Journal of Computing and Information Technology - CIT 15, 2007, 4, 305–311
doi:10.2498/cit.1001137

Training a Genre Classifier for Automatic Classification of Web Pages

Vedrana Vidulin, Mitja Luštrek and Matjaž Gams

Jožef Stefan Institute, Ljubljana, Slovenia

This paper presents experiments on classifying web pages by genre. Firstly, a corpus of 1 539 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, two machine learning algorithms, one for induction of decision trees (J48) and one ensemble algorithm (bagging), were trained and tested on the data set. The ensemble algorithm achieved on average 17% better precision and 1.6% better accuracy, but slightly worse recall; F-measure did not vary significantly. The results indicate that classification by genre could be a useful addition to search engines.

Keywords: genre classification, web page, genre features, ensemble algorithm

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 elephants and search for the keyword "elephant", a search engine will return web pages that describe the life of elephants, but it will also return web pages with elephant picture gallery, newspaper articles about saving the elephants in Africa etc. (see Figure 1). However, if we were able to specify that we want to search only for journalistic materials about elephants, we would get more specific results in accordance with our interest. 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 elephants, but it also tries to communicate that topic in a specific way. To a zoologist, it will give a high number of objective facts about elephants. When wishing to entertain, it will communicate a message about elephants 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.

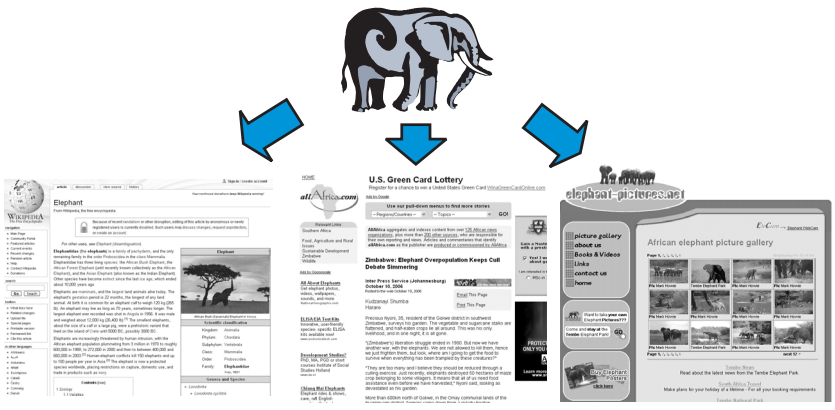


Figure 1. Web pages of different genres obtained by posing topic keyword "elephant".

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.

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

The corpus we experimented with 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 [21] was compiled at Jožef Stefan Institute and consists of 1 539 web pages belonging to 20 genres. The genres are: Blog, Children's, Commercial/Promotional, Community, Content Delivery, Entertainment, Error Message, FAQ, Gateway, Index, Informative, Journalistic, Official, Personal, Poetry, Pornographic, 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 [24]. Our purpose was to build 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.

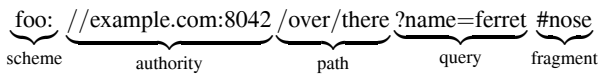
The corpus was manually labeled by two independent annotators. Their labels disagreed on about a third of the web pages in the corpus, so those were reassessed by a third and sometimes even a fourth annotator.

3. Genre Features

There is no generally accepted set of genre features, which can be seen from [2, 5-9, 12, 16-20], particularly since the feature set 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]:



URL content is analyzed by marking the appearances of 54 words most commonly present

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, port, stori, stopic, tripod, user, wallpap*

Table 1. A set of URL 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>, , <cite>, <code>, <del>, <dfn>, , <font>, <h1>, <h2>, <h3>, <h4>, <h5>, <h6>, <i>, <ins>, <kbd>, <pre>, <q>, <s>, <samp>, <small>, <strike>, , <style>, <sub>, <sup>, <tt>, <u>, <var>*
2. **Document Structure** – *
, <caption>, <col>, <colgroup>, <dd>, <dir>, , <dl>, <dt>, <frame>, <hr>, <iframe>,   , <menu>, <noframes>,   , <p>, , <table>, <tbody>, <td>, <tfoot>, <th>, <thead>, <tr>,   *
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>*

Table 2. A set of HTML features.

in URL. The words were stemmed with Porter stemming algorithm [14].

