University of Glasgow

1. Introduction

Genre classification is the process of grouping objects together based on defined similarities such as subject, format, style, or purpose.

Genre classification as a means of managing information is already established in music (e.g. folk, blues, jazz) and text and is used, alongside topic classification, to organise materials in the commercial sector (the children's section of a bookshop) and intellectually (for example, in the Usenet newsgroup directory hierarchy). However, in the case of text, genre is not a well-defined notion (it is better defined in music and arts) and discussions over what exactly constitutes genre abound in the classification community. The characterisation of information using the notion of genre may not be as explicit in other forms of material but, nevertheless, it implicitly permeates the way we view and segment the information space surrounding us.

Despite the fact that the notion of genre classification is still shrouded in ambiguity, it seems clear that with it we are striving towards a document typology which is different from topical classification. Document genre, as with music, pertains to style and/or form. The style and form of a document is constructed to meet the functional requirements within target community in realising predefined objectives of document creation. Thus genre is closely bound to organisational objectives (e.g. describing preliminary results of a research experiment), process, or activity (e.g. publication, conference) as well as the medium of dissemination (e.g. publisher, archive).

Work in genre classification has been discussed as a useful step toward achieving automated semantic metadata extraction. However, automated genre classification tools are still in the developmental stages and a prototype tool for genre classification may involve considerable investigation into the relationship between different feature types and different classes.

2. Short-term Benefits and Long-term Value

- Genre classification will continue to be used to organise objects or information both visually and conceptually to assist users.
- Genres indicate objectives of document creation and use, and, therefore support directed information retrieval ranked in terms of factors other than topical relevance.
- Recently, there has been an increasing interest in automated genre classification. This is an essential step in managing documents according to organisational activities (e.g. meetings, publications, financial records) and domain specific interests. Vast numbers of digital objects are being created daily. Metadata providing succinct descriptions for these objects is a fundamental requirement in the efficient and effective management of the objects within digital repositories. The manual collection of metadata is expensive, labour intensive, time consuming, and inconsistent. In response, automated metadata extraction has been attempted in selected genres such as scientific articles or Web pages.

- Document properties which determine document genre intersect with those that defined dialogue acts (i.e. intentional aspects of discourse such as "statement", "question", and "declaration") and, as such, would enhance document understanding and further mining of information.
- In addition to intelligently searching and retrieving information, genre classification creates conceptual links between different objects which can be used to enhance browsing functionality and can be further developed into personalised retrieval or marketing tools (for example, product recommendations based on previous purchases). This is related to the issue of information retrieval; human users make connections between documents based on several criteria within genre (e.g. objectives at the time of creation) creating defining elements of the document which could be useful to identify for subsequent users and purposes.

Current developments in automated genre classification have the further short-term benefits of:

- Binding together previously developed tools of specific genres to build a generic tool;
- Narrowing down the task of automated metadata extraction to selected genres;
- Creating automated tools for enhancing a deep understanding of documents.

3. e-Science Perspective

By encouraging a modular approach to information extraction which exploits distributed resources (e.g. previously designed tools), automated genre classification conforms to the distributed computing architecture prevalent in the e-Science domain. For example we could exploit distributed resources in the form of the creation of a metadata extraction tools repository or registry which can be called upon by a genre classification manager for the best possible tool for metadata extraction. The distributed architecture leads to focused extraction of information raising the quality level of metadata collected. The technique promotes effective knowledge management in scientific research, by opening up possibilities of intelligent information search and text mining techniques.

4. Issues to be Considered

- In the case of text, the library classification system already reflects some fragments of genre classes (e.g. fiction, non-fiction), whilst the library classification schemas tend to introduce confusion by using a mixture of these classes with subject classes (e.g. mathematics or physics). This raises the question of understanding which properties fundamentally define document genre, and how classifications can differ dependent on whether style, form or content is the primary criterion.
- Genres and schemas are differentiated automatically using different types of features which may be extracted from the text or derived from visual layout as well as contextually honed features (e.g. the source of the collection or context of creation). It is a challenge to design a system which will arrive at a dependable classifier for a range of documents as there can be domain/genre transfer problems for some indicators. Current techniques are effective in single domains but are error-prone when tested on multidisciplinary data.
- Language analysis for genre classification must necessarily vary across different languages and can require different algorithms to be effective.
- The massive volume of data available is both an opportunity and a challenge for genre classification. The deluge of digital information will eventually make it impossible for management to be carried out solely on a manual basis. In response, we need to investigate all avenues of incorporating automated management of information. The use of automated genre classification to collect further metadata is only a first step in a range of possibilities. It could also be incorporated into appraisal of digital material as genre classification can place an object within the context of organisational activities and therefore lead to a measurement of value or the means of arriving at a evaluative process for appraisal. These possibilities need to be explored, tested and scrutinised fully by digital curators and information scientists.

5. Additional Resources

- Biber, D. "*Dimensions of Register Variation: a Cross-Linguistic Comparison*". Cambridge University Press, New York (1995)
- Finn, A. & Kushmerick, N. "*Learning to classify documents according to genre*". Workshop on Computational Approaches to Style Analysis and Synthesis (2003)
- Freund, L., Clarke, C.L.A. & Toms, E.G. "*Genre classification for IR in the workplace*". Information Interaction in Context, Copenhagen, Denmark, October 2006
- Kim, Y. & Ross, S. "Searching for Ground Truth: a Stepping Stone in Automated Genre Classification" in Thanos et al. (eds.) Proceedings DELOS Conference on Digital Libraries, LNCS 4877, Springer (2007), pp. 248-261
- Kim, Y. & Ross, S. "Examining Variations of Prominent Features in Genre Classification" in Proceedings 41st Hawaiian International Conference on System Sciences, IEEE Computer Society Press, (2008)
- Ross, S., Kim, Y. and Dobreva, M. "Preliminary framework for designing prototype tools for assisting with preservation quality metadata extraction for ingest into digital repository". Pisa, DELOS NoE, December 2007
- Rosso, M. "What type of page is this?: genre as web descriptor" (2005)
- Santini, M. " State-of-the-art on Automatic Genre Identification" Technical Report ITRI-04-03 (2004), ITRI, University of Brighton (UK)

- Yoshioka, T. et al. "Genre Taxonomy: A Knowledge Repository of Communicative Actions" in ACM Transactions on Information Systems, 19: 4 (2001), pp. 431-456 Web Genre

Digital Curation Centre
Appleton Tower, 11 Crichton Street, Edinburgh, EH8 9LE | t. +44 (0)131 651 1239

The DCC is funded by the Joint Information Systems Committee