Overview of Automatic Genre Identification[1]

Mitja Luštrek
Jožef Stefan Institute, Department of Intelligent Systems
Jamova 39, 1000 Ljubljana, Slovenia
http://dis.ijs.si/mitjal, mitja.lustrek@ijs.si

Technical report IJS-DP-9735
15 January 2007

1. Introduction

Genre is a category of artistic, musical, or literary composition characterized by a particular style, form, or content (according to Merriam-Webster Online Dictionary [2] – among two other, less relevant definitions). It can be debated whether a genre can be defined by style (and form) alone, or must content (topic) also be taken into account. For example, can science fiction be distinguished from all other literary genres only by style or must one also consider that it talks about spaceships and time travel? But for the purpose of information retrieval, topic component of genre can and should be disregarded, because topic is usually treated separately and is characterized by keywords (and other methods), so genre should be orthogonal to topic. This report is written with a general corpus such as WWW in mind. Genres that cannot easily be distinguished by style are probably too fine-grained anyway for a corpus like that: it would be nice to be able to search for science fiction, but at this granularity, there are simply too many genres. It is probably best to limit oneself to genres such as homepage, FAQ, scientific paper, etc. Another common way to describe the genre of a document is through the purpose of the document, but looking for a way to automatically detect the purpose of a document leads back to style.

Genres can also be viewed through the so-called facets [9, 22]: attributes such as complexity of language, amount of graphics, subjectivity, etc. Genres can be defined by these facets: scientific paper, for example, is a genre with relatively complex language, moderate amount of graphics and low subjectivity. Facets can also be informative by themselves, but unless that is one's goal, it is probably better to identify genres directly. The advantage of facets is that they can tell something about documents with unknown genre. The disadvantage is that common genres are something users are more familiar with than facets. They may get used to them, though – users have been reported to find coloring documents that share stylistic properties useful, even though the color does not correspond to a predefined genre [27].

Methods of automatic genre identification found in the literature can be roughly divided into three classes:

1. traditional (a set of features is extracted from documents, a classification algorithm is used on the features; there is a lot of variation in this class);
2. character-based (documents are modeled as a sequence of characters);
3. visual (documents are represented as bitmaps; typically used on scanned documents).

2. Traditional Methods

The selection of features and classification algorithm are relatively independent tasks, so they will be treated separately. More attention is usually given to the selection of features, because it appears that as long as one has good features, several classification algorithms can be expected to give reasonable results.

2.1. Features

Document features can be grouped by various criteria. There seem to be no established way to do it, but the grouping used in this report is rather common, although some researchers name the groups differently.

2.1.1. Surface Features

These are the features pertaining to the text (rather than formatting) that are easy to extract and require no sophisticated parsing.

Function words are words that have little meaning, such as prepositions, pronouns, grammatical articles etc. They serve to express grammatical relationships with other words or specify the attitude of the speaker. As such

[1] This report is derived from a report prepared for Alvis, a European FP6 STREP (http://www.alvis.info)

they are to a large degree detached from the topic of a document, but not from its style. Function words are often treated as stop words, i.e. words that are ignored in various text processing tasks. Usually a preexisting list of function words is used. Koppel et al. [23], however, developed an automated way of finding function words (and other features for style-based text categorization). They used it for authorship attribution, but it should be useful for genre identification as well. They introduced a measure of feature instability: a feature is unstable if it can be replaced without changing the meaning of a document. This makes unstable features suitable for style-based categorization. To determine instability, multiple versions of the same document were needed. They were acquired by translating a document into several languages and then back to English using a number of machine translation applications. Instability of words and POS trigrams was determined using Reuters-21578 corpus [3]. Instability multiplied by the frequency in Reuters-21578 corpus turned out to be a good measure of feature suitability.

Genre-specific words, phrases, and punctuation marks are very intuitive features. They are usually prepared manually and are the easiest way of using human knowledge for genre identification. But they can also be found automatically. Dewdney et al. [10] employed information gain for this purpose. For text categorization, information gain measures the number of bits of information gained, with respect to deciding the class to which a document belongs, by each word's frequency of occurrence in the document [25]. In other words, words with high information gain are strong indicators that a document belongs to a given genre. Information gain of all the words in the dataset was measured and those for which it was high enough were used as features.

