Evaluation of Different Approaches to Training a Genre Classifier

Vedrana Vidulin, Mitja Luštrek, Matjaž Gams
Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia
vedrana.vidulin@ijs.si, mitja.lustrek@ijs.si, matjaz.gams@ijs.si

Abstract

This paper presents experiments on classifying web pages by genre. Firstly, a corpus of 1539 manually labeled web pages was prepared. Secondly, 502 genre features were selected based on the literature and the observation of the corpus. Thirdly, these features were extracted from the corpus to obtain a data set. Finally, three machine learning algorithms, one for induction of decision trees (J48) and two ensemble algorithms (bagging and boosting), were trained and tested on the data set. Additionally, impact of feature selection on ensemble algorithms was tested. The best performed genre classifiers in terms of precision were selected to obtain the best of set of classifiers. On average the best of set achieved 9% better precision, but slightly worse recall. Accuracy and F-measure did not vary significantly. The results indicate that classification by genre could be a useful addition to search engines.

1. Introduction

A good question to start with is why we want to classify a web page by genre. For example, if we are interested in cats and search for the keyword "cat", a search engine will return web pages that describe the life of cats, but it will also return web pages with cat picture gallery, newspaper articles about controlling the growth of cat's population etc. (see Figure 1). However, if we were able to specify that we want to search only for content delivery type of web pages about cats, we would get more specific results in accordance with our interest, e.g. picture gallery of cats. Classification of web pages by genre would make our life easier.

What exactly is a genre? In general, a genre could be described as a style of a web page [7]. A web page is used to send a message to the user. Message has a topic, for example the life of the cats, but it also tries to communicate that topic in a specific way. To a veterinarian it will give a high number of objective facts about cats. When wishing to entertain, it will communicate the message about cats to amuse the user by presenting pictures and video material. In the light of the previous explanation genre can be described as intentional styling of a web page with the objective to communicate the topic in a specific manner.

Classification of web pages by genre is a challenging task [2, 5-9, 12, 16-20]. Even humans with their advanced semantics and understanding of concepts misclassify some web pages, therefore computer programs face a difficult task indeed.

Figure 1. Web pages of different genres obtained by posing topic keyword "cat".

Another problem is to find appropriate features, i.e. properties of a web page that adequately describe a web page in the context of genre. The quality of classifier strongly depends on the choice of features.

The corpus we experimented on is presented in Section 2. Section 3 lists the features used to describe web pages. Section 4 deals with machine learning (ML) algorithms chosen for training the classifier. Results of the experiments are given in Section 5. A conclusion is presented in Section 6.

2. 20-genre collection of web pages

20-Genre Collection was compiled at Jožef Stefan Institute and consists of 1539 web pages belonging to 20 genres. The genres are: adult, blog, childrens', commercial/promotional, community, content delivery, entertainment, error message, FAQ, gateway, index, informative, journalistic, official, personal, poetry, prose fiction, scientific, shopping, user input. Each page can belong to multiple genres.

The web pages were collected from the Internet using three methods. Firstly, we used highly-ranked Google hits for popular keywords like “Britney Spears”. The keywords were chosen according to Google Zeitgeist statistics. Our purpose was to train a classifier that will not have a problem with recognizing the most popular web pages. Secondly, we gathered random web pages. And finally, we specifically searched for web pages belonging to genres underrepresented to that point.

The corpus was manually labeled by two independent annotators. Their labels disagreed on about a third of the web pages in the corpus, so those were reassessed by a third and sometimes even a fourth annotator.

3. Genre features

There is no generally accepted set of genre features what can be seen from [2, 5-9, 12, 16-20], particularly since it depends on the type of documents under consideration. Most past research dealt with pure text with little additional information (such as formatting), so it used only text-based features. Since we were classifying web pages, we also used URL- and HTML-based features [12].

