Training the Genre Classifier for Automatic Classification of Web Pages

Vedrana Vidulin, Mitja Luštrek, Matjaž Gams
Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana
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, 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 the 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 the 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 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 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.

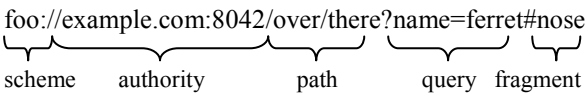
The corpus was manually labeled by two independent annotators. Their labels disagreed on about a third of the web pages in the corpus, so those were reassessed by a third and sometimes even a fourth annotator.

3. Genre Features

There is no generally accepted set of genre features 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]:



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>, , <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>*

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. 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 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 not only to build standalone decision trees, but also to construct bagging ensembles [22]. Although ensemble classifiers are more complex and thus demand more time, only experiments can show the 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.

Accuracy denotes the percentage of correctly classified examples in all the examples [10]. The results of our 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%. 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. Therefore, other standard information retrieval measures are needed: precision, recall and F-measure.

Table 4. Accuracy of the genre classifiers in percent.

	J48	BAGGING	DIFF.
ADULT	97.14	97.79	0.65
BLOG	95.84	97.08	1.24
CHILDRENS'	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
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

Precision is the percentage of examples classified as positive that are in fact positive [21]. 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) precision significantly improved by the use of the bagging algorithm. In two genres (Commercial-promotional and Gateway) situation did improve, but the precision still stayed below 50%. Only in the Informative genre bagging did perform worse, but the difference was insignificant (1%). The overall improvement by bagging was highly significant, on average 17%.

Table 5. Precision of the genre classifiers in percent.

	J48	BAGGING	DIFF.
ADULT	66	78	12
BLOG	61	83	22
CHILDRENS'	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
PROSE-FICTION	46	69	23
SCIENTIFIC	62	85	23
SHOPPING	42	72	30
USER-INPUT	63	83	20
AVERAGE	53	70	17

Recall is the percentage of positive examples that are classified as such [21]. 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 (Adult, Community and Index), but in the case of Community and Index not significantly. Three genres did not manifest any change in recall.

Table 6. Recall of the genre classifiers in percent.

	J48	BAGGING	DIFF.
ADULT	61	71	1
BLOG	56	56	0
CHILDRENS'	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
PROSE-FICTION	39	30	-9
SCIENTIFIC	53	51	-2
SHOPPING	35	33	-2
USER-INPUT	57	57	0
AVERAGE	43	40	-3

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

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

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 Adult and Index genres.

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

	J48	BAGGING	DIFF.
ADULT	63	73	10
BLOG	57	65	8
CHILDRENS'	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
PROSE-FICTION	42	37	-5
SCIENTIFIC	55	63	8
SHOPPING	36	43	7
USER-INPUT	59	67	8
AVERAGE	46	48	2

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.

7. References

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

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

[3] Berners-Lee T, Fielding RT, Masinter L. Uniform Resource Identifier (URI): Generic Syntax. Internet Society. RFC 3986; STD 66; 2005.
[4] Buntine W. Private communication; 2007.
[5] Dewdney N, VanEss-Dykema C, MacMillan R. The Form is the Substance - Classifications of Genres in Text; 1998.
[6] Finn A. Machine Learning for Genre Classification; 2002.
[7] Karlgren J. Stylistic Experiments for Information Retrieval. PhD thesis; 2000.
[8] Karlgren J, Cutting D. Recognizing Text Genres with Simple Metrics Using. Proceedings of the 15th. International Conference on Computational Linguistics; 1994.
[9] Kessler B, Nunberg G, Schütze H. 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] Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. IJCAI; 1995.
[11] Large A, Tedd LA, Hartley RJ. Information seeking in the online age: Principles and practice. London: Bowker; 1999.
[12] Lim CS, Lee KL, Kim GC. Multiple sets of features for automatic genre classification of web documents. Information Processing & Management; 2005.
[13] Mitchell TM. Machine Learning. McGraw-Hill; 1997.
[14] Porter M. The Porter Stemming Algorithm; 2007, http://www.tartarus.org/~martin/PorterStemmer/ [01/10/2007]
[15] Quinlan JR. C4.5: Programs for Machine Learning. Morgan Kaufmann; 1993.
[16] Santini M. 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] Santini M. 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] Santini M. 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] Stamatatos E, Kokkinakis G, Fakotakis N. Automatic Text Categorization in Terms of Genre and Author. Computational Linguistics; 2000.
[20] Vidulin V, Luštrek M, Gams M. 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] Witten IH, Frank E. Data Mining – Practical Machine Learning Tools and Techniques. Elsevier Inc.; 2005.