Classes of words or phrases, such as dates, times, postal addresses and telephone numbers, are somewhat similar to genre-specific words or phrases. They must be chosen manually and are typically described by regular expressions supplied to a suitable parser.

Vocabulary richness can be expressed as the number of different words per the number of words in a document, the number of once- or twice-occurring words in a document or in some similar way. The majority of vocabulary-richness measures are text-length dependent and are considered unstable for documents of less than 1,000 words [32], so they are rarely used.

All words and punctuation marks, possibly stemmed and without stop words, have also seen some use in genre identification, but not much. This is not surprising, since they are the opposite of function words, which are the established feature for genre identification. All words are often used in topical text categorization, which is another reason one would not expect them in genre identification, since genre is sometimes considered to be orthogonal to topic.

Word length is very easy to measure, so it is a common feature. Both average word length and the number of long words (typically over six characters) are used.

Sentence complexity is usually expressed as sentence length, typically in words, but readability statistics, can also be used. Flesch Reading Ease score [15] is an example of such a readability statistic and is defined as 206.835 – (1.015 × average sentence length in words) – (84.6 × average number of syllables per word).

Document length can be measured in characters, words, phrases, sentences or paragraphs.

2.1.2. Structural Features

Structural features require at least a part-of-speech (POS) tagger, if not a more sophisticated parser. Most researchers do not go beyond POS, perhaps because there are robust POS taggers freely available (for example TreeTagger [4]), while more advanced linguistic analysis can be difficult and there is no evidence that is has a significant impact on genre identification.

Parts of speech, such nouns, verbs, adjectives, adverbs ... Frequencies of single POS, POS bigrams and trigrams can be used. Particularly trigrams seem to be popular; Santini [28] showed they outperform bigrams and single POS. Some researchers eliminated POS that were too common or too rare – Argamon et al. [6] only used POS trigrams appearing in between 25% and 75% of the documents in their dataset. Koppel et al. [23] used unstable trigrams as features (feature instability was described when discussing function words).

Phrases, such as noun phrase or verb phrase: their frequencies and sizes.

Tense of verbs: the frequencies of all or selected tenses, and tense transitions from one verb to the next.

Sentence types: the frequencies of declarative, imperative and question sentences.

Parser-specific: the size of the parse tree, the number of alternative parse trees per sentence (which indicate ambiguities the parser was unable to resolve) and the frequency of sentences the parser was unable to parse successfully are the features one would expect to be available in most syntactic parsers.

2.1.3. Presentation Features

These features chiefly describe the appearance of a document (although token type also pertains to the content). As such, must of them cannot be extracted from plain text document – they are most commonly used with HTML or TeX documents.

Token type is the most general presentation feature and can be used with any written text. It measures the percentage of a document taken by upper- and lowercase characters, numbers, punctuation marks, whitespace etc. It can also measure the percentage of different types of words, such as all lowercase, capitalized, all uppercase, numbers etc.

Graphical elements: the frequencies of images, tables, equations etc. The size of these elements can also be taken into account. Which graphical elements are available depends on the document format: TeX, for example, has equations, while HTML does not. This means that some elements cannot be used when the dataset is format-wise heterogeneous.

Links are usually HTML hyperlinks, although HTML is not the only format with links. Their frequency and whether they point to the source document's domain or to another domain can be taken into account.

(Other) HTML tags are can also be used as features, as can virtually any format-specific information in a document (for example the number of styles in a MS Word document or access restrictions in a PDF document).

2.1.4. Other Features

Topic segments: a document can consist of a number of segments covering distinct topics. Karlgren [19] used the number of segments as a feature. The number of segments was determined by as determined by TextTiling [17, 18], a method for partitioning text into multi-paragraph units by using patterns of lexical connectivity to find coherent topical segments.

URL can of course only be used with documents downloaded from the internet and it only makes sense when they are downloaded from diverse sources. Lim et al. [24] used URL depth, presence of a filename at the end of the URL, document type (HTML, ASP, PHP etc.), top-level domain and genre-specific words in the URL (faq, news, board, detail etc.) as features.

