

# Automatic Genre Identification: Towards a Flexible Classification Scheme

Marina Santini  
University of Brighton, Lewes Road, Brighton (UK)  
*MarinaSantini.MS@gmail.com*

**This paper presents an automatic genre classification model that implements a flexible classification scheme, i.e. a scheme capable of performing zero-, one- or multi-genre assignment. I suggest that this scheme is more appropriate for genres on the web, because many web pages have often more than one genre or none at all. The model that I propose relies on the distinction between the concepts of ‘text types’ and ‘genre’, which are both ‘inferred’ and not ‘learned’ from pre-labelled examples. The main drawback of this approach is that it cannot be fully evaluated given the limitations of current genre research. However, I present a partial evaluation that shows that the model performs competitively, and remains stable when re-scaled.**

*Keywords: genre, classification, inference, web pages*

## 1. INTRODUCTION

The term ‘genre’ has a long-standing tradition in many academic disciplines, such as literary studies, linguistics and rhetoric. From Aristotle’s *Poetics* onwards, genre definitions abound. Interestingly, this diversity does not seem to affect the restricted field of automatic genre identification studies. In this field, researchers borrow definitions from genre analysts or define genre in relation to topic, or do not provide any definition. Genre is indeed a difficult concept to handle. First, it is hard to answer the question: what is genre? The plethora of definitions indicates the difficulty of finding any consensus. Second, it is not always clear what genre classes are. Probably, for this reason in the recent ECIR 2007 and SIGIR 2007, the authors of three poster papers preferred the more neutral expression ‘document type’ (Yeung et al., 2007a; Xu et al., 2007; Yeung et al.; 2007b), rather than the overwhelmingly engaging ‘genre’. Braver in this respect is the demo paper co-authored by Yeung, Freund and Clarke, who present an enterprise search engine that exploits the relation between user’s tasks and document genres.

Generally speaking, genres are textual categories like ACADEMIC PAPERS, FABLES, EDITORIALS, or RECIPES. They have often been characterized in terms of <purpose, form>. This means that documents belonging to the same genre share the same purpose (e.g. recipes have all an instructional purpose) and the same form, either for the language and/or the layout (e.g. recipes are often organised as sequences of steps starting with imperatives). The attribute of ‘form’ makes genre classes different from topical classes, like sport, fashion, or politics, because topics or domains can be virtually shaped in many different genres. For example, we can have a political editorial, a political essay, a political statement, and so forth. Consequently, classes like student, faculty, or staff, which can be found in public collections like the Web->KB at CMU, or classes like “virtual hosting services” or “universities and research institutes” (Amitay et al., 2003) are not genres, but rather topical or functional categories.

Although difficult and controversial, the concept of genre has great potential for Information Retrieval (IR). For instance, the idea that genre features can also help the automatic classification of XML documents has been recently put forward by Malcom Clark and appears promising (see Clark and Watt, 2007; Clark, 2007). Another application of genre in IR is its integration in a search engine, enabling users to combine topic-based search with genre-based search.

A considerable amount of research has already been done in automatic genre classification. However, no genre benchmark has been developed so far. Additionally, the collections used in genre classification experiments are small and mostly built with subjective criteria. Although the experiment that I am going to present suffers from the same limitations (hopefully future research will address these two problems), it focuses on a specific aspect of genre classification, namely the need of a more flexible genre classification scheme.

Most previous work has considered genres as mutually exclusive categories, disregarding the fact that many documents, and in particular many web pages, cannot be fitted into a single genre. This approach has been taken for the sake of practicality, but proves inadequate when dealing with complex documents, like web pages. In this paper, I suggest that a more flexible genre classification scheme, i.e. a scheme capable of performing zero-, one- or multi-genre assignment, would be more appropriate for web pages. In particular, two factors affect genre identification on the web: (1) the complexity of web pages, and (2) the fluidity and the fast-paced evolution of the web. First, genres on the web are instantiated in web pages that, from a physical, linguistic and textual point of

view, can be considered documents of a new type, much more unpredictable and individualised than documents on paper. Web pages show a visual organization of the space, where several communicative purposes are included at the same time. In brief, in a web page, not all the elements necessarily belong together.

Second, the web is a recent communication medium, invented only in 1989. Being so recent, it is still fluid and unstable. In addition, it is evolving at a fast pace owing to the continuous introduction of new web technologies. In such an environment, the modification of existing genres or the creation of new ones create classification hurdles.

