Classifying Web corpora into domain and genre using automatic feature identification

Serge Sharoff [1]
Centre for Translation Studies
University of Leeds, LS2 9JT UK

Abstract

Texts in representative corpora are typically classified into their domain and genre. However, it is not clear if existing domain and genre typologies can be applied at all to unlabeled data collected from the Web, for instance, to results of crawling. This study attempts to establish the most suitable categories for describing domains and genres of arbitrary web texts and to estimate the accuracy of their automatic classification using machine learning methods, such as Support Vector Machine (SVM) and clustering (repeated bisections and graph clustering). We also discuss methods for inducing the most discriminative features to perform this classification. The method has been designed to work with few or no linguistic resources and has been validated on a variety of languages: English, German, Chinese and Russian.

Keywords : corpus annotation, domains, genres, clustering.

1. Introduction

Texts in representative corpora (Brown, BNC) are classified according to a number of parameters, as it it important to know what the corpus is composed of. Most typically this classification is created manually and covers at least the domain and genre of constituent texts. The Web is a natural source of linguistic data and many corpora are produced from it, see the Special Issue of *Computational Linguistics* (Kilgarriff & Grefenstette 2003). However, texts collected from the Web, e.g. the SPIRIT collection (Joho & Sanderson 2004), lack metadata describing their domain and genre. As for automatic categorization, the Web presents an additional challenge, since Web texts are much more diverse than text collections like Reuters or Wall Street Journal, which are frequently used as benchmarks. In addition to news items the Web contains personal webpages, discussion forums, academic articles, advertisements, tutorials, and so on, all referring to a wider range of topics.

The rationale of this study is to provide the possibility of telling the domain and genre of an arbitrary web document. This can be used for various purposes, such as improving the relevance of information retrieval or selecting more appropriate language models in POS tagging, parsing, machine translation, or in word sense disambiguation, cf. (Karlgren 2005; McCarthy *et al.* 2004).

However, there is no universally accepted typology for classifying texts in traditional corpora and Internet alike. For instance, the BNC classifies text domains using nine categories, such as applied, natural or social sciences, politics, etc. Texts in the Brown corpus are not classified into domains proper, while EAGLES corpus recommendations present a list of 36 categories (Sinclair 1996). Catalogues available

[1] `s.sharoff@leeds.ac.uk`

Corpus	I-EN	I-DE	I-RU	I-ZH	Brown	BNC	NG20
Documents	71636	44783	30346	71153	500	4054	11952
Tokens (millions)	181	188	188	282	1.16	111	3.66
Average length	2532	4193	6194	3959	2320	27441	306

Table 1. Parameters of corpora used in the study

on the Internet itself also differ in their sets of categories. For instance, the classification used by Yahoo (`dir.yahoo.com`) consists of 14 top-level categories, such as Computers, Education, Government, Recreation, etc, while the classification of 15 domains used by the Open Directory Project (`dmoz.org`) drops Education and Government and adds Games, Kids and Shopping. It is not clear which typology better represents the actual range of Web texts for an individual language.

It is even more difficult to establish any accepted genre classification. The Brown corpus famously uses 15 categories, such as press reportage or religious texts. The BNC contains a typology consisting of 70 genres, e.g. academic texts in the humanities or reportage sections in national broadsheet newspapers (Lee 2001). As yet another example, DWDS, a German corpus comparable to the BNC, classifies texts into four genres: fiction, journalism, instructive and technical texts. Computational research on genre detection propagates the set of categories even further. With few exceptions, e.g. (Karlgren & Cutting 1994), who relied on genres from the Brown corpus, such studies, e.g. (Biber 1988; Kessler *et al.* 1997; Argamon *et al.* 1998; Rehm 2002), design their own typologies.

This paper presents several experiments addressing the following questions:

1. what is the most suitable typology to classify webpages into their domains and genres?
2. how to generate the best set of features for detection of domains and genres within this typology?

