Genres In Formation?
An Exploratory Study of Web Pages using Cluster Analysis

Marina Santini
ITRI
University of Brighton, UK
Marina.Santini@itri.brighton.ac.uk

Abstract

The Web is a new, large and heterogeneous community where the interaction among the users and the possibility offered by technology may modify existing genres or create new ones. In fact, most genres being borrowed from the paper world have undergone adjustments when moving on to the Web (for instance, online newspapers and online manuals). Also, there is a family of genres, which have been created specifically for the Web, e.g. home pages, splash screens, newsletters, hotlists. Besides these, are there other emerging genres on the Web for which a genre label has not been coined yet? Is it possible to capture genres in formation in an automated way? An experiment using cluster analysis has been set up to provide initial answers to these questions. Results show that the main clusters have a shape which is quite well-defined and show a number of regularities. Interestingly, Web pages appear to have been clustered according to their rhetorical/discoursal types (informational, instructional, argumentative, etc.), rather than genre classes (e.g. sermons and editorials, both argumentative, belong to the same cluster). The perception of rhetorical/discoursal types in Web pages has been confirmed by a small-scale Web user study.

1 Introduction

Genres are cultural products, linked to a culture, a society, a community. The Web is a new, large and heterogeneous community where the interaction among the members (the *Internauts*) and the possibility offered by technology might modify existing genres or create new ones, which better satisfy the communication needs of Web users.

The Web has triggered adjustments in many fields. The impact of the Web on the genre system or genre repertoire is an interesting issue, which is worth investigating.

The influence of a new communication medium on genre evolution and creation has been historically proved. For example, the introduction of printing in the XV century, which entailed a passage from hand-written manuscripts to printed books, radically enlarged and transformed the potential for written genres. The steps involved in genre evolution have already been detailed. At first, genres on the new medium are faithful reproductions of existing genres. Then variants are created. These variants become very different from the original genres as time passes. At the same time innovative genres, unrelated to previous ones, might also be created, in order to meet the needs and requirements specific to the new medium. These steps have been traced also on the Web. In their studies, Crowston and Williams (1997) and Shepherd and Watters (1998) identified three main genre categories: 1) reproduced/replicated; 2) adapted/variant; 3) novel/ spontaneous. Many genres on the Web are reproduced or replicated genres, i.e. they are traditional paper genres that have been transplanted into an electronic form. *Academic papers* are one of these. They conform to the rules of academic writing in general (format, style, etc.), and to the rules of a specific domain in particular (for example, academic papers for humanities have different conventions from academic papers for science and technology). Academic papers represent a stable genre. It seems that the electronic transplantation has not influenced any of its conventions. However, most genres coming from the paper world have undergone some adjustments when moving on to the Web. They have been adapted to the functionality of the new medium and variants have been created. Just to mention the most obvious ones, *online newspapers* (with their medley of internal and external links) and *online manuals* (which often have a hypertextual structure) show the adaptation of paper genres to the functionalities provided by the Web (cf. Crowston and Williams 1999, Shepherd and Watters 1999). Only recently, novel and spontaneous Web genres have become fully acknowledged and labels have been invented for them: *home pages* (personal, academic, organizational, etc.), *FAQs*, *newsletters*, *splash screens*, *emails, weblogs*, etc. A set of specific conventions start being built around them. The term “cybergenre” has been coined to label the combination of the concept of genre with the extensive use of the Internet (Shepherd and Watters 1998). These novel genres show stable features and raise expectations. For instance, in a *personal home page* we expect to find a narration of the ‘self’, i.e. a description of personal life, possibly pictures, interests, hobbies, and so on (Dillon and Gushrowski 2000).

Besides these, are there other genres emerging from the Web for which a genre label has not been coined yet? Is there an automated way of detecting genres that are slowly taking shape on the Web? Is it at all possible to capture genres in formation? The experiment presented here provides preliminary answers to these questions.

This paper is organized as follows: Section 2 briefly goes through previous explorations of Web genres; the aim is to identify any references to “unclassifiable” Web pages; Section 3 focuses on the difficulty of applying genre labels to many Web pages; Section 4 illustrates the experiment and the related user study; Section 5 draws some conclusions and suggests directions for future work.

2 Background

So far, nobody has directly addressed the issue of detecting genres in formation, i.e. genres that are not completely formed and acknowledged, but that exist nonetheless, though in an embryonic form. Some useful hints can be derived indirectly from the few surveys carried out so far on Web pages.

Crowston and Williams (1997), whose aim was to document the range of genres used on the Web, sampled and classified 837 randomly selected Web pages (initially they selected 1000 URLs from the URL database created by the developers of Altavista, but not all the URLs were active). For the classification, they started from a list of genres drawn mainly from Oxford English Dictionary. A research assistant did the actual determination of the genre for the majority of the Web pages. After looking at each page, the coder chose the appropriate genres from a pop-up list in a database program. If none of the already defined genres were appropriate, the coder could add new genres to the list. A second rater re-did 40% of classification, in order to check the reliability of the first rater. In a few cases, they found that some of the Web pages that could not be classified because they did not know the name of the genre, or because the pages had not a recognizable genre. In these cases, the raters agreed that there was a genre, but did not know its name. They considered it to be a variant of an accepted genre, missing some features or including new features. Web pages without a genre were labeled as “unclassified”. Interestingly, one of the conclusions was that some of these unclassified pages could be interpreted as “emerging genres”.