In summary, web pages are noisy documents and the web is a noisy environment. If we aim at devising an automatic classification system capable of facing the open web, and not only a closed-world environment, we should account for these factors. The open web accounts for a situation where the population is unknown, where web pages might be evolving, hybrid, individualised, or not following any genre convention. Given these two factors, my suggestion is to look at genres as *named communication artefacts, linked to a society or community, characterized by conventions, raising expectations, showing hybridism or individualization, and undergoing evolution.*

As I am dealing with genres of written documents, I consider genres and web genres as linguistic entities, where natural language is used to convey information to potential receivers. The integration of written language with other multimedia resources, like images or hyperlink structure is one of my future directions. However, genre research has already started within IR in a semiotic prospective or in terms of hyperlink analysis.

In the model presented in this paper I make a clear-cut distinctions between the concepts of text types and genre. Text types are rhetorical/discourse patterns that indicate the purpose of communication. When the purpose is to narrate, the NARRATION text type is used; when the purpose is to instruct, the INSTRUCTION text type is used; and so on. Normally, a text contains several purposes and consequently includes several text types. Text types tend to be universal, and cut across time, cultures and societies. Genres are, on the other hand, culturally defined and socially acknowledged text categories, like EDITORIALS, TUTORIALS, HOME PAGES or BLOGS. The interaction of these two concepts allows for more flexibility because (i) a classification in terms of text types remains possible even if a web page does not belong to any genre (zero-genre assignment), and because (ii) a web page can be labelled with more than one genre (multi-genre assignment) when some genres share the same text types, as in the case of EDITORIALS and SERMONS, which are both argumentative.

In this model, text types and genres are 'inferred' and not 'learned' from examples. For this reason, I refer to this model as the 'inferential model'. One problem with the application of classical machine learning models is that the negative class, i.e. all the genres that that we do not want to automatically classify, is very difficult to modelize, because we do not know the proportions of the genres on the web. Therefore, 'inferring' genres rather than 'learning' genres could be a way to overcome this problem. However, the limitation of the inferential approach is that it cannot be fully evaluated as genre research is still in its infancy. Here, I present a tentative and partial evaluation, and defer to future work investigations into issues related to genre evaluation.

The paper is organized as follows: Section 2 provides a very brief overview of previous work in genre classification; Section 3 describes the approach; and Section 4 contains conclusions.

## 2. PREVIOUS WORK

In automatic genre classification studies, the focus is not on the definition of what genre is, but on the potential usefulness of the concept of genre as a classifying taxonomy. The basic assumption is that people roughly know what genre is and researchers either use genre categories available in electronic corpora, like the Brown corpus, or call 'genre' any non-topical category that could be combined with topic.

Broadly speaking, authors working in automatic genre classification assume that genres are mutually exclusive discrete classes, each document instantiating a single genre. For this reason, the mainstream approach is based on discrete single-genre supervised classification (e.g. Karlgren and Cutting, 1994; Stamatatos et al., 2000; Dewdney et al., 2001). Although the inadequacy of the equation 'one web page = one genre' has been acknowledged by authors working in automatic genre classification (e.g. cf. Lim et al., 2005), still most methodologies rely on the assignment of a single genre label to the individual document.

It is important to stress, however, that not all experiments rely on discrete single-genre supervised classification. Some studies describe more flexible systems, namely Kessler et al. (1997), Müller-Kögler (2001) and Rehm (2005). Although very intriguing in many respects, these three approaches do not seem adequate to cope with a situation of flux, like the current scenario of the open web. For this reason, I suggest a different approach.

## 3. METHODOLOGY

The model presented in this section is based on four elements: (i) a corpus of web pages representative of the web, (ii) a special kind of feature, the facets, (iii) four text types, and (iv) seven web genres. The model is implemented with English web pages and designed to detect text types and genres within the individual web page.

### 3.1 The Web Corpus

Since the web is in constant flux, it is almost impossible to compile a representative sample of the web as a whole (the multi-lingual web), or only of a single language, like the English web. There are estimates of the number of indexed web pages (in April 2005 Google could search 8,058,044,651 web pages), which is a daily growing number, but we do not know the proportions of the different types of text on the web.

From a statistical point of view, when the composition of a population is unknown, the best solution is to extract a large random sample and draw inferences from that sample. However, deciding the size of this random sample is not a trivial issue. The solution that I suggest for this model is to approximate one of the possible compositions of a random slice of the web statistically supported by reliable standard error measures. Therefore, I built a web corpus has a standard error ranging from 0.000002 to 0.00744. Since the standard error measures of the corpus are very small, I consider the sampling distribution of the web corpus mean as approximately normal, following the Central Limit Theorem. According to Central Limit Theorem, the sampling distribution of the sample mean is approximately normal even when the population is not. This fact is very important because the inferential model is based on z-scores, and z-scores assume a normal distribution of the sample.