The goal of this study is to estimate the values of these parameters for large Web-derived corpora. Another goal is to generalise the method by applying it to several languages. The answer to the first question should provide the set of top-level categories for classifying webpages, which might differ from typologies used in traditional corpora and human estimates made in Internet catalogues. The resulting typology might be also language- and culture-specific. The answer to the second question should provide the possibility of rapid development of classifiers for domain and genre detection without involving significant human effort. The possibility of their rapid development is especially important for languages other than English, for which additional resources, such as annotated datasets, POS taggers or even stoplists, might be not available.

The general setup for the experiments reported below involves the following steps:

1. generate a set of features for each text;
2. identify a subset of more indicative features;
3. cluster texts using this subset to identify the resulting typology.

In Section 2. we study the performance of domain detection by extraction of keywords and unsupervised clustering on webcorpora. In Section 3. we study automatic induction of reliable features for genre detection. For Web corpora the study uses freely available random snapshots of the Internet for English, Chinese German and Russian (Sharoff 2006). Their parameters are described in Table 1 (I-DE, I-EN, I-RU and I-ZH refer to German, English, Russian and Chinese Internet corpora respectively). In order to evaluate the performance of feature selection and clustering, we also apply the procedure to three

Entropy	sci	comp	talk	rec	misc	sport
0.241	2	23	0	1	0	0
0.075	151	0	1	3	0	0
0.317	1	0	150	9	12	3
0.038	0	2	0	1	0	266
0.308	9	2	1	127	6	1
0.419	30	291	7	5	27	3

Table 2. Clustering of NG20 (Entropy: 0.248)

annotated corpora: the Brown Corpus, BNC and the 20 Newsgroups corpus, so their parameters are also listed (the token counts include punctuation marks). Machine learning tools used are CLUTO (Zhao & Karypis 2004), Chinese Whispers (Biemann 2006), and Weka (Witten & Frank 2005).

2. Detection of domains

2.1. Method

It is well-known that domains can be identified using keywords (Sebastiani 2002) or keyword N-grams (Mladenic 1998). The accuracy of the state-of-the-art methods in text categorization reaches 90-95% for English (Sebastiani 2002), even though few results are available for other languages, cf. (Kornai *et al.* 2003). However, this performance is achieved using large relatively homogeneous training sets with known domains, while experiments in automatic induction of domain typologies, e.g. (Merkl 1998), have not been conclusive.

The set of keywords describing individual documents in a heterogenous Web corpus is not known in advance (in a Reuters-type corpus it can be derived as a by-product of existing classification). At the same time, the total number of features (words) in a Web corpus is too large to process it without any feature selection. So, we used the following unsupervised method for reducing the number of features. For each document in a text collection we generated keywords by comparing the document frequency list against the frequency list for the whole corpus using the log-likelihood score. In comparison to more commonly used TF-IDF, MI and Chi-scores, the LL-score is more robust in cases of few occurrences of search terms in a document (Rayson & Garside 2000). Given that the procedure was intended to be language-independent, stop words were defined as the 500 most frequent words in each corpus. Then keywords have been sorted by the number of documents for which they are relevant and the top keywords are used as the feature list. Documents represented by selected keyword lists are clustered.

We tried two clustering algorithms: repeated bisections (using CLUTO and the cosine similarity measure between documents) and graph clustering (using Chinese Whispers). The repeated bisections (RB) algorithm needs to know the number of clusters, so it was run in iterations from 7 to 11 clusters to detect top-level domains. For graph clustering, documents are represented as nodes connected to adjacent nodes according to the number of keywords they share. The advantage of graph clustering is that it does not require a preset number of clusters. However, it usually results in a large number of smaller clusters.

For Internet corpora we produce qualitative interpretation of the results by treating each cluster as a subcorpus, which can be described by its own keywords (we again use the LL-score for this task). The positive side effect is that the procedure gives a semi-automatic method for producing lists of keywords to classify new documents when they are added to the corpus (the procedure is semi-automatic, as manual filtering of keyword lists is required).