Similarly in Roussinov et al. (2001)’s explorative user study on Web genres, not all pages could be classified. In some cases the respondents gave answers such as “I don’t know”. But no special conclusions were drawn from this behaviour.

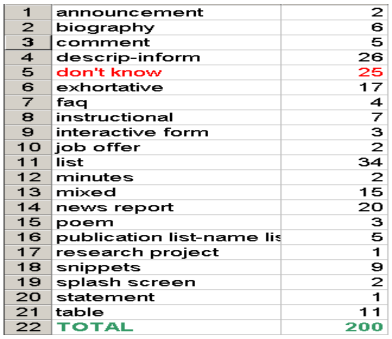
On the other hand, it seems that Haas and Grams (1998) had no unclassified pages in their qualitative survey of 75 Web pages, and nor did Shepherd and Watters (1999) in their analysis of 96 websites.

It is understandable that the main interest of Web pages surveys focuses on what can be classified. However, Web pages that are “unclassified” today, may be found to belong to a new Web genre tomorrow. In this respect, unclassified Web pages can be seen as anticipations or forerunners of novel genres, currently not fully formed.

3 Difficulty in naming a Web Page by Genre

Giving a genre label to a Web page is not always as straightforward as one might think.

As a preliminary step, we tried to assign a genre label manually (one rater) to a random sample of 200 Web pages, extracted from the SPIRIT collection (Clarke et al. 1998, see below). The resulting classification is shown in Figure 1.



1	announcement	2
2	biography	6
3	comment	5
4	descrip-inform	26
5	don't know	25
6	exhortative	17
7	faq	4
8	instructional	7
9	interactive form	3
10	job offer	2
11	list	34
12	minutes	2
13	mixed	15
14	news report	20
15	poem	3
16	publication list-name lis	5
17	research project	1
18	snippets	9
19	splash screen	2
20	statement	1
21	table	11
22	TOTAL	200

Figure 1. Manual Classification of 200 Web pages from the SPIRIT collection

For 25 pages we could not find a label at all, for 15 pages we could say only that they were “mixed”; the rest of the labels shifts from genre labels (*biography, minutes, poems, FAQs, comments, news reports*, etc.), to rhetorical/discoursal labels (exhortative, instructional, descriptive-informational, etc.), to labels that account for the layout structure (tables, lists, etc.).

It seems that there is a wide gap between Web genres which have been acknowledged and which can be easily recognized, like the ones shown in Figure 2, and Web pages, like the ones in Figure3, for which applying a genre label is not easy.

One solution to address this problem is to use statistics to automatically group similar Web pages together. The resulting groups can be then analyzed to see whether there are similarities among the members of a group, and whether these similarities help sketch the profile of an emerging genre.

Housing Assignment Services
STANFORD UNIVERSITY
Frequently A
ON THIS PAGE
Eugene Charniak
Box 1910, Computer Science Department
Brown University
Providence, RI 02912
401-863-7636 (voice)
401-863-7657 (fax)
The New York Times
real jobs
COACH
Violence in Iraq Forces 2 Big Contractors to Curb Work
Kerry Opens New Bush Attack, Focusing on Iraq and Economy
MARKETS

Figure 2. Well-Established Web Genres (not in the SPIRIT collection)

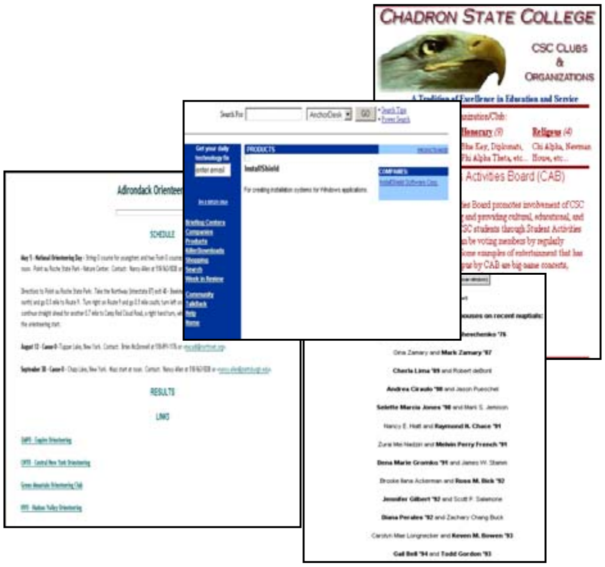


Figure 3. Emerging Web Genres? Some Web Pages in the SPIRIT collection

4 Experiment

4.1 Web page Collection

The SPIRIT collection is a random crawl with an initial seed of university websites carried out in 2001 by a Canadian university (Clarke et al. 1998). It contains single Web pages and not complete websites. The size of the collection is one terabyte, and the number of Web pages (mostly HTML files) is about 95 million. It is multilingual and without any meta-information, except a short header including the original URL, the date and time when the pages were crawled from the Web, and few other details. It represents a genuine slice of the real Web, "frozen" in a collection for research purposes.

Only Web pages written in English were used for this experiment.

4.2 Methodology

Cluster analysis is said to be an objective methodology for quantifying the structural characteristics of a set of observations. It is considered to be a potent analytical tool because it has the potential to reveal structures within the data by grouping homogeneous objects together, so that objects in the same clusters are more similar to one another than they are to objects in other clusters. The primary value of cluster analysis lies in the grouping of data into clusters as suggested by natural groupings of the data themselves. These clusters presumably reflect some mechanism at work, a mechanism that causes some instances to bear a stronger resemblance to one another than they do to the remaining instances.