3.1. URL features

URL features are based on the structure and the content of an URL. Structural features follow URL syntax defined by [3]:

$$\underbrace{\text{foo}}_{\text{scheme}}\text{://}\underbrace{\text{example.com:8042}}_{\text{authority}}\underbrace{\text{/over/there}}_{\text{path}}\text{?}\underbrace{\text{name=ferret}}_{\text{query}}\text{\#}\underbrace{\text{nose}}_{\text{fragment}}$$

URL content is analyzed by marking the appearances of 54 words most commonly present in URL. The words were stemmed with Porter stemming algorithm [14].

76 features were obtained in total, all Boolean except for URL depth, which is numeric. Features and their descriptions are presented in Table 1.

Table 1. A set of URL features.

Feature	**Description**
Https	Indicates whether the scheme is https.
URL depth	Number of directories included in the path.
Document type	Described by four Boolean features, each indicating whether the document type is *static HTML* (document extensions html and htm), *script* (document extensions asp, aspx, php, jsp, cfm, cgi, shtml, jhtml and pl), *doc* (document extensions pdf, doc, ppt and txt) or *other* (the other document extensions).
Tilde	Appearance of “/~” in the URL.
Top-level domain	Described by ten Boolean features, each indicating whether the top-level domain is *com, org, edu, net, gov, biz, info, name, mil* or *int*.
National domain	Indicates whether the top level domain is a national one.
WWW	Indicates if the authority starts with www.
Year	Indicates the appearance of year in the URL.
Query	Indicates the appearance of query in the URL.
Fragment	Indicates the appearance of fragment in the URL.
Appearance of 54 most commonly used words in URL	Indicates the appearance of common content words in URL: *about, abstract, adult, archiv, articl, blog, book, content, default, detail, download, ebai, english, error, fanfic, faq, forum, free, fun, funni, galleri, game, help, home, index, joke, kid, legal, librari, link, list, lyric, main, member, music, new, paper, person, poem, poetri, product, project, prose, pub, public, quiz, rule, search, sport, stori, topic, tripod, user, wallpap*

3.2. HTML features

HTML features correspond to HTML tags. According to the general trend in literature [18] we grouped tags into five categories according to their functionalities. In addition, we counted the hyperlinks in the web page and separated external from internal.

In total, 7 features were chosen, all numeric and normalized (see Table 2).

Table 2. A set of HTML features.

Feature
Number of hyperlinks to the same domain / Total number of hyperlinks
Number of hyperlinks to a different domain / Total number of hyperlinks
Number of tags / Total number of tags for 5 tag groups: *1.* ***Text Formatting*** – *<abbr>, <acronym>, <address>, <b>, <basefont>, <bdo>, <big>, <blockquote>, <center>, <cite>, <code>, <del>,*

<dfn>, <em>, <font>, <h1>, <h2>, <h3>, <h4>, <h5>, <h6>, <i>, <ins>, <kbd>, <pre>, <q>, <s>, <samp>, <small>, <strike>, <strong>, <style>, <sub>, <sup>, <tt>, <u>, <var>

2. ***Document Structure*** *– <br>, <caption>, <col>, <colgroup>, <dd>, <dir>, <div>, <dl>, <dt>, <frame>, <hr>, <iframe>, <li>, <menu>, <noframes>, <ol>, <p>, <span>, <table>, <tbody>, <td>, <tfoot>, <th>, <thead>, <tr>, <ul>*
3. ***Inclusion of external objects*** *– <applet>, <img>, <object>, <param>, <script>, <noscript>*
4. ***Interaction*** *– <button>, <fieldset>, <form>, <input>, <isindex>, <label>, <legend>, <optgroup>, <option>, <select>, <textarea>*
5. ***Navigation*** *- Counting href attribute of tags <a>, <area>, <link> and <base>*

3.2. Text features

In total, 419 text features were extracted from web pages, all numeric and normalized. They are listed in Table 3.

The set of 321 content words is a combination of manually extracted content words and most common content words automatically extracted from our corpus. A punctuation symbol set is obtained equally.

Table 3. A set of text features.

