Testing a Genre-Enabled Application: A Preliminary Assessment

Marina Santini
HATII (University of Glasgow, UK)
MarinaSantini.MS@gmail.com

Mark Rosso
North Carolina Central University (USA)
mrosso@nccu.edu

In this paper we would like to contribute to the discussion about genre-enabled applications, currently engaging many genre researchers, by presenting a preliminary assessment of a web add-on devised to augment the result list of general-purpose search engines with genre labels. For this assessment, we use a small collection of web pages manually annotated with genre labels by a large number of web users. This resource is made up of two sets of web pages created by two independent researchers for their own user-based genre studies. This comparison allows us to provide a preliminary view on the genre add-on performance and to highlight some open issues in genre research.

Genre-enabled applications, web genre, genre annotation, genre labelling, genre evaluation.

1. INTRODUCTION

Genre is a deeply rooted concept in our civilization. Aristotle's *Poetics* started a long-standing discussion about literary text classification by identifying the underlying conventions that differentiate epic, lyric and drama, and the patterns of form and content that characterize tragedy and comedy. Since then, along the centuries, the interest in the conventions typifying textual production has moved from literary criticism, to modern genre analysis, to library science, to online bookshops (e.g. see the *Browse Genres* link in Amazon[1], Figure 1) and finally to digital genres, and genres on the web, a.k.a. web genres. Regardless this uninterrupted tradition of genre studies and practice, the answer to the core question – *what is genre?* – remains basically open due to the number of dissenting definitions, differing characterizations and multiple uses.

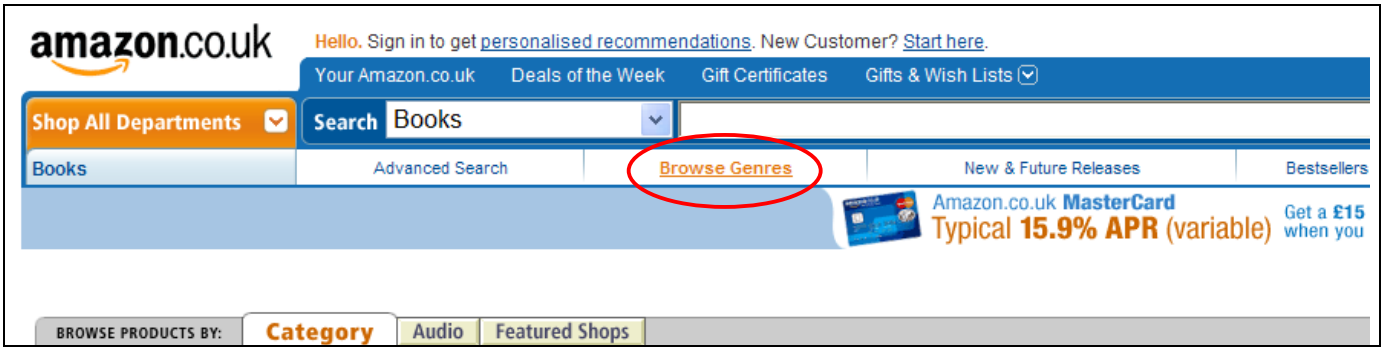

FIGURE 1. The *Browse Genres* link in Amazon UK

In this paper, we focus our attention on the genres that can be found on the web and in other digital environments, such as ESHOPS, HOME PAGES, FAQS, ONLINE FORMS, ONLINE TUTORIALS and LIST OF LINKS. The theoretical and empirical studies of digital and web genres are developed within different but related disciplines, such as genre analysis (e.g. Herring et al., 2005; Askehave and Nielsen, 2005), document management (e.g. Orlikowski and Yates, 1994); text technology (e.g. Rehm, 2008), corpus linguistics (e.g. Sharoff, 2007), information retrieval (e.g. Muresan et al., 2006), social network analysis (e.g. Paolillo et al., 2007), web mining (e.g. Mehler and Wegner, 2008), information extraction (e.g. Gupta et al., 2006), automatic summarization (e.g. Seki, 2005), and authorship attribution (e.g. Karlgren and Ericsson, 2007). All the researchers working with genres in these different areas strongly believe that genre is an important classification principle that could help many real-world applications, since genres could be used as filters, as metadata, for indexing, and in any kind of crawlers, agents, or robots that explore or harvest the web or large document collections. Despite its inherent elusiveness, currently there is a high peak of interest in the concept of genre. Many recent academic initiatives have been set up to foster discussion about this concept[2], because – although controversial and difficult to pin down satisfactorily and unanimously – the concept of genre has been proved to be useful to overcome information overload and increase search relevance and usefulness (Freund, 2008a; Freund, 2008b).

[1] See http://www.amazon.co.uk/b/ref=sv_b_1/203-6819510-2539957?ie=UTF8&node=1025612 (accessed 26 August 2008)

[2] For a list, follow the links Reference materials→Genre-Focussed Academic Events in the WEBGENREWIKI <http://purl.org/net/webgenres>.

Genre can be considered a non-topical descriptor that, together with other non-topical descriptors like style or sentiment, may help formulate or refine the information needs expressed in a query. The specificity of genre with respect of other topical and non-topical descriptors lies in its capacity to convey how information is packaged. For instance the INTERVIEW genre indicates that a document contains a dialogue between (usually) two people, one asking questions, the other providing answers.