Cluster analysis can be used in an exploratory or confirmatory way. Here the aim is exploratory.

It is important to highlight that the success of clustering is measured subjectively in terms of how useful the results appear to be to a human user.

The clustering algorithm chosen for our experiment was K-means, as implemented in SPSS (Norusis 1999). K-means is suitable for large datasets, it is easy to understand and very fast. It is based on the following steps:

1. Pick up a number of cluster centers.
2. Assign every item to its nearest cluster center.
3. Move each cluster center to the mean of its assigned items.
4. Repeat until convergence occurs.
5. Objects may be reassigned if they are closer to another cluster than the original assigned.

There are two main problems with K-means: the number of clusters and the set of initial seeds. Which is the number of clusters that better represent the data? This number must be specified in advance and is not suggested by the algorithm. Which are the optimal initial seeds? The quality and reliability of the final cluster solution heavily depends on the initial seeds. Unfortunately, there is not just a single answer to these questions (cf. Anderberg 1973). The choices for the current experiment are given below.

4.3 Shallow Features

For this experiment we rely on easily extractable linguistic features, previously used in genre identification studies, and on HTML tags. The total number of features is 238, with the following breakdown:

- 50 most frequent words in English (Stamatatos et al. 2000);
- 31 Parts-of-Speech (POS) tag[1] frequencies (Rayson et al. 2002);

[1] POS tags were extracted using Connexor (Tapanainen and Järvinen 1997).

- 100 POS trigram frequencies (Argamon 1998, Santini 2004);
- 57 HTML tag frequencies.

4.4 Experimental Question

Even though the aim of the experiment was merely exploratory, the task was challenging:

Premise 1: having a collection of unclassified Web pages,
Premise 2: using cluster analysis, which is claimed to have the potential to reveal structures within the data,
Premise 3: leveraging on sets of features used in previous successful studies on automatic genre identification,
Question: is it possible to detect patterns that could suggest the formation of emerging genres?

4.5 Steps

1. Extraction of five samples of 200 English Web pages each (1000 Web pages in total) from the SPIRIT collection;
2. Extraction and frequency counts of 238 features;
3. Feature reduction using Principal Component Analysis (PCA);
4. Cluster analysis over the components resulting from PCA;
5. Qualitative analysis of the clusters returned by cluster analysis;
6. Comparison of the clusters across the different samples.

Outliers were considered to be representatives of subpopulations, and left in the samples.

The requirements of normality, linearity, and homoscedasticity have little influence on cluster analysis. But a critical issue is multicollinarity (Hair et al. 1998). Variables that are multicollinear (i.e. the ones that show some degree of correlation among them) are implicitly weighted more heavily: multicollinearity acts as a weighting process that affects the results. In order to address this problem, PCA was run. All the components resulting from PCA were used for cluster analysis.

Many different techniques were tried to decide the optimal initial seeds and the most representative number of clusters for each sample.

After many attempts with random seeds, selected seeds, and seeds suggested from Hierachical Cluster Analysis, a final decision was made for random seeds, also supported by Peña et al. (1999). Random seed selection is also the default in SPSS, which uses a parallel threshold, selects several cluster seeds simultaneously, and assign objects within the threshold distance to the nearest seed.

In order to determine the appropriate number of clusters, K-means was run several times on the datasets, specifying each time a different number of clusters, from 6 to 12. The choice of this range is simply empirical, mainly based on reflections on the manual classification (see Figure 1). For each cluster solution, the distance between clusters and the distance within each cluster were calculated. The cluster solutions were different for each sample. The preferred solution had the maximum distance between clusters and the minimum distance within each single cluster. This approach ensures the highest distinctiveness and the highest compactness of a cluster solution.

Cluster analysis was repeated on five samples in order to see whether the nature of the clusters was consistent among the different samples.

4.6 Results

The graphical representation of the main clusters of the five samples showed that they had distinctive and non-overlapping shapes. Some clusters were more well-defined than others.

As mentioned earlier, the assessment of the cluster is measured subjectively. In our analysis of the clusters, we picked up Web pages which were the nearest to the cluster centroids. Intuitively, Web pages, which are very close to a cluster centroid are the most representative for that cluster.

Some regularities came out from the qualitative analysis of the clusters. Four main rhetorical/discoursal types were identified across the five samples: informational, instructional, argumentative, and, to a smaller extent, narrative.

Informational Web pages show regular traits, such as 3rd person point of view, a detached, and non-emphatic tone, etc. Their goal is to give an objective and impersonal explanation or description, which aims at being reliable. The genres enacting these traits are "*about" pages*, *news releases, book reviews*, etc. In argumentative Web pages, the structure of the argumentation is often implemented through rhetorical questions, the use of the 1st person pronouns, 2nd person pronouns to address and involve the reader in a direct way, etc. The writer's intention to convince the reader is evident. A large variety of genres are used for this purpose: *sermons, editorials, answers to letters*, etc. Narrative texts are present to a lower extent in the samples. Examples of narration are *meeting minutes* and *news reports*. The instructional type includes *guidelines, suggestions, recommendations, FAQs*, etc. There is a proliferation of 2nd person pronouns, *should*, deontic future, etc. This text profiling seems to be cross-genre, very much the same as the concept of "text type" as intended by Biber (1988).