Feature
Average number of characters per word
Average number of words per sentence
Number of characters in hyperlink text / Total number of characters
Number of alphabetical tokens (alphabetical token is a sequence of letters) / Total number of tokens
Number of numerical tokens (numerical token is a sequence of digits) / Total number of tokens
Number of separating tokens (separating token is a sequence of separator characters (space, return…)) / Total number of tokens
Number of symbolic tokens (symbolic token is a sequence of characters excluding alphanumeric and separator characters) / Total number of tokens
Number of content words / Total number of content words for 321 content words (stemmed by Porter stemming algorithm)
Number of function words / Total number of function words for 50 most common function words in the corpus
Number of punctuation symbols / Total number of punctuation symbols for 34 punctuation symbols
Number of declarative sentences / Total number of sentences
Number of interrogative sentences / Total number of sentences
Number of exclamatory sentences / Total number of sentences
Number of other sentences (in most cases list items) / Total number of sentences
Number of date named entities / Total number of words
Number of location named entities / Total number of words
Number of person named entities / Total number of words

4. ML problem

Weka, a collection of ML algorithms [22], was chosen as a tool for genre classification. Since the ML algorithms in Weka do not support multilabeled classification, we divided the problem into 20 binary sub-problems, one for each genre. The task was thus to train 20 classifiers, each to decide whether an input web page belongs to one of the 20 genres.

Several Weka ML algorithms were tested on the domain [20]. On the basis of their performance, J48, the Weka implementation of C4.5 [13], [15], was chosen for constructing the classifier. Besides performance, it was also selected for simplicity, transparency and speed, which were important criteria because the classifier was intended to be integrated into the Alvis search engine [1].

Due to unsatisfactory performance of standalone genre classifiers (e.g. average precision of 53% - see Table 5), metalearning algorithms [22] were employed to train ensembles of J48 decision trees. Bagging and AdaBoostM1, the Weka implementations of bagging and boosting were used.

AttributeSelectedClassifier is another metalearning algorithm that firstly performs feature selection and secondly applies ML algorithm on the subset of features [22]. Feature selection was accomplished by running Rank Search. This searching algorithm uses Gain Ratio single feature evaluator to sort features and than Cfs Subset evaluator to rank feature subsets. From ML algorithms bagging and boosting were chosen.

In total, five classifiers were trained for each genre. Except mentioned changes in algorithm parameters, default values set in Weka were used. 10-fold cross-validation [10] was used for testing, and the classifier performance is presented as accuracy, precision, recall and F-measure.

5. Results

Accuracy denotes the percentage of correctly classified examples in all the examples [10]. The results of our experiments are presented in Table 4.

Table 4. Accuracy of the genre classifiers in percent (FS = feature selection).

ACCURACY	J48	Bagg.	Bagg. – FS	Boos.	Boos. – FS	BEST
Adult	97.14	**97.79**	97.47	97.47	97.01	97.79
Blog	95.84	97.08	97.34	96.82	**97.66**	97.66
Childrens'	94.80	95.45	95.52	**95.71**	95.39	95.71
Commercial/ promotional	89.34	**92.07**	91.42	91.16	90.77	92.07
Community	95.65	96.69	96.56	**97.34**	96.95	97.34
Content delivery	90.38	**91.81**	90.64	90.84	90.45	91.81
Entertainment	94.87	95.78	**96.23**	96.10	95.71	96.23
Error message	97.47	**97.73**	97.60	97.21	97.27	97.73
FAQ	98.38	98.70	**98.83**	98.77	98.70	98.83
Gateway	93.31	**95.32**	94.74	94.54	94.22	95.32
Index	81.94	**87.39**	87.13	84.41	85.90	87.39
Informative	79.73	83.56	**85.25**	83.76	83.82	85.25
Journalistic	85.90	89.54	**89.99**	89.80	89.60	89.99
Official	96.56	**97.01**	96.88	96.88	96.75	97.01
Personal	91.49	93.24	**94.15**	93.96	93.50	94.15
Poetry	97.01	97.27	**97.92**	97.60	97.79	97.92
Prose fiction	95.39	96.17	96.49	**97.01**	96.56	97.01
Scientific	95.78	97.08	97.14	97.21	**97.34**	97.34
Shopping	94.80	96.36	96.10	**96.69**	96.56	96.69
User input	95.71	**97.08**	**97.08**	96.62	96.69	97.08
Average	93.07	94.66	94.72	94.50	94.43	**95.02**
DIFF. - Average	1.95	0.36	0.29	0.52	0.58	**0.74**