Genres are based on more or less tight conventions. The identification of these conventions allows people to reconstruct the context in which texts have been produced, together with their purpose and function. In a word, genre is a contextual factor that can be derived from the documents themselves rather than from external human-computer interactions, like clickthroughs or eye tracking. There would be no problem in retrieving genres if all the documents were annotated with genre labels, or contained genre labels in the title, headings and in the meta-content, or if they could be unambiguously correlated with a limited set of topics. If this was the case, genre labels could be simply treated as "terms", and genres could be unfailingly retrieved by current retrieval models. Unfortunately, the scenario is more complex, especially on the web, where genre colonization and genre contamination seem to be widespread, since the web is the crossroads of many communities. At present, there are still many documents or web pages that belong to a certain genre, and could be relevant and useful for a search need, but they do not explicitly contain the name of the genre, or their topics are not predictably correlated with predefined genres. Although general-purpose search engines do a good job when the genre of a document is mentioned in the document itself or in the meta-content, there are still problems when this does not occur, since genre labels can hardly be derived by synonym expansion. This is why the Automatic Genre Identification (AGI) is not entirely term-based or topic-dependent. For instance, obvious cues that could turn out to be useful in the automatic identification of the INTERVIEW genre are: a high frequencies of questions, second person personal pronouns, first personal pronouns, verbs like "believe", "think", "assume", or expressions like "in my opinion", without neglecting graphical hints, like the visual differentiation between questions and answers through the use of paragraph spacing, and other typographical cues such as the use of bold.

Concrete attempts to implement genre-sensitive retrieval models have been made recently. More specifically, Luanne Freund (2008) has presented X-Site, a search system designed and implemented to "test the practical value of making use of task-genre relationships in real-life work environment" (Freund, 2008b: 114). A demo of X-Site was shown at SIGIR 2007 (Yeung et al., 2007).

While X-Site has been devised for professionals (namely software engineers) who can exploit the concept of genre to rapidly find information that is task-appropriate, situationally-relevant and mission-critical for their job, WEGA (an acronym that stands for **WE**b **G**enre **A**nalysis[3]), has been developed at the Bauhaus University Weimar by Prof Benno Stein's team (Stein et al., 2008) for the web and for common web users. WEGA is an add-on that superimposes genre labels a few seconds after the result list is returned by a general-purpose search engine, namely Mozilla Firefox.

These recent applications show that genre-enabled systems are feasible and that genre classes can help improve productivity in the workplace (in the case of X-Site) and offer additional hints about the nature of the web pages listed in the search results. However, the incapacity of defining genre unambiguously has serious repercussions on genre classification and genre labelling. In practical terms, issues that genre researchers constantly face are the following:

(i) How can we say that a genre is a genre, and not another textual category like topic, domain, or style? Although valuable attempts to define the boundaries between these neighbouring categories were made by Lee (2001) and Stein and Meyer zu Eissen (2006), we still do not have any practical criterion that can help us share a common view on these categories.

(ii) What are the cognitive, semantic or pragmatic criteria that people follow when creating a document of a certain genre, or classifying a document by genre? Scholars and researchers suggest different answers to this question, thus creating a plethora of genre classes, selected and defined following contrasting criteria and geared towards different aims.

In this paper, we would like to emphasize the importance of using existing resources (whenever possible) for comparison and cross-testing, especially in a field like AGI, where there are no established benchmarks and where every decision is left to subjectivity, from the selection of genre taxonomies to the creation of genre collections. Comparison and cross-testing help establish relations or correlations between different views and approaches, thus creating a more fertile ground for future research.

While X-Site was evaluated by a user study carried out in December 2005 and based on 32 software service consultants (Freund, 2008b: 129-157), WEGA has not yet gone through any user evaluation to date, because, although publicly available, it is still under development. Leaving the final evaluation to WEGA's creators, here we propose a transversal and preliminary assessment of the WEGA add-on in order to provide some insights *along the way*, i.e. while WEGA is still in a pre-final stage, hoping that these can contribute to WEGA final version and, more in general, to the discussion about genre.

We propose the re-use of a ***small*** number of web pages (50) annotated with genre labels by a ***large*** number of web users. This resource is made up of two sets of web pages created by two independent researchers for their own user-based genre studies, more specifically 20 web pages were labelled by 135 people, and 30 web pages were validated by 257 people. This collection is unique and the two sets have never been used beyond the studies they were devised for. Here we will use this collection to have an idea of the extent to which the genre labels returned by WEGA match the judgement of these two samples of web users. This comparison will allow us to provide a preliminary view on the genre add-on performance, and to highlight some open issues in genre research.

[3] WEGA is freely downloadable from the WEBGENREWIKI (follow the links Reference Materials→Genre-Enabled Applications).

The paper is organized as follows: Section 2 briefly describes the user studies in which the two sets were created; Section 3 presents a comparative analysis; finally Section 4 draws some conclusions and outlines viable future directions.

2. WEB PAGES LABELLED BY GENRE IN USER STUDIES

Recently, a number of user studies have been carried out in order to understand which genre classes could be useful for web applications, and especially for genre-enabled search engines.
In 2004, Meyer zu Eissen and Stein (2004) carried out a survey through a questionnaire where university students were asked to come up with genres that could meet their information needs. The data were analysed and researchers worked out eight classes that could cover the genres suggested by the users. These eight genre classes have been incorporated in the current implementation of WEGA (see subsection 3.1).
In 2004, Rosso (2008) carried out a series of studies to identify the genre classes that could improve web searches. After the users proposed their own genres, Rosso developed a palette of 18 genres, and validated this palette through a user study with 257 participants classifying 55 web pages.
In 2005, Santini (2008) set up an online study to investigate the level of disagreement in genre labelling. She presented 25 web pages to the users (135 participants) and suggested 21 genre labels, plus two additional labels (*Add a new type* and *I don't know*) to be used by the participants when they were not satisfied with the suggested labels.
The tangible outcome of Rosso's and Santini's studies is represented by two sets of web pages annotated with genre labels by a large number of web users. We wish to use this collection of two sets to explore to what extent the genre classification performed by WEGA corresponds to human genre labelling. Our empirical study is described in the next section.

3. WEGA'S PRELIMINARY ASSESSMENT

The rationale of our preliminary assessment is to investigate WEGA performance in classifying the retrieved documents by genre while it is still under development, since, at this stage, a full evaluation would still be premature. This preliminary assessment will also allow us to highlight some problems that currently affect AGI research.
For this small study, we took all the URLs of the web pages annotated by users in Santini's and Rosso's studies, and specified them in Mozilla Firefox browser with WEGA activated. Figure 2 shows this procedure: the URL was typed in the search box (LHS), the results were labelled by genre by WEGA (see coloured flags next to the heading of the snippets). When it was not possible to retrieve the URL, a simple query, based on the web page headings, was typed in the search box (RHS).