However, some noise was also returned. For example, some clusters could be divided into two subgroups, i.e. a single cluster grouped at the same time Web pages with continuous text, the majority, and Web pages with little continuous text (for example, Web pages similar to those shown in Figures 5 and 7), that we assessed to be noise because they were a minority in the cluster. No plausible explanations could be found for this phenomenon.

4.7 Rhetorical/Discoursal Types and Web Users

In order to validate our analysis of the rhetorical/discoursal types perceived in the clusters, a small study with Web users was set up. Five PhD students in Computational Linguistics and a researcher were chosen as subjects. Six texts from our samples (shown in the Appendix) were distributed. Only six texts were used because we wanted a thoughtful answer without burdening the subjects.

Two tests were set up. In Test 1, the subjects were given the six texts in random order and were instructed to group them according to their type by hand. No training was provided. Subjects were neither given examples nor a list of possible labels. They were only told to ignore the topic, i.e. their groupings could be based on any criteria except the topic. They were allowed to create at least two but fewer than six groups. The subjects were then asked to assign a label (existing or invented) to each of their groups. In Test 2, the subjects received the same Web pages, but this time they were asked to group them according to three labels: ARGUMENTATIVE, INFORMATIONAL, OTHER. Table 1 reports our personal classification of the six Web pages for Test 1 and Test 2:

File Name	Test 1	Test 2
SPRT_025_058_104_0051906	sermon	argumentative
SPRT_025_058_105_0052155	e-shop	other
SPRT_022_009_162_0080733	editorial	argumentative
SPRT_003_002_167_0083008	introduction	informational
SPRT_010_049_112_0055827	news release	informational
SPRT_010_049_112_0055878	search page	other

Table 1. Personal classification for Test 1 and Test 2

Here are the labels suggested by the subjects for Test 1 (Table 2 and Table 3):

File Name	S_1	S_2	S_3
SPRT_025_058_104_0051906	debate or controversy	short paragraphs	Philosofical Dissertation
SPRT_025_058_105_0052155	search/user's interface	links	e-commerce portals
SPRT_022_009_162_0080733	debate or controversy	long paragraphs	Philosofical Dissertation
SPRT_003_002_167_0083008	news	short paragraphs	Organization's Information
SPRT_010_049_112_0055827	news	long paragraphs	Business Information
SPRT_010_049_112_0055878	search/user's interface	user input	e-commerce portals

Table 2. Classification of Subjects 1, 2 and 3 for Test 1

File Name	S_4	S_5	S_6
SPRT_025_058_104_0051906	analysis	editorial	commentary
SPRT_025_058_105_0052155	merchant page	information index	selling
SPRT_022_009_162_0080733	analysis	editorial	commentary
SPRT_003_002_167_0083008	announcement/press release	organization introduction	introduction
SPRT_010_049_112_0055827	announcement/press release	news	fact description
SPRT_010_049_112_0055878	search page	information index	search

Table 3. Classification of Subjects 4, 5 and 6 for Test 1

When the subjects could choose freely their non-topical labels, there was a wide range of variation (Table 2 and Table 3). However, some of the categories suggested could be seen as belonging to the same family, for example *sermons*, *debates*, *controversies*, *commentaries*, and *editorials* have all common rhetorical/discoursal traits, because in all of them there is an argument being supported. This seems to confirm the idea that a predominant rhetorical/discoursal strategy is enacted within different, but related genres (cf. Werlich 1976). Some degree of consistency came out also for the labels suggested for Figure 5 and Figure 7: Figure 5 is mostly related to "search", and Figure 7 to "selling". When the subjects were asked to choose among the restricted number of labels suggested for Test 2, their choices were extremely consistent. Their classification is shown in Table 4 (arg=argumentative; inf=informational; oth=other):

File Name	S_1	S_2	S_3	S_4	S_5	S_6
SPRT_025_058_104_0051906	arg	arg	arg	arg	arg	arg
SPRT_025_058_105_0052155	oth	oth	oth	oth	inf	oth
SPRT_022_009_162_0080733	arg	arg	arg	arg	arg	arg
SPRT_003_002_167_0083008	inf	inf	inf	inf	inf	inf
SPRT_010_049_112_0055827	inf	inf	inf	inf	inf	inf
SPRT_010_049_112_0055878	oth	oth	oth	oth	inf	oth

Table 4. Classification of the 6 subjects for Test 2

There is almost unanimous agreement in classifying the Web pages; only Subject 5 used the label INFORMATIONAL for Figure 5 and Figure 7, while all other subjects classified them as OTHER. In order to assess the degree of agreement beyond any random chance, we computed the K statistic as an inter-rater measure, and the value returned was 0.85, indicating a very good level of agreement among the subjects (Carletta 1996).

This user study suggests that the rhetorical/discoursal properties of a text are recognized and applied consistently by the Web users when their label is hinted; when no indications are given, except the restriction of not using the topic as a category, a wide range of different labels come out, some of them showing a degree of consistency from a rhetorical/discoursal point of view. As for Figure 5 and Figure 7, there is a tendency to classify them according to their function ("search page") and purpose ("selling page"*)*.