The web corpus includes four BBC web genres (20 web pages each) – EDITORIALS, DIY MINI-GUIDES, SHORT BIOGRAPHIES, and FEATURE ARTICLES – and seven novel web genres (200 web pages each) – BLOGS, ESHOPS, FAQs, FRONT PAGES, LISTINGS, PERSONAL HOME PAGES, and SEARCH PAGES. These web genres represent the known part of the web, i.e. about 60% of the sample. The SPIRIT collection (Joho and Sanderson, 2004), which contains random and unclassified web pages, amounts to about 40% (1,000 web pages) and represents the unknown part of the web. The selection of the genres and their proportions are purely arbitrary. Future work includes the testing of the model on a different slice of web genres with different proportions.

### 3.2 The Facets

Broadly speaking, the word 'facet' indicates an 'aspect' of a situation, a concept, and so on. My facets are macro-features, i.e. they contain several micro-features. I used the word 'facet' because each facet represents an 'aspect' of communication. For example, the first person facet includes first person pronouns, singular and plural. The first person facet indicates that the communication context is related to the text producer. A high frequency of first person facets in a text signals an impressionistic or subjective stance of the text producer. While in previous genre classification approaches, pronouns were used individually without any further interpretation, with the first person facet my aim is to interpret, or assess, whether first person pronouns indicate a particular stance in communication, and if this stance is linked to a genre or text type. For instance, a high frequency of first person facet is often used in ARGUMENTATIVE genres, like COMMENTS and OPINIONS that can be found in newspapers and magazines. I created these 100 facets to linguistically represent the four text types that I use in the model. For their creation, I drew inspiration from Werlich (1976), Biber (1988) and Biber et al. (1999).

### 3.3 Text Types Inferred With Odds-Likelihood

In this first implementation of the inferential model (presented from a different perspective in Santini et al. 2006) I included four broad intuitive text types: DESCRIPTIVE\_NARRATIVE, EXPLICATORY\_INFORMATIONAL, ARGUMENTATIVE\_PERSUASIVE and INSTRUCTIONAL. These text types are coarse conglomerates, and can be refined in future. As far as I am aware, an inferential approach has never been applied to automatic genre identification. Inferential approaches have been used extensively, instead, in Artificial Intelligence. In particular, the inferential model adopted here, i.e. the form of Bayes' theorem called odds-likelihood or subjective Bayesian method, was suggested by Duda et al. (1981) to handle uncertainty in PROSPECTOR, a rule-based system for classifying mineral exploration prospects. The main reason for choosing the odds-likelihood form of Bayes' theorem is that the model is very simple, but allows for more complex reasoning. Like the standard Bayesian version, the odds-likelihood method is based on probabilities. Odds and probabilities contain the same information and are interconvertible. But odds are not limited to the range 0-1, like probabilities. In other words, odds is a number (without any limitation) that tells us how much more likely one hypothesis is than the other. The main difference between the regular Bayesian models and the subjective one is that in the latter attributes are NOT considered to be equally important, but are weighted according to their probability value. These weights are confidence measures: Logical Sufficiency (LS) and Logical Necessity (LN). LS is used when the evidence is known to exist, while LN is used when evidence is known NOT to exist.  $LS(e|h)$  expresses how much the prior odds  $O(h)$  in presence of a clear evidence of  $e$ , has to be multiplied in order to get the posterior odds  $O(h|e)$ .  $LN(e|h)$  expresses how much the prior odds  $O(h)$ , in presence of a clear evidence against  $e$ , has to be multiplied in order to get the posterior odds  $O(h|e)$ . In this implementation, LS was set to 1.25 and LN was set to 0.8 on the basis of previous experience and empirical adjustments (see details in the poster and in the slides).