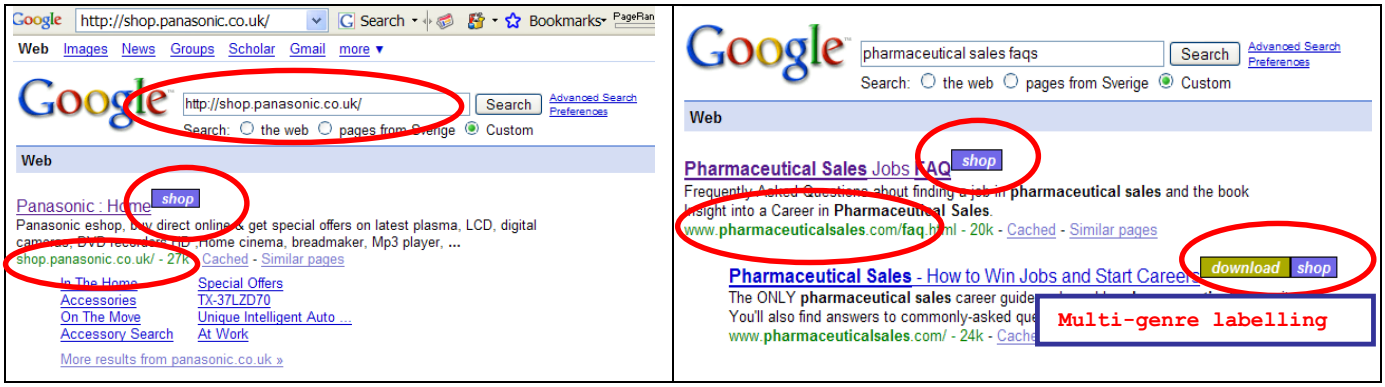

FIGURE 2: Searching by URL (LHS) and searching by keywords (RHS)

3.1 WEGA

WEGA performs the classification of search results using eight genre classes for German and English web pages. WEGA implementation is described in Stein et al. (2008). The genre classes worked out as described in Meyer zu Eissen and Stein (2004), and implemented in the version of WEGA used in this paper are the following:

1. ARTICLES. Documents with long passages of text, such as research articles, reviews, technical reports, or book chapters.
2. DISCUSSIONS. All pages that provide forums, mailing lists or discussion boards.
3. DOWNLOADS. Pages on which freeware, shareware, demo versions of programs etc. can be downloaded.
4. HELPS. All pages that provide assistance, e.g. Q&A or FAQ pages.
5. LINK LISTS. Documents which consist of link lists for the main part.
6. PORTRAYAL (NON-PRIV). Web appearances of companies, universities, and other public institutions. I. e., home or entry or portal pages, descriptions of organization and mission, annual reports, brochures, contact information, etc.
7. PORTRAYAL (PRIV). Private self-portrayals, i.e. typical private homepages with informal content.
8. SHOP. All kinds of pages whose main purpose is product information or sale.

WEGA is provided also with other non-genre classes, such as “non classifiable”, “unsupported language”, “offline”, etc. WEGA follows a multi-labelling scheme, i.e. the same web page can receive several genre labels (see Figure 2).
We assessed WEGA classification following the mapping proposed in Table 1 and Table 2, and used an assessment scheme similar to the one employed in Freund et al. (2006):

1 = Exact match – when WEGA applies only label(s) assigned by users.
2 = Good match – when WEGA applies all the labels assigned by users plus more.
3 = Fair match – when WEGA includes some of the labels assigned by users.
4 = No match, when WEGA applies no labels assigned by users.

It is worth reminding that WEGA is still under development and a couple of version have already been released to date. The assessment reported here is based on the version issued in March 2008.

3.2 SANTINI’s (2008) web pages

The aim of Santini’s study was to explore the need of adopting a multi-labelling genre classification scheme when devising genre-enabled applications, because many experiments in automatic genre classification still focus on only a single label per web page. Her claim was that the single label does not match the view of web users because web pages are often multi-functional and composite. Consequently, when users are forced to select only a single label, they focus on different things, thus generating a large disagreement in genre labelling. Santini’s aim was not to investigate the usefulness of genre for web searches, but to explore the familiarity of web users with labels taken from the web pages themselves. She proposed 21 labels and two escape options (*Add a new type* and *I don’t know*) that the users could use when they were not satisfied with the 21 suggested labels. Labels were suggested in order to reduce the fragmentation that is common with spontaneous users’ terminology (e.g. see the range of variants reported in footnotes 4-14). Her assumption was the following: if users can find an appropriate label in the proposed list, they would gladly use it because this reduces both their cognitive effort and the number of disparate labels (for more details on the sample of participants and the selection of web pages, see Santini, 2008). The labels suggested in Santini’s study are as follows:

1. about page
2. blog (weblog)
3. clog (community log)
4. eshop (online store)
5. email message
6. ezine
7. FAQs
8. home page (academic)
9. home page (corporate)
10. home page (organizational)
11. home page (personal)
12. hotlist
13. howto
14. net advertising (banner)
15. newsletter
16. online frontpage
17. online form
18. online tutorial
19. search page
20. sitemap
21. splash screen
22. *Add a new type*
23. *I don’t know*

In Santini’s study, 135 web users labelled 25 web pages using these labels or their own labels (using the *Add a new type* option), or saying *I don’t know*. Since WEGA palette is coarser-grained than Santini’s labels, we mapped Santini’s labels to WEGA labels following the scheme shown in Table 1. The only direct match was shop-eshop.