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

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.

We have chosen 7 HTML features, all numeric and normalized (see Table 2).

3.3. Text Features

From web pages, 419 text features were extracted, 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.

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

Table 3. A set of text features.

4. ML Problem

Weka, a collection of ML algorithms [23], was chosen as the 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. Each page was typically assigned 2–3 genres, but the number varied from 1 to 10 or more.

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 the 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].

We used J48 to build standalone decision trees and to construct bagging ensembles [23]. Although ensemble classifiers are more complex and thus demand more time, they are often beneficial in terms of tradeoff between additional time and improved performance.

5. Results

For the experiments, J48 and bagging were run with the default Weka parameters. 10-fold cross validation [10] was used for testing. The classifier performance was measured by accuracy, precision, recall and F-measure. The statistical significance test of J48 vs. bagging was performed using the corrected resampled paired t-test with significance level of 5%.

Accuracy is the degree of conformity of a measured quantity to its actual value. In our experiments it denotes the percentage of correctly classified web pages [10]. The results of the experiments are presented in Table 4. The differences between the performance of classifiers built by J48 and by bagging in terms of accuracy are on average 1.58%. Considering the results of paired t-test, bagging outperformed J48, performing significantly better in 7 genres and equally well in other 13 genres. However, accuracy is not the most suitable performance measure in our setting, because for each genre the corpus contains higher number of web pages that do not belong to observed genre. 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. Therefore, other standard information retrieval

	J48	Bagging	Diff.
Blog	95.84	97.08	1.24
Children's	94.80	95.45	0.65
Commercial/Promotional	89.34	92.07	2.73
Community	95.65	96.69	1.04
Content Delivery	90.38	91.81	1.43
Entertainment	94.87	95.78	0.91
Error Message	97.47	97.73	0.26
FAQ	98.38	98.70	0.32
Gateway	93.31	95.32	2.01
Index	81.94	87.39	5.45
Informative	79.73	83.56	3.83
Journalistic	85.90	89.54	3.64
Official	96.56	97.01	0.45
Personal	91.49	93.24	1.75
Poetry	97.01	97.27	0.26
Pornographic	97.14	97.79	0.65
Prose Fiction	95.39	96.17	0.78
Scientific	95.78	97.08	1.3
Shopping	94.80	96.36	1.56
User Input	95.71	97.08	1.37
Average	93.07	94.66	1.58
Paired T-Test		(7/13/0)	

Table 4. Accuracy of the genre classifiers in percent.

	J48	Bagging	Diff.
Blog	61	83	22
Children's	71	81	10
Commercial/Promotional	21	40	19
Community	63	76	13
Content Delivery	40	64	24
Entertainment	53	69	16
Error Message	83	87	4
FAQ	85	98	13
Gateway	35	45	10
Index	38	63	25
Informative	31	30	-1
Journalistic	43	62	19
Official	56	73	17
Personal	39	72	33
Poetry	72	76	4
Pornographic	66	78	12
Prose Fiction	46	69	23
Scientific	62	85	23
Shopping	42	72	30
User Input	63	83	20
Average	53	70	17
Paired T-Test		(5/15/0)	

Table 5. Precision of the genre classifiers in percent.

measures are needed: precision, recall and F-measure.

Precision is the proportion of retrieved and relevant web pages to all web pages retrieved from the corpus [22]. In our experiments it denotes the percentage of web pages classified as positive, that are in fact positive. It is presented in Table 5. In 11 genres the precision of both classifiers was higher than 50%, which sounds reasonable, having in mind that there are 20 genres and the process is multilabeled. For 6 genres in particular (Content Delivery, Index, Journalistic, Personal, Prose Fiction and Shopping) the precision was significantly improved by the use of the bagging algorithm. In two genres (Commercial/Promotional and Gateway), the situation did improve, but the precision still stayed below 50%. Only in the Informative genre bagging performed worse, but the difference was insignificant (1%). The overall improvement by bagging was highly significant, on average 17%. Considering results of paired t-test, bagging outperforms J48, performing significantly better in 5 genres and equally well in other 15 genres.