Comp	I-EN	file, software, object, file, computer, digital, system, user, text, Fibonacci, Windows, server, code, MPlayer, image, page, int, Internet, model, mobile, output, number, Fortran, BibTeX, interface, network, programming, button, function, HTML, C, map	8.72%
	I-DE	Datei (file), Software, Rechner (Computer), Version, Internet, Server, Gerät (device), Windows, Benutzer (user), Apple, verwenden (to use), Linux, installieren (install), Datenbank (database), PC, speichern (save), Installation, E-Mail	9.69%
	I-RU	сеть (network), устройство (device), диск (disk), сервер (server), функция (function/feature), версия (version), сообщение (message), Windows, компьютерный (computer), программный (program), ноутбук (laptop), документ (document), процессор (processor)	7.84%
Law	I-EN	employer, employee, labour, tribunal, act, disabilities, discrimination, collective, maternity, equality, job, complaint, legislation, entitled, shall, agreement, training, dismissal, tax, safety, government, policy, appellate, arrangement, statutory, provision, bill	4.26%
Left	I-DE	Revolution, Arbeiterklasse (working class), Demokratie (democracy), Regierung (government), Bewegung (movement), Linke (Left), Kapitalismus, Politik, Lenin, Kommission, Mitgliedstaat (member state), EU, Kampf (fight), sozialistisch (socialist), Globalisierung (globalisation), Trotzki, Bourgeoisie	4.09%
WWII	I-RU	противник (enemy), аэродром (airfield), фронт (front), армия (army), войска (troops), офицер (officer), сбивать (to shoot down), авиация (aviation), товарищ (comrade), майор (major), атака (assault), немец (German), фашистский (fascist)	4.27%
Travel	I-ZH	亚 (Asia), 山 (mountain), 长城 (Great Wall), 桥 (bridge), 桂林 (Guilin), 门票 (entry ticket),寺 (pagoda), 景点 (scenic spot), 酒店 (pub), 司机 (driver), 导游 (tour guide), 旅游 (travel), 岛 (island), 汤 (soup), 房 (room), 公里 (km), 丽江 (Lijiang), 南 (south), 啤酒 (beer), 船 (boat)	9.69%

Table 3. Examples of domain clustering using RB

2.2. Results

Table 2 presents the results of six-way clustering of the NG20 corpus using the RB algorithm (rows corresponds to clusters, columns in the right part to the number of documents per cluster). As our task was to evaluate the ability of our method to generalise over top-level categories, the 20 newsgroups were combined into six larger groups. A class that has not been found by this method is the `misc.forsale` newsgroup, which became absorbed partly into the `comp` cluster (computer games and hardware are also advertised there), partly into the `talk` cluster. At the same time the class of `comp` newsgroups has been split into two clusters. The smaller one (Cluster 1 in Table 2) corresponds to `comp.os.ms-windows` only, while the cluster corresponding to the rest of computing (Cluster 6) also absorbs some messages from `sci.cryptography` and has the highest entropy score.

This proves that our unsupervised feature selection method can produce reasonable results, at least for English (no large domain-annotated Web corpus is readily available for other languages). Another goal of this study was to detect the set of categories reflecting data found on the Web. Existing Internet classifications, such as `dmoz.org` or Google, cannot be used for validation, as each of them uses a significantly different set of categories, and this is the parameter we would like to test.

Table 3 shows keyword lists for some clusters produced by RB clustering (the proportion these clusters take in the entire text collections is in its last column). It is typically possible to interpret them in terms of traditional corpus categories. Domain clusters consistently produced in our clustering experiments

Corpus	Descriptive words	Size
I-DE	Infektion, Virus, Krankheit (disease), Erreger (pathogen), Antikörper (antibody), HIV, Impfung (vaccine), Patient, Medikament (medicine), Bakterie, Immunsystem (immune system)	353
I-RU	заболевание (disease), вирус (virus), инфекция (infection), препарат (medicine), терапия (therapy), спид (AIDS), вакцина (vaccine), иммунный (immune), инфицировать (to infect)	298
I-ZH	SARS, 肺炎 (pneumonia), 病人 (patient), 疫情 (epidemic), 感染 (infection), 隔离 (isolate), 病毒 (virus), 卫生 (hygiene), 消毒 (sterilise)	327

Table 4. An example of clusters produced by Chinese Whispers

in the four languages are those referring to domestic and international news (it is frequently split into language-specific subdomains, e.g. Australian and British news for English, Swiss and Austrian news for German, former USSR news for Russian, Taiwan news for Chinese), Commerce, Health, Computers, Sciences, Arts, Sports and Religion. At the same time, clustering found clusters more specific to individual languages, such as the legal cluster for English, travel cluster for Chinese or WWII cluster for Russian.