TABLE 1: Expected matches (Santini’s study)

Santini’s (2008)	**WEGA**
eshop	shop
about page, corporate home page, organizational home page,	portrait non priv
academic home page, personal home page, personal blog	portrait priv
sitemap, hotlist	linklist
FAQs, tutorials, howtos	help
clog, email, newsletter, other blogs	discussion
---	download
---	article
online form	*non classifiable*
ezine	*non classifiable*
net advertising	*non classifiable*
newspaper online frontpage	*non classifiable*
search page	*non classifiable*
splash screen	*non classifiable*
Add a new type	*(as appropriate)*
I don’t know	*non classifiable*

20 of Santini’s 25 original pages could be retrieved and classified by the WEGA add-on The list of the labels assigned by the 135 users and the labels assigned by WEGA is reported in Table 3 (see Appendix). Comparison shows that out of the 20 web pages that could be reached by Mozilla Firefox, there were: **Exact Matches=0; Good Matches=1; Fair Matches=8; No Match=11**.

3.3 ROSSO'S (2008) web pages

The genre palette developed by Rosso was the result of three user studies. The first study asked experimental participants to group a set of web pages by genre and assign names and definitions to each genre. In the second study, another set of participants labelled the same set of web pages but their choices were mostly constrained to the 48 labels collected during the first study. Participants were allowed to suggest new labels if none of the other labels were deemed adequate. Rosso used the results of the two studies to create a palette of 18 genres and definitions, derived mostly from user-terminology and preferences. The 18 labels are shown below. A third study with new pages and participants "validated" the palette, achieving over an average of over 70% agreement by the 257 participants for the set of 55 pages (for more details on the sample of participants and the selection of web pages, see Rosso, 2008). The labels suggested in Rosso's study are as follows:

1. article
2. course description
3. course list
4. diary, weblog or blog
5. FAQ/help
6. form
7. forum/interactive discussion
8. index/table of contents/links
9. job listing
10. other instructional materials
11. personal website
12. picture/photo
13. poetry
14. product for sale/shopping
15. search start
16. speech
17. welcome/homepage
18. NONE OF THE ABOVE

Table 2 shows how we expected the web page labels assigned by Rosso's study participants to match up against the labels assigned by the WEGA classifier.

TABLE 2: Expected matches (Rosso's study)

Rosso (2008)	**WEGA**
article	article
FAQ/help/ course description	help
forum/interactive discussion archive	discussion
index/table of contents/links/course list	linklist
personal website/diary, weblog or blog	portrayal priv
product for sale/shopping	shop
welcome/homepage	portrayal non-priv
---	download
NONE of the ABOVE	*non classifiable*
speech	*non classifiable*
form	*non classifiable*
search start	*non classifiable*
job listing	*non classifiable*
other instructional materials	*non classifiable*

Only 30 of Rosso's 55 original pages were able to be classified by the WEGA add-on. This was mostly due to pages that no longer exist, or pages that still exist but are no longer in Google's index. We used the same match criteria as with Santini's pages, and the results are reported in Table 4 (see Appendix). Out of the 30 web pages that could be reached by Mozilla Firefox, there were: **Exact Matches=2; Good Matches=0; Fair Matches=11; No Match=17**.

3.5 DISCUSSION

Although the user studies were performed between 2004 and 2005 and WEGA web pages were retrieved in March 2008, these web pages had often the same layout and similar content. As already noted by Boese and Howe (2005), "pages in some genres change rarely if at all and can be used in present-day research experiments without requiring an updated version".

Looking at the matches with the two sets of pages combined (i.e. 50 web pages all in all), there were: **Exact Matches= 4%; Good Matches=2%; Fair Matches=38%; No Match=56%** (see Figure 2 for a charted overview of raw counts).

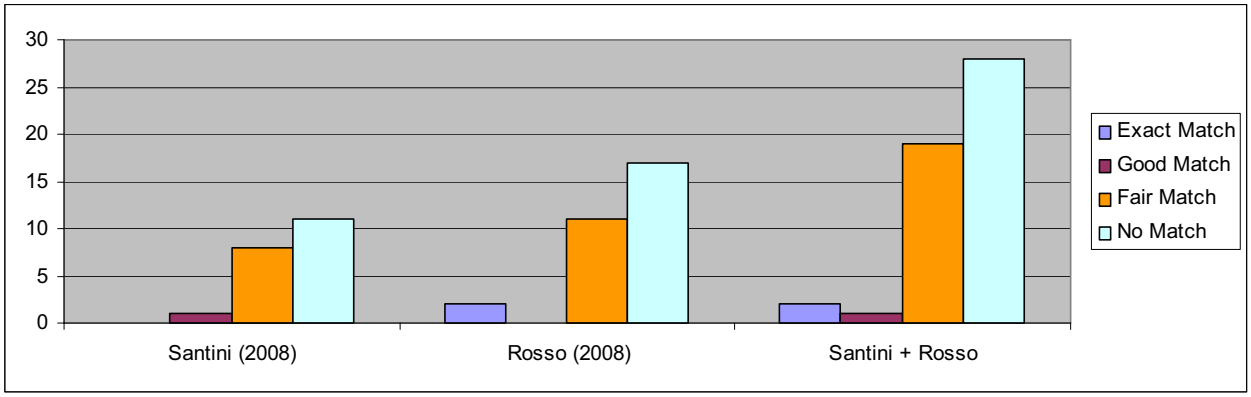

FIGURE 2: Overview of the matches (raw counts)

The sum of Exact Matches, Good Matches and Fair Matches is about 45%, which is a promising achievement since genre classification in an open environment like the web is overly difficult for a number of reasons. One reason concerns the distribution of genres on the web. As we do not know how genres are distributed on the web, it is very difficult to approximate web genre population in any computational and statistical model. Another reason is related to feature representativeness, because the relationship between automatically extractable features and web genres is still under exploration. To date, a number of experiments have been carried out to investigate the efficiency and effectiveness of a range of genre features, but always on very small genre collections, containing only a restricted number of genres and a limited number of documents (e.g. cf. Dong et al., 2008; Kim and Ross, 2008; Kanaris and Stamatatos, 2007). In particular, WEGA has been trained and tested on two small but widely used genre corpora, the KI-04 corpus (Meyer zu Eissen and Stein, 2004) and the 7-web-genre collection (Santini, 2007), both containing fewer than 2000 web pages.