5 Conclusions and Future Work

Even if cluster analysis did not return any evident textual patterns that could suggest emerging genres, we found out that the traditional rhetorical/discoursal types are extensively used on the Web. Regardless of all the uncertainties related to decision-making in cluster analysis[2], regardless of some noise in the clusters, consistent rhetorical/discoursal profiles could be identified. These profiles are easily recognized by Web users: probably the needs to inform or describe, instruct, convince, narrate are universal, and these types are easily produced and recognized (cf. also Faigley and Meyer 1983). If this is true, it is also true that some Web pages do not fit well into them, for example Figure 5 and Figure 7.

One interesting point that comes out from this exploratory study is the correlation and interaction between genres and rethorical/discoursal types. This interaction needs a better understanding. Biber (1988) showed that his 'text types' (corresponding to our rethorical/discoursal types) cut across genre categories. Other linguists (for example, Werlich 1976) highlighted that there is usually one predominant text type within a genre, and that this main rethorical/discoursal type is intermingled with other types. We suggest that the dosing of different rhetorical devices might account for the variation among different genres.

An important remark to make is about features. We wonder whether the features used in this experiment were too "shallow" to identify patterns that may lie at a deeper level of the textual structure. There is probably a correlation between shallow features and the noise returned. Deeper feature can possibly return more robust clusters. The role of HTML tags in text profiling remains unclear.

Future work includes:

1. Cluster analysis with a new set of features, i.e. deep linguistic features, such as syntactic patterns and lexical items that can be interpreted functionally, for example *communication verbs* (widely used in reported speech) or *that omission* (extensively employed in "informal" genres).
2. Exploration of rhetorical/discoursal variations within a single Web page (in order to investigate the mixed nature of a text), and among Web pages (in order to explore the interaction between different discourse strategies and different genres).
3. Investigation of the interaction between linguistic features and layout (and other visual cues). The connection between genres, rhetorical/discoursal types and layout remains to be explored for Web pages.
4. A more extensive study on Web users' agreement on genre labels and rethorical/discoursal labels for Web pages.

6 References

[1] Anderberg, M. (1973), *Cluster Analysis for Application*, Academic Press, New York and London.

[2] Argamon S., Koppel M., and Avneri G. (1998), "Routing documents according to style", *Proceedings of the First International Workshop on Innovative Internet Information Systems* (IIIS-98).

[3] Biber, D. (1988), *Variations across speech and writing*, Cambridge University Press, Cambridge.

[4] Carletta, J. (1996), "Assessing agreement on classification tasks: the kappa statistic", *Computational Linguistics*, Vol. 22, 2, 249-254.

[5] Clarke C., Cormack G., Laszlo M., Lynam T., and Terra E. (1998), "The Impact of Corpus Size on Question Answering Performance", *Proc. of the 25th Annual International ACM SIGIR Conference on Research and Development in IR*, Tampere, Finland.

[6] Crowston, K., and Williams, M. (1997), "Reproduced and emergent genres of communication on the World-Wide Web" *Proc. of the 30th Hawaii International Conference on System Sciences,* Hawaii, USA.

[7] Crowston, K., and Williams, M. (1999), The Effects of Linking on Genres of Web Documents, *Proc. of the 32nd Hawaii International Conference on System Sciences,* Hawaii, USA.

[8] Dillon A. and Gushrowski B. (2000), "Genres and the Web: is the personal home page the first uniquely digital genre?", Journal of the American Society for Information Science, 51, 2.

[9] Faigley, L., and Meyer, P. (1983), "Rhetorical theory and readers' classification of text types", *Text*, Vol. 3, 305-325.

[2] In cluster analysis, a high number of alternatives is available and it is hard to know which is the best for the actual data. The choice between the different clustering methods should be dictated by the nature of the mechanisms that are thought to underlie the particular clustering phenomenon. However, these mechanisms are rarely known at the beginning of the investigation.

[10] Haas, S., and Grams, E. (1998), "Page and Link Classifications: Connecting Diverse Resources", *Proceedings of Digital Libraries '98 – Third ACM Conference on Digital Libraries, 99-107.*

[11] Hair, J., Anderson, R., Tatham, R., and Black, W. (1998), *Multivariate Data Analysis*, Prentice-Hall International Inc, 5th Edition.

[12] Norusis, M. (1999), *SPSS Base 10.0 Application Guide*, SPSS Inc. USA.

[13] Peña, J., Lozano, J., and Larrañaga, P. (1999). "An Empirical Comparison of Four Initialization Methods for the K-Means Algorithm", *Pattern Recognition Letters*, Vol. 20, 10, 1027-1040.

[14] Rayson, P., Wilson, A., and Leech, G. (2002), "Grammatical word class variation within the British National Corpus Sampler", in Peters, P., Collins, P., Smith, A., (eds.), *New Frontiers of Corpus Research*, Rodopi, Amsterdam, New York.

[15] Roussinov D., Crowston K., Nilan M., Kwasnik B., Cai J., Liu X. (2001), "Genre Based Navigation on the Web", *Proceedings of the 34th Hawaii International Conference on System Sciences* (HICSS-34).

[16] Santini, M. (2004), "A Shallow Approach To Syntactic Feature Extraction For Genre Classification", *Proc. of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*, Birmingham, UK.

[17] Shepherd, M., and Watters, C. (1998), "The Evolution of Cybergenre", *Proceedings of the 31st Hawaii International Conference on System Sciences*, Hawaii, USA.