### 3.4 Web Genres inferred with *if-then* Rules

Once the gradations/probabilities have been calculated, they are ranked in descending order. At this point, simple if-then rules combine inferred predominant text types with additional traits – namely genre-specific-word facets, HTML facets and web page length – for determining genres in web pages. The main reason for not applying odds-likelihood here is that the combination of text types with other features is more tentative; inference rules allow for a better understanding of how a conclusion is reached. Rules are drawn up on the basis of previous genre studies or cursory qualitative analyses. For example, rules for PERSONAL HOME PAGES are based on Roberts (1998), and Dillon and Gushrowski (2000). The number of positive rules, i.e. those that confirm the presence of attributes for the genre under assessment, is very limited. More precisely, I used: 4 rules for BLOGS; 7 rules for ESHOPS; 5 rules for FAQs; 8 rules for FRONT PAGES; 5 rules for listings; 4 rules for PERSONAL HOME PAGE; 9 rules for SEARCH PAGE (see details in the poster and in the slides). The number of negative rules, i.e. those that disconfirm the presence of attributes for the genre under assessment, is very low for BLOGS, and slightly higher for other web genres.

### 3.5 Evaluation

As mentioned in the Introduction, the model cannot be fully evaluated because of the current state of genre research. Additionally, this model proposes a classification scheme that envisages the possibility of no genre assignment and multiple genre assignment. In this respect, also appropriate evaluation metric is lacking.

In the next subsections, I will perform a tentative and partial evaluation of the model. More precisely, I will (i) report on the single-genre evaluation, and (ii) test the scalability of model.

#### 3.5.1 Single-Genre Evaluation

In order to assess the performance of the inferential model, I compared its classification accuracy with the accuracy of standard single-label classifiers. More precisely, I compared the outcome of the model on a single genre with the SVM and Naive Bayes classifiers from the Weka workbench (Witten and Frank, 2005). SVM and Naive Bayes classification models were built using the seven novel web genres listed in Section 3.1 – 200 web pages per web genre, amounting to a total of 1,400 web pages – and the frequencies of facets normalized to the page length. The stratified 10-fold-cross-validated accuracy returned by these classifiers for seed 1 is about 89% for SVM, and about 67% for Naive Bayes. The accuracy achieved by the inferential model is about 86% (see Table 1).

Although results in Table 1 are not directly comparable because both Naive Bayes and SVM standard classifiers have been run on 1,400 web pages, while the inferential model has been built on a corpus of 2,480 web pages, they give an idea of the accuracy that can be achieved by the inferential model. The accuracy of 86% returned by the inferential model is a good achievement for a first implementation, and it also shows that this model can stand a certain level of noise, represented by web pages that have not been classified by genre, i.e. the 1,000 web pages of the SPIRIT collection, and the 80 web pages belonging to the BBC web genres.

**TABLE 1.** Comparison of accuracies

Web genres	SVM Classifier (1,400 web pages)	Naive Bayes Classifier (1,400 web p.)	Inferential Model (2,480 web p.)
Blogs	96%	92%	91%
Eshops	88%	76%	83%
FAQs	94.5%	67%	88.5%
Front Pages	100%	98%	97%
Listings	80%	29%	75.5%
Personal Home Pages	79%	27%	77%
Search Pages	85%	82%	88%
TOTAL	about 89%	about 67%	about 86%

Although theoretically unsound, for explanatory purposes I built an ‘unknown’ class. This means that I built an SVM classifier with the 2,480 web pages of the web corpus and labelled as DONTKNOW all the web pages belonging to the BBC collection and the SPIRIT collection, i.e. 1,080 web pages. The stratified 10-fold-cross-validated accuracy (seed 1) on the SVM model built with eight classes is about 76%, i.e. about –13% of the accuracy achieved with a corpus of 1,400 web pages. In this respect, the inferential model appears much more corpus independent, returning an accuracy of about 86% on 2,480 web pages, i.e. about +10% more than the model built with SVM on eight classes. Corpus independence is important when dealing with genre classes, because, as explained above, annotating a web page by genre is not straightforward, and is often debatable. The inferential model tries to make the best of theoretical and empirical findings documented by genre analysts by encoding them in hard-coded rules, rather than blindly relying on the learning from small genre collections assembled with subjective criteria. While supervised models needs large quantities of pre-labelled data in order to be reliable, the inferential model does not require any pre-labelled document. The web corpus described in Section 3.1 contains pre-classified web pages mainly for evaluation purposes. Ideally, the inferential model could be built on a totally unclassified corpus (a pure

random slide of the web), and it would indeed help identify those web pages that either belong to one or more of the genres hard-coded in the rules, or which do not belong to any of them.

### 3.5.2 Scalability

Broadly speaking, scalability refers to the ability of an application to continue to function well when it is re-scaled to a larger size. The scalability of a genre model can have several aspects, such as the enlargement of the corpus or the increase of the number of genres. In the current implementation of the inferential model, I tested the former aspect, simulating the increase in size of the web.