An important factor that causes bias in our comparison is the misalignment of Santini's and Rosso's palettes with WEGA palette. Both Santini's and Rosso's palettes contain genres that had no obvious correspondence with WEGA palette. Clearly, WEGA palette is generally at a higher level of abstraction than the other two palettes. In this respect, it would be very helpful for future research to start creating a network of relationships between genres. For example, a new resource could be designed and implemented similar to the hierarchical framework adopted in Wordnet, or following an ontology-like structure. This would permit not only more straightforward comparisons among different genre palettes and different collections, but also a deeper understanding of the cognitive criteria or constraints underlying genre classes.

As far as WEGA is concerned, it seems to be a good idea, for search or browsing purposes, to have more coarse-grained genre classes like HELP including FAQS, TUTORIALS and HOWTOS, or a LINKLIST genre including different kind of listing genres, like SITEMAPS and HOTLISTS. Therefore, a certain level of abstraction in the genre classes presented to web users is very welcomed. However, it is not clear how to distribute more fine-grained genres like BLOGS (usually divided into several subclasses, like PERSONAL BLOGS or NEWS BLOGS, characterised by different purposes and audience) within more general classes. Maybe, BLOGS should have their own place in any genre palette. While BLOGS have a very strong genre identity, classes like PORTRAYAL PRIVATE and PORTRAYAL NON-PRIVATE seem to be rather opaque. For this reason, they have been replaced by other labels, namely PERSONAL HOMEPAGE and NON-PERSONAL HOME PAGE respectively, in the version of WEGA released in April 2008.

Given the small size of the user-labelled web page collection, we cannot draw any final conclusions. As mentioned earlier, a full evaluation (maybe task-based) will probably be carried out by WEGA's creators when this application reaches its completion. With this preliminary assessment, we would like to emphasise that there is still a lot to know about the tradeoffs involved in genre labelling, the variations across different genres, and the cognitive implications in the use of genres (see also Freund et al., 2006).

4. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we presented a preliminary assessment of WEGA, a genre add-on. Since no genre test collections or genre benchmarks are currently available, we assessed WEGA performance using 50 web pages annotated with genre labels by the participants to two genre studies. On this small collection, WEGA performance in March 2008 was below 50%. Apparently, there is a wide gap between laboratory tests of genre classifiers and the performance in real-world conditions. Recorded performance measurements for automatic genre classifiers can be higher than 90% (cf. Santini, 2007; Kanaris and Stamatatos, 2007; Dong et al., 2008). If web retrieval by genre is ever to become a widespread reality, this gap must be filled by future research.

Obviously, a reliable and convincing evaluation cannot be based only on a collection of 50 web pages. We consider the assessment described in this paper only a first step towards a larger evaluation. Nonetheless, it is important to stress that this collection, although tiny, is unique because it is labelled by a large number of web users. It is also worth noting that existing genre collections are all annotated with disparate genre criteria, and the annotation is commonly decided on the agreement of a very small number of annotators (at most 4), or decided by the individual researchers themselves. This high degree of subjectivity leads also to diversified genre palettes, and their mapping introduces a bias in subsequent re-use. We suggest that the creation of a genre resource capable of spelling out and encoding the inter-relationships among genres would be useful and would permit a more effective re-utilization of existing resources.

From a methodological viewpoint, WEGA is a web browser add-on. Therefore the retrieval of relevant documents is decided by the underlying search engine on the basis of topical keywords, and WEGA applies genre labels on the search results. A viable and complementary line of genre research would be the integration of topic and genre as combined search criteria, as in Vidulin et al., 2007.

REFERENCES

Askehave, I. and Nielsen, A. E. (2005). What are the Characteristics of Digital Genres? – Genre Theory from a Multi-modal Perspective. *Proceedings of Hawaii International Conference on System Sciences (HICSS-2005)*.

Belkin, N., Chaleva, I., Cole, M., Li, Y.-L., Liu, L., Liu, Y.-H., Muresan, G., Smith, C., Sun, Y., Yuan, X.-J., Zhang, X.-M. (2005). Rutgers' HARD Track Experiences at TREC 2004. *Proceedings of TREC-2004*.

Boese, E. and Howe, A. (2005). Effects of Web Document Evolution on Genre Classification. *Proceedings of the ACM 14th Conference on Information and Knowledge Management* (CIKM 2005).