2.2. Classification Algorithms

It is beyond the scope of this report to describe the classification algorithms themselves. This is the list of algorithms reported in the literature to have been used for classification of document into genres:

- decision trees etc.: Ripper, C4.5 (most popular);
- naive Bayes;
- SVM;
- discriminant analysis;
- regression: logistic regression, multiple regression;
- neural networks: two- and three-layer perceptrons;
- nearest neighbor: TiMBL;
- clustering: self-organizing map.

2.3. Features / Classification Algorithms Table

The following table gives an overview of features and classification algorithms used in the work surveyed for this report.

		Decision trees etc.	Naive Bayes	SVM	Discriminant analysis	Regression	Neural networks	Nearest neighbor	Clustering
Surface	function words	[6, 12, 19]			[21, 29]	[22, 29]			[27]
	genre-specific words, phrases, and punctuation	[8, 10, 12, 19]	[10]	[5, 10]	[21, 29]	[22, 29]	[22]	[24]	[27]

	classes of words							[24]	
	vocabulary richness							[24]	
	all words	[12]			[30]				
	word length	[8, 12, 19]			[21]	[22]	[22]	[24]	[27]
	sentence complexity	[10, 12, 19]	[10]	[10]	[21, 29]	[22, 29]	[22]	[24]	[27]
	document length	[12]			[21]			[24]	[27]
Structural	POS	[6, 10, 12]	[10, 28]	[10]	[21, 29]			[24]	
	phrases					[29]		[24]	
	tense	[10]	[10]	[10]	[21]				
	sentence type							[24]	
	parser-specific	[19]			[29]	[29]		[24]	
Presentation	token type	[8, 10, 12, 19]	[10]	[10]	[29]	[22, 29]	[22]	[24]	
	graphical elements	[8]						[24]	[27]
	links	[8]						[24]	[27]
	(other) HTML tags							[24]	
Other	topic segments	[19]							
	URL							[24]	

2.4. Datasets, Genres and Accuracy Table

The following table gives an overview of the datasets used and the resulting classification accuracy reported in the work surveyed for this report.

Experiment	**Dataset and genres**	**Accuracy and notes**
Argamon and Dodick [5]	460 papers in experimental and historical science (on average 230 per genre).	83 % – 91 % (SVM)
Argamon et al. [6]	800 articles belonging to four journalistic genres (200 per genre).	78.1 % – 84.3 % (Ripper) Function words were in most cases better than POS.
Bretan et al. [8] Dewe et al. [11] Karlgren et al. [20]	1,358 web pages belonging to 11 web genres (on average 123 per genre).	~ 70 % (C4.5)
Dewdney et al. [10]	9,705 texts belonging to seven diverse genres (on average 1,386 per genre): advertisements, bulletin	83.1 % (naive Bayes) 87.8 % (C4.5) 92.1 % (SVM)