Graph-based clustering by Chinese Whispers (Table 4) produced in total about 6,000 clusters for each language, which varied in size from about 1,000 documents to clusters consisting of single documents (for English there were 5570 clusters, but only 47 of them were larger than 100 documents). In comparison to RB clustering, CW clusters are more specific in terms of their domains, e.g. news from Iraq, contagious diseases or star wars. They cannot help in choosing the set of top-level categories for corpora or webportals, but they can be useful as a tool for creating keyword lists for individual domains without the need of using large annotated datasets. Again, even if the two clusters can be compared across languages in terms of their general label, their content can concern country-specific topics, such as SARS and sanitation for the contagious diseases cluster in Chinese.

3. Genre detection

3.1. Method

Punctuation marks and POS trigrams have proven their effectiveness for detection of genres (Argamon *et al.* 1998; Santini 2007). However, these studies, as well as other studies on genre or style detection, e.g. (Biber 1988; Karlgren & Cutting 1994; Kessler *et al.* 1997; Finn & Kushmerick 2006), used hand-crafted feature lists which were known to provide a better diagnostics for English, while very few studies have been done for other languages, (Braslavski 2004) is one of few exceptions. In this study we attempted to apply the same feature generation mechanism as used in our domain studies to genre detection. We also attempted to test genre detection on languages other than English. As mentioned above, the field of genre studies is plagued with confusing terminologies and typologies; human annotators also find it sometimes difficult to assign a unique genre label to a given webpage (Santini 2007). The evaluation of clustering results is also more difficult for genres: in domain detection it is possible to evaluate a cluster informally by looking at the list of its keywords, while the set of POS trigrams describing a cluster gives little information for intuitive interpretation.

As one way of evaluating our method we replicated the three classic (Karlgren & Cutting 1994) experiments with the Brown corpus. Karlgren and Cutting created a hierarchy of the 15 genres listed in the Brown corpus. In one experiment they joined all fiction texts (e.g. humour or romance) to form one group and left other categories intact; this made 10 categories (in the tables below it is referred to as KC-10). Then (experiment KC-4) they kept fiction as a single category and created three more

Experiment	No test set	Cross-validation
KC-2	95.6%	–
BC-2	96.6%	95.0%
BC-2m	100.0%	96.6%
KC-4	73.2%	–
BC-4	81.6%	73.2%
BC-4m	100%	78.4%
KC-10	64.4%	–
BC-10	80.8%	67.1%
BC-10m	100.0%	59.0%
BNC-5	100.0%	94.0%
BNC-5m	100.0%	95.8%
I-EN5	99.1%	82.1%
I-EN5m	99.1%	61.2%
I-RU5	81.0%	58.6%
I-RU5m	100.0%	58.6%

Table 5. Genre detection accuracy

general categories: press, non-fiction (it included government documents and academic texts) and miscellaneous (all other genres). Finally, they grouped the texts into two categories: fiction and informative texts (KC-2). Their detection procedure was based on linear discriminant analysis using a hand-crafted list of 20 features, such as counts of adverbs or present participles, or counts of individual words, such as *therefore* or *which*.

In our experiments with the Brown corpus we used exactly the same set of genres as they did, used Weka's implementation of SVM for supervised learning, and designed two methods for producing feature sets. One involves unsupervised collection of POS trigrams that are more specific for individual texts in the same way as used for domain detection (BC-2, BC-4, BC-10). Another method (BC-2m, BC-4m, and BC-10m) tests the case when a POS tagger for the language we would like to work with is not available. In this case we can treat the first N most frequent word forms as their own POS tags. All other words receive the empty tag (this follows the assumption that function words are typically more frequent than content words). For English such trigrams look like: `X is X`, `X to the`, `X X that`. A similar technique of replacing POS tags with function words was used for learning to distinguish between original texts and translations in (Baroni & Bernardini 2006).