Dong, L., Watters, C., Duffy, J. and Shepherd, M. (2008). An Examination of Genre Attributes for Web Page Classification. *Proceedings of Hawaii International Conference on System Sciences (HICSS-2008)*.
Freund, L. (2008a). Situating relevance through task-genre relationships. *Bulletin for the American Society for Information Science and Technoloy*, 34 (5), 23-26.
Freund, L. (2008b). *Exploiting task-document relations in support of information retrieval in the workplace*, doctoral dissertation, Faculty of Information Studies, University of Toronto, Canada <http://faculty.arts.ubc.ca/lfreund/Publications/Freund_Luanne_S_200811_PhD_thesis.pdf>.
Freund, L., Clarke, C.L.A. & Toms, E.G. (2006). Genre classification for IR in the workplace. *Proceedings of Information Interaction in Context* (IIiX 2006).
Gupta, S., Becker, H., Kaiser, G., and Stolfo, S. (2006). Verifying genre-based clustering approach to content extraction. *Proceedings of WWW '06*.
Herring, S., Scheidt, L., Bonus, S. and Wright, E. (2005). Weblogs as a bridging genre. *Information, Technology & People*, 18 (2).
Kanaris, I. and Stamatatos E. (2007). Webpage Genre Identification Using Variable-length Character n-grams. *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence*.
Karlgren, J. and Eriksson, G. (2007). Authors, Genre, and Linguistic Convention. *Proceeding of SIGIR Workshop on Plagiarism Analysis*, 30th International ACM SIGIR Conference, Amsterdam.
Kim Y. and Ross S. (2008). Examining Variations of Prominent Features in Genre Classification. *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS-2008)*.
Lee, D. (2001). Genres, Registers, Text types, Domains, and Styles: Clarifying the concepts and navigating a path through the BNC Jungle. *Language Learning & Technology*, 5(3).
Mehler, A. and Wegner A. (2008). Unifying Content and Structure Learning: A Model of Semi-Supervised Hypertext Zoning. *Abstract Prodeedings of the Processing Text-technological Resources Conference*, Bielefeld University, Germany <http://coli.lili.uni-bielefeld.de/Texttechnologie/Forschergruppe/PTTR/abstracts/Abstract-Mehler-Wegner.pdf>.
Meyer zu Eissen, S. and Stein, B. (2004). Genre Classification of Web Pages: User Study and Feasibility Analysis. In Biundo S., Fruhwirth T. and Palm G. (eds.). *KI 2004: Advances in Artificial Intelligence*, Springer.
Muresan, G., Smith, C., Cole, M., Liu, L. and Belkin, N. (2006). Detecting Document Genre for Personalization in Information Retrieval. *Proceedings of Hawaii International Conference on System Sciences (HICSS-2006)*.
Orlikowski, W. and Yates, J. (1994). Genre Repertoire: The Structuring of Communicative Practices in Organizations. *Administrative Science Quarterly*, 39 (4).
Paolillo, P., Warren, J. and Kunz, B. (2007). Social Network and Genre Emergence in Amateur Flash Multimedia. *Proceedings of Hawaii International Conference on System Sciences (HICSS-2007)*.
Rehm, G. (2008). Hypertext Types and Markup Languages. In Dieter Metzing, Andreas Witt (eds.) *Linguistic Modelling of Information and Markup Languages*, Springer.
Rosso, M. (2008). User-based Identification of Web Genres. *JASIST*, 59(7).
Santini, M. (2007). *Automatic Identification of Genre in Web Pages*. PhD Thesis, University of Brighton, Brighton.
Santini, M. (2008). Zero, single, or multi? genre of web pages through the users' perspective. *IPM*, 44(2).
Seki Y. (2005). Automatic Summarization Focusing on Document Genre and Text Structure. Doctoral Abstract. National Institute of Informatics (NII). Tokyo.
Sharoff, S. (2007) Classifying Web corpora into domain and genre using automatic feature identification. *Proceedings of Web as Corpus Workshop*, Louvain-la-Neuve.
Stein, B. and Meyer zu Eissen, S. (2006) Distinguishing Topic from Genre. *Proc.of I-KNOW 06*.
Stein, B., Meyer zu Eissen, S. and Lipka, N. (2008). Web genre analysis: Use cases, retrieval models, and implementation issues (in preparation).
Vidulin, V., Luštrek, M., Gams, M. (2007). Using genres to improve search engines. *Proceedings of the workshop Towards Genre-enable Search Engines: The Impact of Natural Language*. RANLP-2007.
Yeung, P., Freund, L. and Clarke, C. (2007) X-Site: a workplace search tool for software engineers. System demo presented at the 30th International ACM SIGIR Conference, Amsterdam.

APPENDIX

TABLE 3: Classification of web pages from Santini's (2008) study

URLs (web page name between brackets, see Appendix in Santini, 2008)	Web Users' Genre Labelling	WEGA Classification	Assessment (1-4)
http://shop.panasonic.co.uk/ (webpage_01)	**ESHOP=88.15%** **NETAD=5.19%**	***SHOP***	**2**
http://www.satansbarber.co.uk/ (webpage_02)	**PERS. HOMEPAGE=88.89%** **BLOG=7.41%**	***PORTRAIT PRIV,*** ***PORTRAIT NON PRIV***	**3**

http://torvald.aksis.uib.no/corpora/2004-3/0239.htm (webpage_03) **New**: http://gandalf.aksis.uib.no/corpora/2004-3/0239.html	**EMAIL=48.89%** ***ADD LABEL=25.19%***[4] **ABOUT PAGE=14.81%** **BLOG=4.44%**	***PORTRAIT PRIV***	**3**
http://www.nytimes.com/ (webpage_04)	**FRONT PAGE=40.74%** ***ADD LABEL=22.96%***[5] **NEWSLETTER=11.11%** **EZINE=8.15%** **ORG. HOMEPAGE=6.67%** **CORP. HOMEPAGE=5.93%**	***SHOP***	**4**
http://www.dogpile.com/ (webpage_05)	**SEARCH PAGE=83.7%** **ADD LABEL=4.44%**[6]	***DOWNLOAD***	**4**
http://www.thebritishmuseum.ac.uk/sitemap/sitemap.html (webpage_06) **New**: http://www.britishmuseum.org/about_this_site/site_map.aspx	**SEARCH PAGE=47.41%** **SITEMAP=34.07%** **HOTLIST=7.41%**	***SHOP*** ***LINKLIST***	**3**
http://journals.aol.com/brucer5150/AGimpsLife/ (webpage_07)	**BLOG=66.67%** **ABOUT PAGE= 18.52%** **ADD LABEL=7.41%**[7] **DON'T KNOW=4.44%**	***DISCUSSION*** ***PORTRAIT NON PRIV***	**3**
http://www.cs.brown.edu/people/ec/ (webpage_08)	**ACAD. HOMEPAGE=58.52%** **PERS. HOMEPAGE=23.70%** **ABOUT PAGE=8.15%** **ADD LABEL=4.44%**[8]	***DOWNLOAD*** ***PORTRAIT PRIV***	**3**
http://www.infogistics.com/about.html (webpage_10)	**CORP. HOMEPAGE=69.63%** **ABOUT PAGE=23.70%**	***DOWNLOAD***	**4**
http://www.intel.com/index.htm?iid=Homepage+Header_UShome& (webpage_11)	**CORP. HOMEPAGE=88.15%**	***PORTRAIT PRIV***	**4**
http://www.pharmaceuticalsaleshelp.com/faq.php (webpage_12)	**FAQs=83.7%**	***SHOP***	**4**
http://www.splendidezine.com/ (webpage_13)	**EZINE=60%** **NEWSLETTER=11.85%** **FRONT PAGE=11.11%** **ORG. HOMEPAGE=5.19%**	***DOWNLOAD*** ***LINKLIST***	**4**
http://kycares.ky.gov (webpage_14)	**ORG. HOMEPAGE=51.1%** **CORP. HOMEPAGE=9.63%** **ABOUT PAGE=8.89%** **FRONT PAGE=5.93%** **SEARCH PAGE=5.19%**	***PORTRAIT PRIV***	**4**
http://www.fi.edu/tfi/hotlists/insects.html (webpage_15) **New**: http://www.fi.edu/learn/hotlists/insects.php	**ADD LABEL=23.70%**[9] **HOTLIST=21.48%** **SITEMAP=17.04%** **TUTORIAL=8.15%** **ACAD. HOMEPAGE=5.93%** **SEARCH PAGE=5.19%** **DON'T KNOW=5.19%**	***PORTRAIT NON PRIV***	**3**