The best genre classifiers are marked by highlighting a table cell and are abstracted in the last column. The last two rows show the average performance of all genre classifiers trained by particular ML algorithm and the difference between that value and averaged performance of the best of set of genre classifiers.

The differences between ML algorithms are insignificant. Slight advantage can be given to bagging. Feature selection did not have considerable impact on the performance. The best of set of classifiers would achieve on average only 0.74% better performance, which is also insignificant.

However, accuracy is not the most suitable performance measure in our setting, because for each genre, negative examples far outnumber positive examples. A classifier that would assign no genre to any web page would have a high accuracy, because most web pages indeed do not belong to most genres.

Because of the unbalanced datasets and small change in accuracy other standard information retrieval measures were used, i.e. precision, recall and F-measure.

Precision is the percentage of examples classified as positive that are in fact positive [21]. It is presented in Table 5.

Table 5. Precision of the genre classifiers in percent (FS = feature selection).

PRECISION	J48	Bagg.	Bagg. – FS	Boos.	Boos. – FS	BEST
Adult	66	78	**81**	79	73	81
Blog	61	83	**86**	84	85	86
Childrens'	71	81	82	**87**	74	87
Commercial/ promotional	21	**40**	33	31	33	40
Community	63	76	81	**90**	86	90
Content delivery	40	**64**	49	49	45	64
Entertainment	53	69	**80**	72	60	80
Error message	83	87	**88**	84	78	88
FAQ	85	**98**	97	**98**	95	98
Gateway	35	**45**	37	26	38	45
Index	38	63	**64**	46	54	64
Informative	31	30	41	**42**	**42**	42
Journalistic	43	62	**69**	64	61	69
Official	56	73	**81**	65	53	81
Personal	39	**72**	66	71	64	72
Poetry	72	76	83	**85**	84	85
Prose fiction	46	69	75	**87**	69	87
Scientific	62	85	85	**87**	84	87
Shopping	42	72	59	**78**	70	78
User input	63	**83**	81	79	80	83
Average	53.50	70.30	70.90	70.20	66.40	**75.35**
DIFF. - Average	21.85	5.05	4.45	5.15	8.95	**9.09**

Precision of ensembles of J48 decision trees is significantly higher that the precision of standalone J48 decision tree (13 to 17%). Differences between bagging and boosting are insignificant. Only feature selection has small negative impact on boosting. However, if we compose the set of the best of genre classifiers we would obtain the precision of 75.35%, which is an average improvement of 9%. In comparison with the best performing ensemble methods 5% advancement was obtained and in comparison with standalone J48 decision tree, the improvement of 22%.

The best of set is composed from genre classifiers trained by bagging with and without feature selection and boosting. Precision below 50% is present only in the three genres (Commercial/promotional – 40%, Gateway – 45% and Informative – 42%). This is reasonable having in mind that there are 20 genres and the process is multilabeled.

Recall is the percentage of positive examples that are classified as such [21]. It is presented in Table 6.

Table 6. Recall of the genre classifiers in percent (FS = feature selection).