Recall is the proportion of relevant web pages that are retrieved out of all relevant web pages available in the corpus [22]. In our experiments it denotes the percentage of positive web pages classified as such. Recall and precision are inversely related: as you attempt to increase one, the other tends to decline [11]. This can be seen in Table 6. The increase in precision gained by using the bagging algorithm resulted in a decline of recall in 14 genres. Recall was improved only for three genres (Community, Index and Pornographic), but in the cases of Community and Index not significantly. Three genres did not manifest any change in recall. Considering results of paired t-test bagging performed worse in one genre and had the same level of performance in other 19 genres.

	J48	Bagging	Diff.
Blog	56	56	0
Children's	49	48	-1
Commercial/Promotional	13	4	-9
Community	52	55	3
Content Delivery	25	23	-2
Entertainment	30	27	-3
Error Message	68	68	0
FAQ	80	73	-7
Gateway	19	12	-7
Index	32	37	5
Informative	27	9	-18
Journalistic	40	36	-4
Official	29	27	-2
Personal	26	16	-10
Poetry	63	61	-2
Pornographic	61	71	10
Prose Fiction	39	30	-9
Scientific	53	51	-2
Shopping	35	33	-2
User Input	57	57	0
Average	43	40	-3
Paired T-Test		(0/19/1)	

Table 6. Recall of the genre classifiers in percent.

	J48	Bagging	Diff.
Blog	57	65	8
Children's	56	58	2
Commercial/Promotional	16	7	-9
Community	56	62	6
Content Delivery	30	33	3
Entertainment	36	39	3
Error Message	73	75	2
FAQ	81	83	2
Gateway	22	18	-4
Index	34	46	12
Informative	28	14	-14
Journalistic	41	45	4
Official	37	37	0
Personal	31	24	-7
Poetry	66	67	1
Pornographic	63	73	10
Prose Fiction	42	37	-5
Scientific	55	63	8
Shopping	36	43	7
User Input	59	67	8
Average	46	48	2
Paired T-Test		(1/18/1)	

Table 7. F-measure of the genre classifiers in percent.

F-measure is the weighted harmonic mean of precision and recall [22]. The purpose of using F-measure is to obtain a single measure that characterizes performance of the classifier [23]. It is calculated as presented in Eq. 1.

$$\frac{2 \times recall \times precision}{recall + precision} \quad (1)$$

F-measure is higher than 50% in 9 genres (see Table 7). In spite of using the bagging algorithm, F-measure did not improve enough to pass the 50% value in 11 genres. The use of bagging resulted in poorer performance of 5 genre classifiers (Commercial/Promotional, Gateway, Informative, Personal and Prose Fiction). F-measure remained unchanged only for the Official genre. Performance did improve in 14 genres, but the improvement was significant only for the Index and Pornographic genres. Considering results of paired t-test bagging performed better in one genre, worse in one genre and had the same level of performance in other 18 genres.

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. However, we have managed to achieve a reasonable precision, particularly with bagging, which brought a 17% improvement over standalone J48. 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.

References

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

[2] S. ARGAMON, M. KOPPEL, G. AVNERI Routing Documents According to Style. Presented at the *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: classification of genres in text. Presented at the *Proceedings of the workshop on Human Language Technology and Knowledge Management – Volume 2001*, (2001) Toulouse, France.

[6] A. FINN Machine Learning for Genre Classification. MSc. Thesis, University College, Dublin, Ireland, 2002.

[7] J. KARLGREN Stylistic Experiments for Information Retrieval. PhD. Thesis, Swedish Institute of Computer Science, Stockholm, Sweden, 2000.