The Internet collections we used had small samples of 170 documents for English and for Russian classified into five classes of the main aim of text production according to the EAGLES guidelines: discussion, recommendation (advertising and reviews), instruction (tutorials), information (reference materials) and texts for recreational reading. The genre classification in the BNC uses 70 categories, some of which can be mapped onto the EAGLES classes, for instance, academic and non-academic papers can be treated as ‘discussions’, fiction, biographies and popular lore as ‘recreational’ texts. Thus, for evaluation purposes we also use 828 BNC files which can be unambiguously mapped onto the EAGLES text classes to study the possibility of their detection on a text set larger than the annotated Internet samples.

3.2. Results

Table 5 lists the results of supervised learning of genre detection. Karlgren and Cutting used the whole Brown corpus for both training and testing, so the figures for our experiments in the middle column of Table 5 (testing on the training set) are artificially high. However, even with cross validation (the

BC-4m

a	b	c	d	<– classified as
67	2	0	19	a = press
3	72	0	35	b = nonfiction
1	0	120	5	c = fiction
10	26	7	133	d = miscellaneous

BNC-5m

a	b	c	d	e	<– classified as
195	0	0	3	10	a = discussion
1	23	3	3	5	b = information
1	1	13	0	0	c = instruction
1	1	1	50	1	d = recommendation
3	0	1	0	509	e = recreation

Table 6. Confusion matrices for genre detection

Entropy	recreation	instruct	discuss	recomm	info
0.000	0	0	0	0	33
0.000	274	0	0	0	0
0.027	137	0	0	1	0
0.596	10	0	27	3	2
0.132	86	0	5	0	0
0.488	3	2	70	19	1
0.573	3	13	106	31	2

Table 7. Clustering of genres in BNC-5 (Entropy: 0.212)

last column) automatically detected features are as good as those selected manually by Karlgren and Cutting (though this can be also explained by the better efficiency of SVM). The confusion matrix for our experiment BC-4m (Table 6) shows that categories in general are well separated with the exception of ‘miscellaneous’, which actually is a collection of several genres, as well as domains (it combines the Brown categories corresponding to religion, hobbies, popular lore and belles lettres). The confusion matrix for distinguishing between text aims for the BNC shows a better separation between them. However, the classification accuracy drops for Internet corpora, as samples used for training are too small, e.g. 112 for I-EN with only six examples of recreational texts, which is not enough for reliable genre detection.

The number of word forms retained as POS tags (N) depends on the language. For instance, inflective languages, such as Russian, have more word forms, so it is sensible to use a longer list of ‘POS tags’ to cover their morphological variation. In general the greater granularity of function words used instead of tags provides more information for the classification algorithm, but it increases data sparseness. To evaluate the influence of this factor, we measured the accuracy of cross-validation for different values of N (Figure 1). The maximum for English is achieved at around 120 and 140 (for BC and I-EN) and again 700 and 800, for Russian the maximal values were achieved by 80 and 600, but the variation is not significant and might be attributed to the difference in data used in each corpus.

Finally, we clustered the 828 texts from the BNC using CLUTO and the same set of POS trigrams as in our experiment BNC-5 (Table 7). The samples from the Internet corpora used in the study did not produce reasonable results because of the small number of training instances. Even if this experiment did not test the genre distribution of Internet corpora, at least it evaluated the validity of categories used to describe genres.

Clustering managed to identify the class of informative texts, and produced three separate clusters for recreational texts. It was less successful in separating discussions, instructions and recommendations. This further proves the discriminative power of POS trigrams, as the recreation class has been produced by combination of fiction, popular biographies and popular lore from the BNC genre set. However, even if RB clustering separated academic and non-academic articles (cluster 6 consists mostly of academic articles), the entropy of these clusters is higher.