RECALL	J48	Bagg.	Bagg. – FS	Boos.	Boos. – FS	BEST
Adult	61	**71**	61	61	57	71
Blog	56	56	57	49	**66**	66
Childrens'	49	48	46	44	**55**	55
Commercial/ promotional	13	4	6	10	**15**	15
Community	52	55	47	**56**	51	56
Content delivery	25	23	22	**26**	24	26
Entertainment	30	27	30	31	**36**	36
Error message	68	68	65	61	**70**	70
FAQ	**80**	73	77	74	76	80
Gateway	19	12	10	11	**28**	28
Index	32	37	**39**	32	38	39
Informative	**27**	9	10	25	26	27
Journalistic	40	36	37	39	**41**	41
Official	**29**	27	24	24	28	29
Personal	26	16	**33**	29	**33**	33
Poetry	63	61	**73**	60	66	73
Prose fiction	**39**	30	35	38	35	39
Scientific	53	51	55	53	**59**	59
Shopping	35	33	29	35	**41**	41
User input	57	57	**60**	54	52	60
Average	42.70	39.70	40.80	40.60	44.85	**47.20**
DIFF. - Average	4.50	7.50	6.40	6.60	2.35	**5.47**

Recall and precision are inversely related: as you attempt to increase one, the other tends to decline [11]. This can be seen from the Table 6. Genre classifiers trained by J48 and boosting with feature selection did show the highest average recall. These classifiers also manifested the lowest precision.

If we compose the set of the best of genre classifiers in terms of recall we would obtain on average 5% better recall.

F-measure is the weighted harmonic mean of precision and recall [21]. It is calculated as presented in Eq. 1.

$$\frac{2 \times recall \times precision}{recall + precision} \tag{1}$$

Table 7. F-measure of the genre classifiers in percent (FS = feature selection).

F-MEASURE	J48	Bagg.	Bagg. – FS	Boos.	Boos. – FS	BEST
Adult	63	**73**	67	68	62	73
Blog	57	65	67	60	**73**	73
Childrens'	56	58	58	57	**62**	62
Commercial/ promotional	16	7	9	15	**20**	20
Community	56	62	58	**68**	63	68
Content delivery	30	**33**	29	**33**	30	33
Entertainment	36	39	43	43	**44**	44
Error message	73	**75**	74	69	73	75
FAQ	81	83	**85**	84	83	85
Gateway	**22**	18	15	16	31	22
Index	34	46	**47**	37	44	47
Informative	28	14	16	31	**32**	32
Journalistic	41	45	47	**48**	**48**	48
Official	**37**	**37**	34	34	36	37
Personal	31	24	**43**	41	42	43
Poetry	66	67	**76**	69	73	76
Prose fiction	42	37	43	**51**	44	51
Scientific	55	63	65	64	**68**	68
Shopping	36	43	38	47	**50**	50
User input	59	67	**68**	62	62	68
Average	45.95	47.80	49.10	49.85	52.00	**53.75**
DIFF. - Average	7.80	5.95	4.65	3.90	1.75	**4.81**

F-measure is presented in Table 7. In terms of F-measure, boosting with feature selection is the best choice with the average F-measure of 52%. Ensembles of J48 decision trees outperformed standalone J48 decision trees (2 to 6%). The set of the best of genre classifiers performed 5% better.

Considering demands of the genre classification task, to train the classifier that will be precise in labeling web pages, the set of the best of genre classifier in terms of precision will be used. Table 8 presents the implication of that choice on accuracy, recall and F-measure.

Table 8. Performance of the set of genre classifiers with the highest precision (in percent).

	Accuracy	Precision	Recall	F-measure
Adult	97.47	81	61	67
Blog	97.34	86	57	67
Childrens'	95.71	87	44	57
Commercial/ promotional	92.07	40	4	7
Community	97.34	90	56	68
Content delivery	91.81	64	23	33
Entertainment	96.23	80	30	43
Error message	97.60	88	65	74
FAQ	98.74	98	74	84
Gateway	95.32	45	12	18
Index	87.13	64	39	47
Informative	83.79	42	26	32
Journalistic	89.99	69	37	47
Official	96.88	81	24	34
Personal	93.24	72	16	24
Poetry	97.92	85	60	69
Prose fiction	96.49	87	38	51
Scientific	97.14	87	53	64
Shopping	96.10	78	35	47
User input	97.08	83	57	67
Average	**94.77**	**75.35**	**40.50**	**49.95**
Overall improvement	**0.49**	**9.09**	**-1.23**	**1.01**

Accuracy, recall and F-measure did not change significantly. Accuracy and F-measure slightly increased and recall slightly decreased (on average 1%).