To test the scalability, I merged the web corpus made of 2,480 web pages described in Subsection 3.1 with the KI-04 corpus (Meyer zu Eissen and Stein, 2004), made of 1,205 web pages. The genres contained in the KI-04 corpus are: ARTICLES, DISCUSSIONS, DOWNLOADS, HELPS, LINKLISTS, PORTRAYAL (PRIVATE), PORTRAYAL (NON-PRIV.) and SHOPS. In other words, this experiment well simulates a situation where additional web pages with existing or new genres are introduced in a pre-existing setting (here represented by the web corpus). The enlarged corpus contains 3,685 web pages. This means that the corpus has been enlarged by about 35% and that z-scores and inferential process are based on 3,685 web pages instead of 2,480 web pages. For the rest, the model remains unchanged.

Accuracy results achieved by the inferential model with the re-scaled web corpus on single-label classification are shown in Table 2. The second column in this table shows the accuracies on the initial web corpus, while the third column indicates the accuracies with the re-scaled web corpus. The accuracy achieved with the enlarged web corpus is only about <5% (86% vs. 81%). This decrease in accuracy is statistically significant, but quite small if we consider that the corpus has been increased by about 35%.

**TABLE 2.** Accuracies on the initial and rescaled corpus

Web genres	Initial web corpus	Enlarged web corpus
Blogs	91%	72%
Eshops	83%	78.5%
FAQs	88.5%	84%
Front Pages	97%	96.5%
Listings	75.5%	74%
Personal Home Pages	77%	77.5%
Search Pages	88%	85.5%
TOTAL	about 86%	about 81%

### 3.6 Discussion

The accuracy of the inferential model on a single label is competitive with the accuracy of standard classifiers. It achieves 86% accuracy on seven web genres (1,400 web pages) although the model is build on 2,480 web pages, i.e. 1,400 web pages annotated by seven genres used to evaluate the model + 1,000 without any genre annotation + 80 BBC web pages.

The model appears to be scalable in size. It remains basically stable and robust on the accuracy of the seven web genres. Although both the original corpus and the enlarged corpus are relatively small, these sizes can be considered large if compared with other genre experiments with web pages, like Kennedy and Shepherd (2005), who used 321 web pages or Lim et al. (2005), who included 1,224 web pages. Unfortunately, annotating web pages by genre is extremely controversial and time consuming, and no standard criteria have been set so far. The lack of shared genre annotation criteria makes comparisons difficult and tentative.

The inferential approach is implemented in a computational model that is purposely very simple. This computational simplicity should help better understand the validity of the theoretical stance. However, the possibility of replacing hand-crafted rules with some learning methodology (e.g. Winnow) is under exploration.

One could argue that in a standard classifier text types are implicitly learned from the features and directly mapped into a genre, without the need for any middle layer. Unfortunately, the mapping between surface cues and genres is never unambiguous. A relatively small number of linguistic combinations serve to represent many different types of text. With the inferential model the effort is concentrated on reconstructing of the context of communication.

Finally, it is important to stress that although text types are hard-coded in the model, there is no manual annotation of web pages by genre. Web pages belonging to the seven genres were randomly downloaded from genre-specific portals or archives (Santini 2006). Web pages were parsed using the tagger-parser Connexor (Tapanainen and Järvinen, 1997), and linguistic features and facets were automatically extracted and counted from the parsed outputs, while frequencies of HTML tags were automatically counted from the raw web pages.

## 4. CONCLUSIONS

In this paper I advocated for a more flexible genre classification scheme, i.e. zero-to-multi genre classification. I proposed using a simple inferential approach in order to implement this scheme. Unfortunately, given the current

state of genre research, and especially the lack of genre benchmarks, I cannot fully evaluate the model. Some initiatives have been recently started to overcome this absence that holds back computational approaches to genres, e.g. Colloquium "Towards a reference corpus of web genres".

In this paper I offered a partial and tentative evaluation of the inferential model, which focused on single-genre evaluation and scalability. Findings show that the inferential model (a) has a competitive accuracy on single-genre classification; (b) proves stable, showing a decrease in accuracy of only -5% when the corpus on which it is based upon is rescaled to a larger size (+35%); and finally (c) does not require any manual annotation of web pages by genres (an operation that is expensive and time-consuming), because it is not based on the learning from pre-labelled examples but on the inference from general linguistic insights.