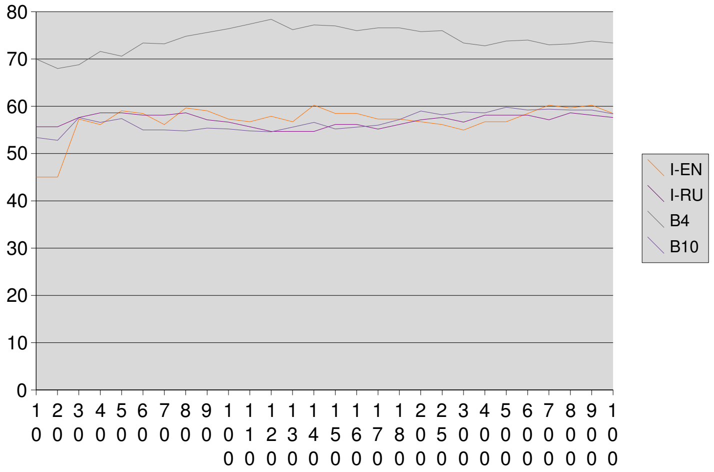

Figure 1. Tagging using function words

4. Conclusions

The goal of this study was to establish the usefulness of language-independent methods for classifying unrestricted Web documents into their domain and genre. The study proves that a language-independent automatic procedure can produce reliable classifiers of domains and genres of Web texts.

The second goal of this study was to find the set of categories for Web classification depending on the nature of the corpus and culture. A culture-independent typology is advisable to compare results produced for different languages, but it might miss certain dimensions important for some languages.

The study also gives some statistics on the relative proportion of texts in individual domains. For instance, texts clustered as the computers and hardware cluster constitute about 8 % of I-EN and I-RU, and 10% of I-DE, on the basis of a random snapshot of the Internet in those languages. Nevertheless, data clustering is not a substitute for development of a typology: in the end the categories depend not only on texts, but also on ways they are going to be used. The category **Reference** (referring to encyclopedias, dictionaries, etc) cannot be described in terms of keywords and, therefore, it is unlikely to be found by their clustering (actually it corresponds to the *genre* of informative texts), but there can be other reasons for keeping it on the front page of `dmoz.org`. In a similar way, clustering can justify reasons for having the category Sport on the front page by showing that it is one of the most frequent topics with its own clearly defined lexicon, but from the user viewpoint there can be reasons for considering it as a subordinate concept within other types of entertainment. Anyway, a keyword list for this category (and its subcategories) can be reliably produced by clustering.

Classification of Web texts into genres proved to be a more challenging task, because of the small number of genre-annotated documents. Even if the figures in Table 5 for Internet samples are impressive, this is down to the fact that the majority of them are examples of discussions, so the baseline for choosing the most frequent genre is around 60% for these samples (this is unlike the Brown and BNC samples). Clustering reported in Table 7 also shows that POS trigrams, which are commonly used in genre classification, are indeed associated with generic genres: clustering did not use any knowledge about the categories to be used for genre classification, yet it identified several clusters, such as fiction, instructions, reference texts and, to a smaller degree, academic papers.

Finally the study shows that it is possible to do computational research on genres in languages with few or no resources. Further research is needed to identify features for detecting other classification categories (such as the reading difficulty), as well as to investigate more complicated methods of unsupervised feature selection, such as SVD (Berry *et al.* 1999) or rough sets (Chouchoulas & Shen 2001).

Acknowledgements

I am grateful to Marco Baroni, Adam Kilgarriff, and Katja Markert for very useful comments on the earlier drafts of the paper.

References

ARGAMON S., KOPPEL M. et AVNERI G. (1998), “Routing documents according to style”, in *Proc. International Workshop on Innovative Internet Information Systems (IIIS-98)*,Pisa.

BARONI M. et BERNARDINI S. (2006), “A new approach to the study of translationese: Machine-learning the difference between original and translated text”, in *Literary and Linguistic Computing*, n^{o} 3, vol. 21.

BERRY M., DRMAČ Z. et JESSUP E. (1999), “Matrices, Vector Spaces, and Information Retrieval”, in *SIAM Review*, n^{o} 2, vol. 41.