[4] The added labels for this page were: *bulletin board, discussion group, discussion list, discussion page, email (within a web-based mailing-list archive), email archive, email discussion list message, email newsgroup archive, forum, forum posting, list serve web posting, listserv, listserv message, listserve message, mailing list, mailing list archive, message board, message board entry, message from newsgroup, newsgroup, online forum, online forum/community interactive page, web forum/discussion list*.

[5] The added labels for this page were: *e-newspaper, electronic newspaper, entry point of a regular newspaper, home page (newspaper), home page (publication), home page newspaper, info webportal, news, news site, newspaper, newspaper front page, on-line newspaper, online magazine, online news source, online news website, online newspaper, online newspaper, periodical front page*.

[6] The added labels for this page were: *home page (search engine), meta-searchengine, portal, search engine, search engine front page*.

[7] The added labels for this page were: *bulletin board, chat page, diary, discussion forum (chat room), entries within a blog, forum, message board, online forum, online journal, someones bull shit* (sic).

[8] The added labels for this page were: *academic's personal home page, contact, online cv, organizational sub-link, personal page on academic institution website*.

[9] The added labels for this page were: *academic document, catalog, classification page, contents page, database listing, encyclopaedia* (sic), *expert information perhaps, index, index of links, index page, information page, itemization page, knowledge directory entry, link list, links page, menu page, navigation page, online encyclopedia, online reference, online table of contents, online textbook, primary navigation tool, reference, reference page, select from list, table of contents, topic indices*.

	ABOUT PAGE=4.44%		
http://faculty.plattsburgh.edu/nancy.allen/aok.htm (webpage_18)	**ORG. HOMEPAGE=38.52%** **ABOUT PAGE=16.30%** **NEWSLETTER= 10.37%** **DON'T KNOW=10.37%** **HOTLIST=5.93%** **ADD LABEL=5.93%**[10]	***DOWNLOAD***	**4**
http://wt.xpilot.org/publications/linux/howtos/cd-writing/html/ (webpage_20) **New**: http://tldp.org/HOWTO/CD-Writing-HOWTO-4.html	**HOWTO=54.07%** **TUTORIAL=22.22%** **FAQs=19.26%**	***DOWNLOAD***	**4**
http://www.citidex.net/896.htm (webpage_21) **New**: http://www.citidex.com/	**SEARCH PAGE=57%** **ONLINE FORM=9.63** **ADD LABEL=6.67%**[11] **DON'T KNOW= 4.44%** **ESHOP=4.44%**	***SHOP***	**3**
http://www.intap.net/~drw/cpp / (webpage_22) **New**:http://www.intap.net/~drw/cpp/cpp03_02.htm	**TUTORIAL=65.19%** **HOWTO=21.48%** **ADD LABEL=5.93%**[12]	***ARTICLE***	**4**
http://www.oceanoptics.com/products/ach.asp (webpage_23) **New**:http://www.oceanoptics.com/Products/74ach.asp	**ADD LABEL=26.67**[13] **ABOUT PAGE=20.74%** **ESHOP=14.81%** **DON'T KNOW=10.37%** **TUTORIAL=6.67%**	***SHOP*** ***ARTICLE***	**3**
http://www.lotekk.net/index.php?page=moz&sub=splash (webpage_24)	**SPLASH SCREEN=45.19%** **DON'T KNOW=17.78%** **ADD LABEL=15.56%**[14] **NET AD=4.44%**	***SHOP*** ***HELP***	**4**

TABLE 4: Classification of web pages from Rosso's (2008) study

URLs	Web Users' Genre Labelling	WEGA Classification	Assessment (1-4)
http://themis.law.ualr.edu:81/	**INDEX=50%** **SEARCH START=41%**	***SHOP*** ***DISCUSSION***	**4**
http://www.matsci.ucdavis.edu/	**WELC./HOMEPAGE=90%**	***DOWNLOAD***	**4**
http://www.hnet.uci.edu/mposter/syllabi/readings/viruses.html	**ARTICLE=89%**	***ARTICLE***	**1**
http://otto.cmr.fsu.edu/~kelley_r/justtonnetz.htm **New**: http://www.robertkelleyphd.com/justtonnetz.htm	**ARTICLE=50%** **PERS. WEBSITE=22%**	***PORTRAIT PRIV*** ***ARTICLE***	**1**
http://www.cs.wpi.edu/Research/aidg/CS540/aid.html	**COURSE DESCR.=93%**	***PORTRAIT PRIV*** ***ARTICLE***	**4**
http://blogs.law.harvard.edu/ethan/	**DIARY, BLOG=95%**	***ARTICLE***	**4**

[10] The added labels for this page were: *calendar page, content page, events page, listings page, mixed, organisational page (not home), results page, search result page*.