[18] Shepherd, M., and Watters, C., (1999), "The Functionality Attribute of Cybergenres", *Proceedings of the 32nd Hawaii International Conference on System Sciences*, Hawaii, USA

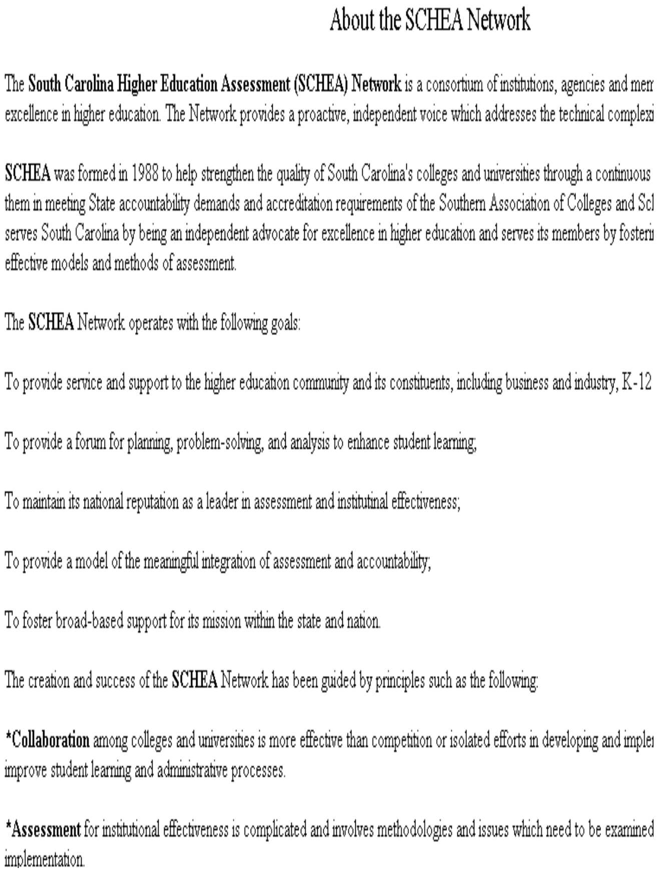
[19] Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2000), "Text Genre Detection Using Common Word Frequencies", *COLING 2000, 18th International Conference on Computational Linguistics*, Vol. 2, 808-814, Luxembourg.

[20] Tapanainen, P. and Järvinen, T. (1997), "A non-projective dependency parser", *Proceedings of the 5th Conference on Applied Natural Language Processing*.

[21] Werlich, E. (1976), *A Text Grammar of English*, Quelle & Meyer, Heidelberg, Germany.

Appendix

The Web pages used for the Web users' study are shown in this Appendix.



About the SCHEA Network

The **South Carolina Higher Education Assessment (SCHEA) Network** is a consortium of institutions, agencies and mem
excellence in higher education. The Network provides a proactive, independent voice which addresses the technical complexi

SCHEA was formed in 1988 to help strengthen the quality of South Carolina's colleges and universities through a continuous
them in meeting State accountability demands and accreditation requirements of the Southern Association of Colleges and Sc
serves South Carolina by being an independent advocate for excellence in higher education and serves its members by fosteri
effective models and methods of assessment.

The **SCHEA** Network operates with the following goals:

To provide service and support to the higher education community and its constituents, including business and industry, K-12

To provide a forum for planning, problem-solving, and analysis to enhance student learning;

To maintain its national reputation as a leader in assessment and institutinal effectiveness;

To provide a model of the meaningful integration of assessment and accountability;

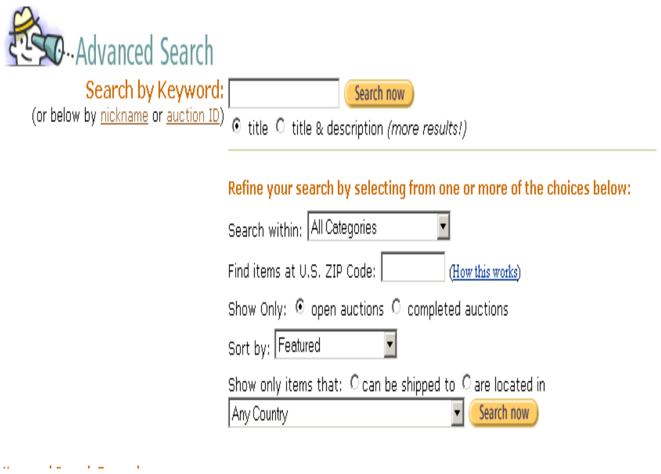
To foster broad-based support for its mission within the state and nation.

The creation and success of the **SCHEA** Network has been guided by principles such as the following:

***Collaboration** among colleges and universities is more effective than competition or isolated efforts in developing and imple
improve student learning and administrative processes.

***Assessment** for institutional effectiveness is complicated and involves methodologies and issues which need to be examined
implementation.