	boards, radio news...	With naive Bayes, words with high information gain were better than other features; with C4.5 and SVM, other features are better.
Finn [12] Finn and Kushmerick [13] Finn et al. [14]	796 subjective and objective articles (on average 398 per genre).	85 % – 88 % when trained and tested on the same topic 67 % – 82 % when trained and tested on different topics, POS best (C4.5)
Finn [12] Finn and Kushmerick [13]	1,372 positive and negative reviews (686 per genre).	61.3 % – 82.7 % when trained and tested on the same topic, all words best 47.1 % – 47.8 % when trained and tested on different topics (C4.5)
Karlgren and Cutting [21]	500 texts from Brown Corpus [16] belonging to 2 first-level genres, 4 second-level (sub-)genres and 15 third-level (sub-sub-)genres (on average 33 per third-level genre).	95.6 % for the first-level genres 73.2 % for the second-level genres 52 % for the third-level genres (discriminant analysis)
Kessler et al. [22]	499 texts from Brown Corpus belonging to six diverse genres (on average 83 per genre): reportage, scientific and technical, fiction...	61 % (logistic regression) 75 % (two-layer perceptrons) 71 % (three-layer perceptrons)
Kessler et al. [22]	499 texts from Brown Corpus belonging to four brows (on average 125 per genre).	44 % (logistic regression) 47 % (two-layer perceptrons) 54 % (three-layer perceptrons)
Kessler et al. [22]	499 narrative and non-narrative texts from Brown Corpus (on average 249 per genre).	78 % (logistic regression) 82 % (two-layer perceptrons) 86 % (three-layer perceptrons)
Lim et al. [24]	1,224 Korean web pages belonging to 16 web genres (on average 76 per genre).	73.9 % – 75.6 % depending on the choice of features and parts of HTML documents taken into account (TiMBL)
Rauber and Müller-Kögler [27]	1,000 newspaper articles, genres not specified in advance.	No formal evaluation, but results appear to be sensible and users find color-coding genres useful. (self-organizing map)
Santini [28]	150 documents from British National Corpus [1] belonging to 2 first-level genres and 10 second-level (sub)genres (15 per second-level genre).	98.4 % – 99.3 % for the 2 first-level genres 81.1 % – 87 % for all 10 genres (naive Bayes) In most cases it was better to ignore punctuation. Using only some of the trigrams was also helpful. Written genres turned out to be more difficult to categorize. Trigrams performed better than bigrams, which performed better than single POS.
Stamatatos et al. [29]	250 texts belonging to 10 diverse genres (25 per genre): press editorial, academic prose, recipes...	82 % (both discriminant analysis and multiple regression)
Stamatatos et al. [30]	Part of the Wall Street Journal	≤ 97 % (discriminant analysis)

	corpus belonging to four journalistic genres.	Best results are achieved with 8 most frequent punctuation marks and 15–35 most frequent words.

3. Character-Based Methods

The advantage of this class of methods is that they view documents merely as a sequence of characters, which makes them easy to implement compared to other methods that use more complex features. They are language-independent and can be used on languages with no word delimitation, such as Chinese or Japanese. They can be used for many text categorization tasks, such as language identification or authorship attribution.

Peng et al. [26] used character-level *n*-grams to build a model of each genre. P ($c_i \mid c_{i-n+1} \ldots c_{i-1}$), where c_i is the *i*-th character in a document, was approximated by frequencies of character sequences appearing in the training documents. This was done for each genre in the set *G* to obtain the genre's model. New document of length *N* was supplied to each genre's model and it was classified into genre *g**, the one according to whose model the document has the highest probability:

$$g^* = \prod_{i=0}^{N-1} \arg\max_{g \in G} \left(P_g \left(c_i \mid c_{i-n+1} \ldots c_{i-1} \right) \right)$$

Experiments were performed on documents belonging to 10 different genres with 20 documents per genre. Genres were press editorial, academic prose, recipes... Classification accuracy with $n = 2$, which was the best setting for this dataset, was 86% and compares favorably to 82% reported on the same dataset in [29].

Teahan [31] categorized documents using cross-entropy. Entropy of a language with probability distribution *L* is:

$$H(L) = -\lim_{m \to \infty} \frac{1}{m} \sum P(c_0 \ldots c_{m-1}) \log P(c_0 \ldots c_{m-1})$$

Since it can only be computed when message length approaches infinity, it is generally not known. Its upper bound, however, can be computed if a model M of the language is available:

$$H(L, M) = -\lim_{m \to \infty} \frac{1}{m} \sum P_M (c_0 \ldots c_{m-1}) \log P_M (c_0 \ldots c_{m-1})$$

This upper bound is called cross-entropy. A model was constructed for each genre. New document was supplied to each genre's model and it was classified into genre whose model fit best, i.e. it had the lowest cross-entropy on the document.

Models were constructed using prediction by partial match (PPM) text compression scheme. PPM builds a series of models where each model assigns a probability to a character based on *k* characters before it; $-1 \leq k \leq n$. Probabilities are expressed as frequencies of previously encountered character sequences. For every $k > -1$, there is an escape probability indicating a length-*k* context that has not been seen before, which invokes the model with the next smaller value of *k*. Model for $k = -1$ assigns equal probability to every character in the coding alphabet, including those not encountered before.