All in all, the inferential model looks promising. The main advantage of the model is that it exploits linguistic insights coming from genre studies that give generality and stability to the approach, and make it more corpus-independent. While standard classifiers act mostly like black boxes, where it is often hard to decode how knowledge is handled by the mathematical algorithm, with the inferential model linguistic insights are coded as facets – i.e. interpreted features – and universal text types. Facets and text types interact in a very straightforward rules. A full evaluation of the model and an adequate evaluation metrics are the object of future research.

## REFERENCES

- Amitay, E., Carmel, D., Darlow, A., Lempel, R., and Soffer A. (2003). The Connectivity Sonar: Detecting Site Functionality by Structural Patterns. *ACM Hypertext 2003*.
- Biber, D. (1988). *Variations across speech and writing*. Cambridge University Press, Cambridge.
- Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Longman, Harlow.
- Clark M. (2007). Structured Text Retrieval by means of Affordances and Genre. FDIA 2007, Glasgow.
- Clark M. and Watt S. (2007). Classifying XML Documents by Using Genre Features. TIR-07 4th International Workshop on Text-based Information Retrieval (DEXA 2007). Regensburg, Germany.
- Dewdney N., Vaness-Dikema C. and Macmillan R. (2001). The form is the Substance, *ACL 2001*.
- Dillon, A. and Gushrowski, B. (2000). *Genres and the Web: is the personal home page the first uniquely digital genre?*, *JASIS*, 51(2).
- Duda, R., Hart, P. and Nilsson, N. (1981). Subjective Methods for Rule-Based Inference System. In Weber B., Nilsson N. (eds.), *Readings in Artificial Intelligence*, Tioga. Palo Alto.
- Joho, H. and Sanderson, M. (2004). The SPIRIT collection: an overview of a large web collection, *SIGIR Forum*, 38(2).
- Karlgren, J. and Cutting, D (1994). Recognizing Text Genre with Simple Metrics Using Discriminant Analysis, *COLING 1994*.
- Kennedy, A. and Shepherd, M. (2005). Automatic Identification of Home Pages on the Web, *HICSS 38*.
- Kessler, B., Nunberg, G. and Shütze, H. (1997). Automatic Detection of Text Genre, 35th *ACL* and 8th *EACL*.
- Lim, C., Lee, K. and Kim, G. (2005) Automatic Genre Detection of Web Documents, in Su K., Tsujii J., Lee J., Kwong O. Y. (eds.) *Natural Language Processing*, Springer, Berlin.
- Meyer zu Eissen, S. and Stein, B. (2004). Genre Classification of Web Pages, in Biundo S., Fruhwirth T., Palm G. (eds.), *Advances in Artificial Intelligence*, Springer, Berlin.
- Rauber, A. and Müller-Kögler, A. (2001). Integrating Automatic Genre Analysis into Digital Libraries. *ACM/IEEE Joint Conference on Digital Libraries*.
- Rehm, G. (2005). Language-Independent Text Parsing of Arbitrary HTML-Documents. *LDV Forum* 20(2).
- Roberts, G. (1998). The Home Page as Genre: A Narrative Approach. *HCSS31*.
- Santini M. (2006). Common Criteria for Genre Classification: Annotation and Granularity". Workshop on Text-based Information Retrieval (TIR-06) (held in conjunction with ECAI 2006). Riva del Garda.
- Santini, M., Power, R. and Evans, R. (2006) Implementing a Characterization of Genre for Automatic Genre Identification of Web Pages, *COLING - ACL 2006*.
- Stamatatos, E., Fakotakis, N., Kokkinakis, G. (2000). Text Genre Detection Using Common Word Frequencies. *COLING 2000*.
- Tapanainen P. and Järvinen T. 1997). A non-projective dependency parser. *5th Conf. on Applied Natural Language Processing*.
- Werlich, E. 1976. *A Text Grammar of English*. Quelle & Meyer, Heidelberg.
- Witten, I., Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers.
- Xu, J., Cao, Y., Li, H., Craswell, N., and Huang, Y. (2007). Searching Documents Based on Relevance and Type, *ECIR 2007*.
- Yeung, P., Büttcher, S., Clarke, C. and Kolla, M. (2007). A Bayesian Approach for Learning Document Type Relevance. *ECIR 2007*.
- Yeung, P., Clarke, C. and Büttcher, S., (2007). Improving Retrieval Accuracy by Weighting Document Types with Clickthrough Data. *SIGIR 2007*.
- Yeung, P., Freund and Clarke, C. (2007). X-Site: A Worksite Search Tool for Software Engineers. *SIGIR 2007*.