Figure 4. SPRT_003_002_167_0083008



Search by Keyword:

(or below by nickname or auction ID)

● title ○ title & description *(more results!)*

Refine your search by selecting from one or more of the choices below:

Search within: All Categories

Find items at U.S. ZIP Code: (How this works)

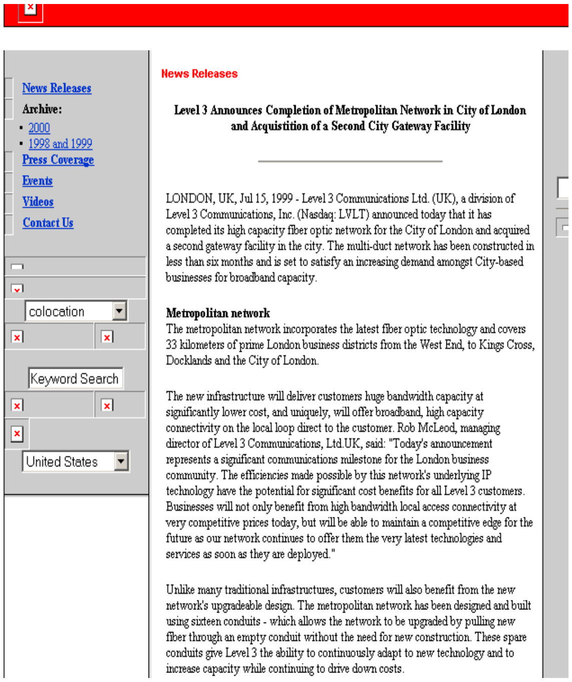
Show Only: ● open auctions ○ completed auctions

Sort by: Featured

Show only items that: ○ can be shipped to ○ are located in

Any Country

Figure 5. SPRT_010_049_112_0055878



News Releases

Level 3 Announces Completion of Metropolitan Network in City of London and Acquistition of a Second City Gateway Facility

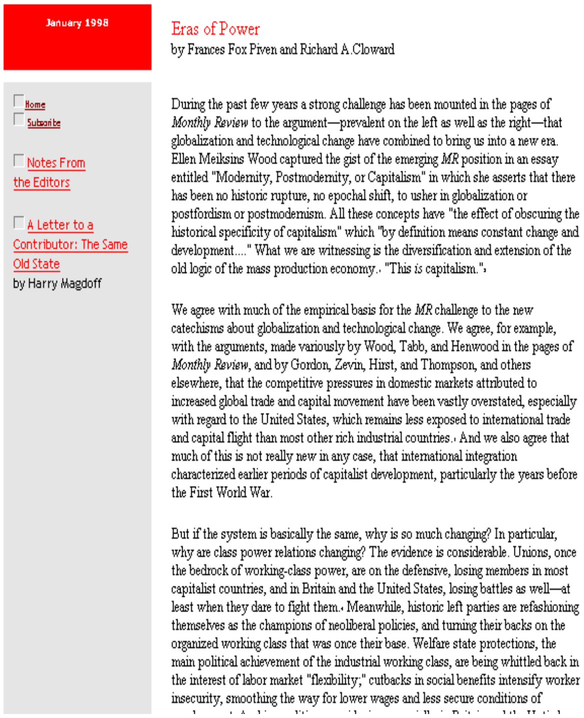
LONDON, UK, Jul 15, 1999 - Level 3 Communications Ltd. (UK), a division of Level 3 Communications, Inc. (Nasdaq: LVLT) announced today that it has completed its high capacity fiber optic network for the City of London and acquired a second gateway facility in the city. The multi-duct network has been constructed in less than six months and is set to satisfy an increasing demand amongst City-based businesses for broadband capacity.

Metropolitan network
The metropolitan network incorporates the latest fiber optic technology and covers 33 kilometers of prime London business districts from the West End, to Kings Cross, Docklands and the City of London.

The new infrastructure will deliver customers huge bandwidth capacity at significantly lower cost, and uniquely, will offer broadband, high capacity connectivity on the local loop direct to the customer. Rob McLeod, managing director of Level 3 Communications, Ltd.UK, said: "Today's announcement represents a significant communications milestone for the London business community. The efficiencies made possible by this network's underlying IP technology have the potential for significant cost benefits for all Level 3 customers. Businesses will not only benefit from high bandwidth local access connectivity at very competitive prices today, but will be able to maintain a competitive edge for the future as our network continues to offer them the very latest technologies and services as soon as they are deployed."

Unlike many traditional infrastructures, customers will also benefit from the new network's upgradeable design. The metropolitan network has been designed and built using sixteen conduits - which allows the network to be upgraded by pulling new fiber through an empty conduit without the need for new construction. These spare conduits give Level 3 the ability to continuously adapt to new technology and to increase capacity while continuing to drive down costs.

Figure 6. SPRT_010_049_112_0055827



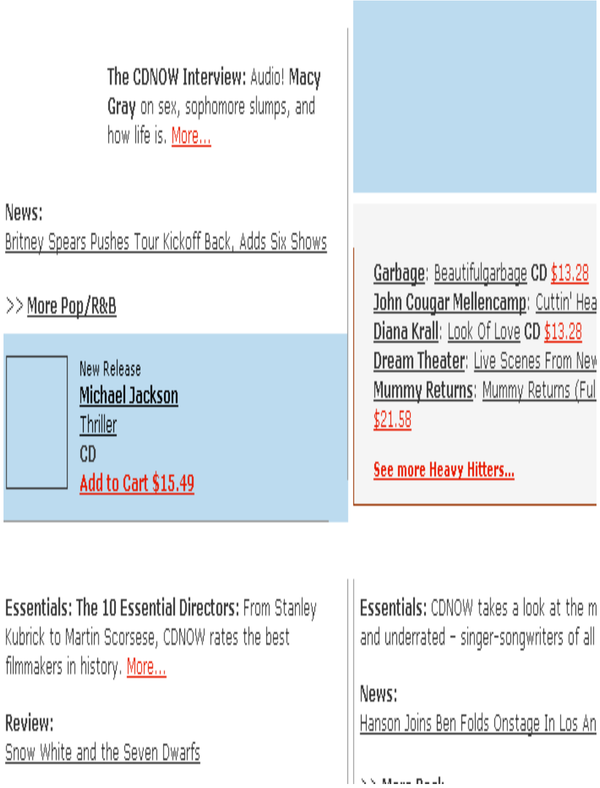
Eras of Power
by Frances Fox Piven and Richard A.Cloward