Experiments were performed on 20.000 articles evenly distributed between 20 newsgroups. The author claimed these to belong to different genres, but they should probably be considered different topics. Nevertheless, the method could just as easily work on genres. Classification accuracy with $n = 5$ was 82.1%.

4. Visual Methods

Typical usage of this class of methods is in a document retrieval system in office environment. Genre identification is used to allow further steps in the document-processing pipeline to use genre-specific features. The input are scanned documents without OCR processing.

Bagdanov & Worring [7] divided each document into text zones and determined their position, size, and font size. Relations between neighboring zones were described by attributed relational graphs (ARGs), where nodes are text zones and edges Voronoi neighbor relations. ARGs of training documents of each genre were generalized into a random ARG (actually into a first order random graph – FORG – a simplified version of random ARG), which is an ARG with probabilistic node and edge interpretations. To classify a new document, its ARG was computed. Genre whose FORG could be instantiated to the new document's ARG with the highest

probability was assigned to the document. Experiments were performed on 150 documents belonging to 11 genres. Genres were different kinds of business reports and product brochures. Classification accuracy was 91%.

5. Discussion

Traditional methods work well on running text. Most useful features in such cases tend to be POS. Relatively high accuracy can apparently be reached with small training sets [28, 29]. However, in most experiments each genre consisted of very homogenous documents. If all the documents of a genre come from, say, a single magazine, one must wonder whether the classifier that is supposed to recognize the genre does not recognize the magazine instead. The experiments did not commit quite so blatant errors, but still, there is little information on genre identification of very heterogeneous documents. Also, not much work has been done on genres not involving running text, where presentation features are probably more important than surface and structural. An exception to both are Bretan, Dewe, Karlgren et al. [8, 11, 20], Dewdney et al. [10] and Lim et al. [24]. Information on the first group's work beyond the preparation of the dataset is scarce. Dewdney et al. [10] were quite successful, but they used a larger dataset than most. The results of Lim et al. [24] are worse, but their dataset seems more difficult than Dewdney et al.'s, so they could also be considered successful; this can likely be attributed to their extremely comprehensive set of features. So the conclusion is that most genre-identification tasks can be tackled with traditional methods, but more work is required for the difficult ones than one may expect judging from most of the literature.

Character-based methods are pretty straightforward to implement and seem to be useful for just about any text categorization task. They can probably be trusted to perform well on any running text characterized by features that may be subtle, but appear throughout the text. While characters-based methods do look very promising, there is not enough evidence to judge whether they are generally useful for genre identification.

Little information is available on visual methods. They appear to be reasonably successful, but they are used in a very different fashion than traditional and character-based methods, so comparison is difficult. The idea of splitting a document into sections is something one would expect to also see outside of visual methods, but for some reason this is not the case. Possibly because most of the experiments are performed on documents that tend to contain no significantly different sections.

References

[1] British National Corpus. http://www.natcorp.ox.ac.uk/

[2] Merriam-Webster Online Dictionary. http://www.m-w.com/

[3] Reuters-21578 text categorization test collection. http://www.daviddlewis.com/resources/testcollections/reuters21578/

[4] University of Stuttgart, Institute for Computational Linguistics. TreeTagger. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[5] Argamon, S. and Dodick, J. (2004). Conjunction and Modal Assessment in Genre Classification: A Corpus-Based Study of Historical and Experimental Science Writing. AAAI Spring Symposium on Attitude and Affect in Text

[6] Argamon, S., Koppel, M. and Avneri, G. (1998). Routing Documents According to Style. First International Workshop on Innovative Information Systems

[7] Bagdanov, A. D. and Worring, M. (2001). Content-Free Document Genre Classification Using First Order Random Graphs. Proceedings of Seventh Annual Conference of the Advanced School for Computing and Imaging, Heijen, The Netherlands

[8] Bretan, I., Dewe, J., Hallberg, A., Wolkert, N. and Karlgren, J. (1998). Web-Specific Genre Visualization. WebNet, Orlando, USA

[9] Crowston, K. and Kwasnik, B. H. (2004). A Framework for Creating a Facetted Classification for Genres: Addressing Issues of Multidimensionality. Proceedings of Hawaii International Conference on System Science (HICSS), Big Island, Hawaii, USA