[11] The added labels for this page were: *entry within the yellow pages, online directory, online service, results page, search result, search results page, specific search page (regional), yellow pages*.

[12] The added labels for this page were: *detailed information, instruction manual, manual, page/section of software documentation, reference, reference page, technical resource/ nerdy/geeky, technical test*.

[13] The added labels for this page were: *(technical) product information page, content page, content page (corporate), corporate web site content, information, information page, normal webpage, online product information, product catalogue, product documentation, product info, product information, product information page - technical specification of a product, product manual, product specification page, product specification sheet, product/ service info (details) page, specification sheet - tech info page, sub page of an online store3, tech spec, tech specs, technical description, technical document, technical documentation, technical documentation/product description, technical information about a product, technical information page, technical instructions, technical product description, technical spec, technical specification, technical specification/product description, technical specifications, technical specifications document, technical specs, webpage*.

[14] The added labels for this page were: *browser, browser loader, dialog box, flash, flash page, flash website, game, game site, loading message, loading page, loading prompt, org loading page, placeholder for a flash app, software download, splash screen, tick-tock page (see comment), wait page, web application*.

New: http://www.ethanzuckerman.com/blog/			
http://helpdesk.wisc.edu/page.php?id=2836	**FAQ/HELP=76%**	***DOWNLOAD***	**4**
http://ls.berkeley.edu/mail/webnet/2004/0046.html	**FORUM/INT. DISC.=85%**	***HELP PORTRAIT PRIV***	**4**
http://www.cropsci.uiuc.edu/faculty/long/	**PERS. WEBSITE=63% NONE of the ABOVE=19%**	***DOWNLOAD ARTICLE***	**4**
http://www.english.uiuc.edu/maps/poets/s_z/cdwright/burt.html	**ARTICLE=72% POETRY=49%**	***ARTICLE LINKLIST***	**3**
http://books.nap.edu/books/0309072786/html/20.html	**SHOPPING=33% ARTICLE=31%**	***ARTICLE***	**3**
http://www.unt.edu/untpress/titles/davisrod.htm	**SHOPPING=83% ARTICLE=10%**	***ARTICLE PORTRAIT PRIV***	**3**
http://pharmacy.ucsf.edu/alumni/address/4/ **New**: http://pharmacy.ucsf.edu/facultyandstaff/address/4/	**ARTICLE=28% NONE of the ABOVE=25% SPEECH=18% SEARCH START=16%**	***PORTRAIT PRIV***	**4**
http://asucd.ucdavis.edu/organizations/other/mar/ **New**: http://www.ucdmaar.org/	**WELC./HOMEPAGE=75%**	***SHOP***	**4**
http://www.uphs.upenn.edu/pahedu/library/	**WELC./HOMEPAGE=60%**	***PORTRAIT PRIV PORTRAIT NON PRIV***	**3**
http://www.biosci.ohio-state.edu/	**WELC./HOMEPAGE=93%**	***PORTRAIT PRIV PORTRAIT NON PRIV***	**3**
http://www.math.ucsd.edu/~williams/bandwidth/kwfluid.html	**ARTICLE=63% NONE of the ABOVE=20%**	***PORTRAIT NON PRIV***	**4**
http://www.med.unc.edu/alcohol/prevention/quiz/quiz.html	**OTHER INSTRUCT.=48% FAQ/HELP=14% NONE of the ABOVE=14% FORM=10% ARTICLE=9%**	***LINKLIST ARTICLE***	**3**
http://iitc.tamu.edu/1998and2000/lessons/lesson20.html	**OTHER INSTRUCT.= 78% COURSE DESCR.=15%**	***PORTRAIT PRIV LINKLIST***	**4**
http://home.case.edu/~mss42/2003/11/matchcom-connecting-people-until-they.html **New**: http://liquidschwartz.wordpress.com/2003/11/09/matchcom-connecting-people-until-they-die/	**DIARY, BLOG=86%**	***PORTRAIT NON PRIV SHOP***	**4**
http://undergrad-catalog.buffalo.edu/coursedescriptions/index.php?frm_abbr=HIS&frm_num=161	**COURSE LIST=82% COURSE DESCR.= 12%**	***HELP ARTICLE***	**3**
http://www.sunysb.edu/philosophy/new/courses/curr_grad_courses.html	**COURSE LIST=67% COURSE DESCR.=30%**	***ARTICLE LINKLIST***	**3**
http://www.su.edu/conservatory/scon/Courses/MUCH.IDC	**COURSE LIST=80%**	***LINKLIST PORTRAIT PRIV***	**3**
http://mason.gmu.edu/~lrockwoo/Sample%20Exam%204.htm	**OTHER INSTRUCT.= 81% FORM=12%**	***ARTICLE***	**4**
http://www.kennesaw.edu/communication/outofto wninternships.shtml	**JOB LISTING=91%**	***PORTRAIT PRIV***	**4**
http://www.brook.edu/comm/events/20040310iraq.htm	**ARTICLE=51% SPEECH=30%**	***ARTICLE PORTRAIT PRIV***	**3**
http://web.princeton.edu/sites/chapel/112303.htm **New**: http://web.princeton.edu/sites/chapel/Sermon%20Files/2003_sermons/112303.htm	**SPEECH=67% ARTICLE=20%**	***ARTICLE PORTRAIT NON PRIV***	**3**
http://www.sfsu.edu/~jtolson/vgarden/1996/garden96.htm	**PERS. WEBSITE=66%**	***LINKLIST POTRAIT NON PRIV***	**4**
http://all.successcenter.ohio-state.edu/epl259-su2001/module-03-forms/self-survey-3-1.asp New: http://all.successcenter.ohio-state.edu/tmtnm/tmtnm.htm	**FORM=62% OTHER INSTRUCT.=19% NONE of the ABOVE=16%**	***ARTICLE DISCUSSION***	**4**
http://web.cornell.edu/redesign/blog/index.php?p=32	**BLOG=75%**	***ARTICLE***	**4**