[8] J. KARLGREN, D. CUTTING Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. Presented at the *Proceedings of the 15th. International Conference on Computational Linguistics – Volume 2*, (1994) Kyoto, Japan.

[9] B. KESSLER, G. NUMBERG, H. SCHÜTZE Automatic Detection of Text Genre. Presented at the *Proceedings of the 35th annual meeting on Association for Computational Linguistics*, (1997) Madrid, Spain.

[10] R. KOHAVI A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Presented at the *IJCAI*, (1995) Montréal, Québec, Canada.

[11] A. LARGE, L. A. TEDD, R. J. HARTLEY *Information seeking in the online age: Principles and practice*. Bowker-Saur, London, UK, 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*, Vol. 41, No. 5 (2005), pp. 1263–1276.

[13] T. M. MITCHELL *Machine Learning*. McGraw-Hill, USA, 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, San Francisco, CA, USA, 1993.

[16] M. SANTINI A Shallow Approach to Syntactic Feature Extraction for Genre Classification. Presented at the *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*, (2003) Brighton, UK.

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

[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*, Vol. 26, No. 4 (2000), pp. 471–495.

[20] V. VIDULIN, M. LUŠTREK, M. GAMS Comparison of the Performance of Genre Classifiers Trained by Different Machine Learning Algorithms; Presented at the *Proceedings of the 9thInternational Multiconference Information Society*, (2006) Ljubljana, Slovenia.

[21] WEB GENRE DATASET, 2007, http://dis.ijs.si/mitjal/genre/ [04/30/2007].

[22] WIKIPEDIA – INFORMATION RETRIEVAL, 2007, http://en.wikipedia.org/wiki/Information_retrieval [01/28/2007].

[23] I. H. WITTEN, E. FRANK *Data Mining-Practical Machine Learning Tools and Techniques*. Morgan Kaufmann-2nd edition, San Francisco, CA, 2005.

[24] GOOGLE ZEITGEIST, 2005, http://www.google.com/press/zeitgeist.html [06/25/2005].

Received: June, 2007
Accepted: September, 2007

Contact addresses:
Vedrana Vidulin
Jožef Stefan Institute
Jamova 39
SI-1000 Ljubljana
e-mail: vedrana.vidulin@ijs.si

Mitja Luštrek
Jožef Stefan Institute
Jamova 39
SI-1000 Ljubljana
e-mail: mitja.lustrek@ijs.si

Matjaž Gams
Jožef Stefan Institute
Jamova 39
SI-1000 Ljubljana
e-mail: matjaz.gams@ijs.si

VEDRANA VIDULIN received her university degree in 2005 from the Faculty of Philosophy, University of Rijeka, Croatia, by defending her thesis "Neural Networks: Algorithms and Applications in Education". She is continuing her education at the Jožef Stefan International Postgraduate School, study program New Media and e-Science, and is receiving scholarship from the Jožef Stefan Institute, Ljubljana, Slovenia. Her research interests include machine learning and data mining, especially text mining.

MITJA LUŠTREK received his Ph. D. degree in computer and information science in 2007 from the Faculty of Computer and Information Science, University of Ljubljana, Slovenia. He is now a researcher at Jožef Stefan Institute, Department of Intelligent Systems. His research interests include heuristic search, computer game playing and machine learning, particularly for text processing. He has published his papers in a number of scientific journals, including Artificial Intelligence and ICGA Journal. He is an executive editor of Informatica journal and a member of executive board of Slovenian Artificial Intelligence Society.

MATJAŽ GAMS is an Associate Professor of computer and information science at the University of Ljubljana and a Senior Researcher at the Jožef Stefan Institute, Ljubljana, Slovenia. He teaches several courses in computer science at graduate and postgraduate levels at Faculties of Computer and Information Science, Economics, etc. His research interests include artificial intelligence, intelligent systems, intelligent agents, machine learning, cognitive sciences, and information society. In his publication list there are over 300 items, 50 of them in scientific journals. He has headed several major artificial intelligence applications in Slovenia, including the major national employment agent on the Internet, and the Slovenian text-to-speech system donated to several thousand users.