6. Conclusion

Because of the huge variety of the web and because genres are difficult to define in a machine-understandable way, classification of web pages by genre is a challenging task. Until now our approach was to choose single ML algorithm and to train 20 genre categories with the same algorithm. However, improvement in performance can be obtained by training each genre classifier with different ML algorithm. By combining bagging, bagging with fetaure selection and boosting, 9% improvement in precision is obtained. For the use in a search engine, where web pages need to be labeled with a genre, precision is much more critical than recall, because it is more problematic if a page is mislabeled than if it is not labeled at all. Independent real-life experiments with the Alvis prototype [4], where the genre classifier was implemented as part of the search engine, confirmed that the classifier's performance is satisfactory.

7. References

[1] Alvis, 2007, http://www.alvis.info/alvis/ [01/23/2007]

[2] S. Argamon, M. Koppel, G. Avneri, "Routing Documents According to Style", *First International Workshop on Innovative Information Systems*, 1998.

[3] T. Berners-Lee, R.T. Fielding, L. Masinter, Uniform Resource Identifier (URI): Generic Syntax, Internet Society, RFC 3986, STD 66, 2005.

[4] W. Buntine, Private communication, 2007.

[5] N. Dewdney, C. VanEss-Dykema, R. MacMillan, "The Form is the Substance - Classifications of Genres in Text", 1998.

[6] A. Finn, "Machine Learning for Genre Classification", 2002.

[7] J. Karlgren, *Stylistic Experiments for Information Retrieval*, PhD thesis, 2000.

[8] J. Karlgren, D. Cutting, "Recognizing Text Genres with Simple Metrics Using", *Proceedings of the 15th. International Conference on Computational Linguistics*, 1994.

[9] B. Kessler, G. Nunberg, H. Schütze, "Automatic Detection of Text Genre", *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 1997.

[10] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", *IJCAI*, 1995.

[11] A. Large, L.A. Tedd, R.J. Hartley, *Information seeking in the online age: Principles and practice*, London: Bowker, 1999.

[12] C.S. Lim, K.L. Lee, G.C. Kim, "Multiple sets of features for automatic genre classification of web documents", *Information Processing & Management*, 2005.

[13] T.M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.

[14] M. Porter, "The Porter Stemming Algorithm", 2007, http://www.tartarus.org/~martin/PorterStemmer/ [01/10/2007]

[15] J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.

[16] M. Santini, "A Shallow Approach to Syntactic Feature Extraction for Genre Classification", *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*, 2003.

[17] M. Santini, "Common Criteria for Genre Classification: Annotation and Granularity", *Workshop on Text-based Information Retrieval (TIR-06),* In Conjunction with ECAI 2006, Riva del Garda, 2006.

[18] M. Santini, "Description of 3 feature sets for automatic identification of genres in web pages", 2006, http://www.itri.brighton.ac.uk/~Marina.Santini/three_feature_sets.pdf [12/19/2006]

[19] E. Stamatatos, G. Kokkinakis, N. Fakotakis, "Automatic Text Categorization in Terms of Genre and Author", *Computational Linguistics*, 2000.

[20] V. Vidulin, M. Luštrek, M. Gams, "Comparison of the Performance of Genre Classifiers Trained by Different Machine Learning Algorithms", *Information Society*, Ljubljana, 2006.

[21] Wikipedia – "Information retrieval", 2007, http://en.wikipedia.org/wiki/Information_retrieval [01/28/2007]

[22] I.H. Witten, E. Frank, *Data Mining – Practical Machine Learning Tools and Techniques*, Elsevier Inc., 2005.