BIBER D. (1988), *Variations Across Speech and Writing*, Cambridge University Press.

BIEMANN C. (2006), “Chinese Whispers — an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems”, in *Proc. HLT-NAACL-06 Workshop on Textgraphs-06*.

BRASLAVSKI P. (2004), “Document style recognition using shallow statistical analysis”, in *ESSLLI 2004 Workshop on Combining shallow and deep processing for NLP* : 1–9.

CHOUCHOULAS A. et SHEN Q. (2001), “Rough Set-Aided Keyword Reduction for Text Categorisation”, in *Applied Artificial Intelligence*, n^{o} 9, vol. 15.

FINN A. et KUSHMERICK N. (2006), “Learning to classify documents according to genre”, in *Journal of the American Society for Information Science and Technology*, n^{o} 5, vol. 7.

JOHO H. et SANDERSON M. (2004), “The SPIRIT collection: an overview of a large web collection”, in *SIGIR Forum*, n^{o} 2, vol. 38.

KARLGREN J. (2005), “The Whys and Wherefores for Studying Textual Genre Computationally”, in *Proc. AAAI Fall Symposium on Style and Meaning in Language, Art and Music*,Arlington, USA.

KARLGREN J. et CUTTING D. (1994), “Recognizing text genres with simple metrics using discriminant analysis”, in *Proc. of the 15th. International Conference on Computational Linguistics (*COLING 94*)*,Kyoto, Japan : 1071 – 1075.

KESSLER B., NUNBERG G. et SCHÜTZE H. (1997), “Automatic Detection of Text Genre”, in *Proceedings of the 35th ACL/8th EACL* : 32–38.

KILGARRIFF A. et GREFENSTETTE G. (2003), “Introduction to the Special issue of the Web as Corpus”, in *Computational Linguistics*, n^{o} 3, vol. 29.

KORNAI A., KRELLENSTEIN, MULLIGAN, TWOMEY, VERESS et WYSOKER (2003), “Classifying the Hungarian Web”, in *Proc. of the European Association of Computational Linguistics, EACL 2003*,Budapest : 203–210.

LEE D. (2001), “Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle”, in *Language Learning and Technology*, n^{o} 3, vol. 5.

MCCARTHY D., KOELING R., WEEDS J. et CARROLL J. (2004), “Finding Predominant Word Senses in Untagged Text”, in *Proc. 42nd Meeting of the Association for Computational Linguistics (ACL’04)*,Barcelona : 279–286.

MERKL D. (1998), “Text classification with self-organizing maps: Some lessons learned”, in *Neurocomputing*, n^{o} 1/3, vol. 21.

MLADENIC D. (1998), “Turning Yahoo to Automatic Web-Page Classifier”, in *Proc. European Conference on Artificial Intelligence* : 473–474.

RAYSON P. et GARSIDE R. (2000), “Comparing corpora using frequency profiling”, in *Proc. of the Comparing Corpora Workshop at ACL 2000*,Hong Kong : 1–6.

REHM G. (2002), “Towards automatic web genre identification – a corpus-based approach in the domain of academia by example of the academic’s personal homepage”, in *Proc. of the Hawaii Internat. Conf. on System Sciences*.

SANTINI M. (2007), *Automatic Identification of Genre in Web Pages*, PhD thesis, University of Brighton.

SEBASTIANI F. (2002), “Machine learning in automated text categorization”, in *ACM Computing Surveys*, n^o 1, vol. 34.

SHAROFF S. (2006), “Creating general-purpose corpora using automated search engine queries”, in Baroni M. & Bernardini S. (Eds), *WaCky! Working papers on the Web as Corpus*, Gedit, Bologna, http://wackybook.sslmit.unibo.it.

SINCLAIR J. (1996), *Preliminary recommendations on corpus typology*, Expert Advisory Group on Language Engineering Standards document EAG–TCWG–CTYP/P, `http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html`.

WITTEN I. et FRANK E. (2005), *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco.

ZHAO Y. et KARYPIS G. (2004), “Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering”, in *Machine Learning*, n^o 3, vol. 55.