During the past few years a strong challenge has been mounted in the pages of *Monthly Review* to the argument—prevalent on the left as well as the right—that globalization and technological change have combined to bring us into a new era. Ellen Meiksins Wood captured the gist of the emerging *MR* position in an essay entitled "Modernity, Postmodernity, or Capitalism" in which she asserts that there has been no historic rupture, no epochal shift, to usher in globalization or postfordism or postmodernism. All these concepts have "the effect of obscuring the historical specificity of capitalism" which "by definition means constant change and development...." What we are witnessing is the diversification and extension of the old logic of the mass production economy. "This *is* capitalism."

We agree with much of the empirical basis for the *MR* challenge to the new catechisms about globalization and technological change. We agree, for example, with the arguments, made variously by Wood, Tabb, and Henwood in the pages of *Monthly Review*, and by Gordon, Zevin, Hirst, and Thompson, and others elsewhere, that the competitive pressures in domestic markets attributed to increased global trade and capital movement have been vastly overstated, especially with regard to the United States, which remains less exposed to international trade and capital flight than most other rich industrial countries. And we also agree that much of this is not really new in any case, that international integration characterized earlier periods of capitalist development, particularly the years before the First World War.

But if the system is basically the same, why is so much changing? In particular, why are class power relations changing? The evidence is considerable. Unions, once the bedrock of working-class power, are on the defensive, losing members in most capitalist countries, and in Britain and the United States, losing battles as well—at least when they dare to fight them. Meanwhile, historic left parties are refashioning themselves as the champions of neoliberal policies, and turning their backs on the organized working class that was once their base. Welfare state protections, the main political achievement of the industrial working class, are being whittled back in the interest of labor market "flexibility;" cutbacks in social benefits intensify worker insecurity, smoothing the way for lower wages and less secure conditions of

Figure 8. SPRT_022_009_162_0080733



The CDNOW Interview: Audio! Macy Gray on sex, sophomore slumps, and how life is. More...

News:
Britney Spears Pushes Tour Kickoff Back, Adds Six Shows

>> More Pop/R&B

New Release
Michael Jackson
Thriller
CD
Add to Cart $15.49

Garbage: Beautifulgarbage CD $13.28
John Cougar Mellencamp: Cuttin' Hea
Diana Krall: Look Of Love CD $13.28
Dream Theater: Live Scenes From New
Mummy Returns: Mummy Returns (Ful
$21.58

See more Heavy Hitters...

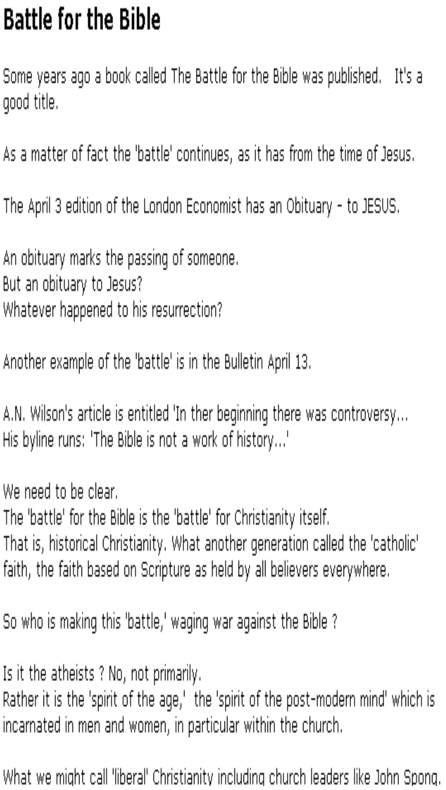
Essentials: The 10 Essential Directors: From Stanley Kubrick to Martin Scorsese, CDNOW rates the best filmmakers in history. More...

Review:
Snow White and the Seven Dwarfs

Essentials: CDNOW takes a look at the m
and underrated - singer-songwriters of all

News:
Hanson Joins Ben Folds Onstage In Los An

Figure 7. SPRT_025_058_105_0052155



Battle for the Bible

Some years ago a book called The Battle for the Bible was published. It's a good title.

As a matter of fact the 'battle' continues, as it has from the time of Jesus.

The April 3 edition of the London Economist has an Obituary - to JESUS.

An obituary marks the passing of someone.
But an obituary to Jesus?
Whatever happened to his resurrection?

Another example of the 'battle' is in the Bulletin April 13.

A.N. Wilson's article is entitled 'In ther beginning there was controversy...
His byline runs: 'The Bible is not a work of history...'

We need to be clear.
The 'battle' for the Bible is the 'battle' for Christianity itself.
That is, historical Christianity. What another generation called the 'catholic' faith, the faith based on Scripture as held by all believers everywhere.

So who is making this 'battle,' waging war against the Bible ?

Is it the atheists ? No, not primarily.
Rather it is the 'spirit of the age,' the 'spirit of the post-modern mind' which is incarnated in men and women, in particular within the church.

What we might call 'liberal' Christianity including church leaders like John Spong.

Figure 9. SPRT_025_058_104_0051906