[10] Dewdney, N., VanEss-Dykema, C. and MacMillan, R. (2001). The Form is the Substance: Classification of Genres in Text. Workshop on Human Language Technology and Knowledge Management, ACL

[11] Dewe, J., Karlgren, J. and Bretan, I. (1998). Assembling a Balanced Corpus from the Internet. 11th Nordic Computational Linguistics Conference, Copenhagen, Denmark

[12] Finn, A. (2002). Machine Learning for Genre Classification. MSc thesis, University College Dublin

[13] Finn, A. and Kushmerick, N. (2003). Learning to Classify Documents According to Genre. Workshop on Computational Approaches to Text Style and Synthesis, IJCAI, Acapulco, Mexico

[14] Finn, A., Kushmerick, N. and Smyth, B. (2002). Genre Classification and Domain Transfer for Information Filtering. Proceedings of European Colloquium on Information Retrieval Research, Glasgow, UK

[15] Flesch, R. (1974). The Art of Readable Writing. Harper and Row, New York

[16] Francis, W. N. and Kučera, H. (1964). Brown Corpus. http://helmer.aksis.uib.no/icame/newcd.htm
[17] Hearst, M. (1997). TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. Computational Linguistics 23(1), pp. 33-64
[18] Hearst, M. A. TextTiling. http://elib.cs.berkeley.edu/src/texttiles/
[19] Karlgren, J. (1999). Stylistic Experiments in Information Retrieval. In Natural Language Information Retrieval (ed. T. Strzalkowski), pp. 147-166
[20] Karlgren, J., Bretan, I., Dewe, J., Hallberg, A. and Wolkert, N. (1998). Iterative Information Retrieval Using Fast Clustering and Usage-Specific Genres. Proceedings of 8th DELOS Workshop on User Interface in Digital Libraries, Stockholm, Sweden
[21] Karlgren, J. and Cutting, D. (1994). Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. Proceedings of 15th International Conference on Computational Linguistics, Kyoto, Japan, pp. 1071-1075
[22] Kessler, B., Nunberg, G. and Schuetze, H. (1997). Automatic Detection of Text Genre. Proceedings of ACL/EACL, Madrid, Spain, pp. 32-38
[23] Koppel, M., Akiva, N. and Dagan, I. (2003). A Corpus-Independent Feature Set for Style-Based Text Categorization. Proceedings of Workshop on Computational Approaches to Style Analysis and Synthesis, IJCAI, Acapulco, Mexico
[24] Lim, C. S., Lee, K. J. and Kim, G. C. (2005). Multiple Sets of Features for Automatic Genre Classification of Web Documents. Information Processing and Management 41(5), pp. 1263-1276
[25] Mitchell, T. (1997). Machine Learning. McGraw-Hill, Boston, USA
[26] Peng, F., Schuurmans, D. and Wang, S. (2003). Language and Task Independent Text Categorization with Simple Language Models. Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Edmonton, Canada, pp. 110-117
[27] Rauber, A. and Müller-Kögler, A. (2001). Integrating Automatic Genre Analysis into Digital Libraries. Proceedings of 1st ACM/IEEE-CS Joint Conference on Digital Libraries, Roanoke, USA
[28] Santini, M. (2004). A Shallow Approach To Syntactic Feature Extraction For Genre Classification. Proceedings of 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics, Birmingham, UK
[29] Stamatatos, E., Fakotakis, N. and Kokkinakis, G. (2000). Automatic Text Categorization in Terms of Genre and Author. Computational Linguistics 26(4), pp. 471-495
[30] Stamatatos, E., Fakotakis, N. and Kokkinakis, G. (2000). Text Genre Detection Using Common Word Frequencies. Proceedings of 18th International Conference on Computational Linguistics, Luxembourg, pp. 808-814
[31] Teahan, W. J. (2000). Text Classification and Segmentation Using Minimum Cross-Entropy. Proceedings of 6th International Conference "Recherche d'Information Assistee par Ordinateur" (RIAO), Paris, France
[32] Tweedie, F. and Baayen, H. (1998). How Variable May a Constant Be? Measures of Lexical Richness in Perspective. Computers and the Humanities 32(5), pp. 323-352