Clustering Web Pages to Identify Emerging Textual Patterns

Marina Santini

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

Mots-clefs – Keywords

genre textuels sur le Web, typologie des pages Web, analyse de groupement

Web genres, Web Text Types, Web pages, Cluster Analysis

Résumé – Abstract

Le Web a causé beaucoup de changements dans plusieurs domaines. Il a aussi influencé l'inventaire des genres textuels traditionnels. De nouveaux genres ont été créés, par exemple les *blogues* et *foires aux questions*. Il est probable que d'autres genres soient en train de se former, parce que le Web est un médium qui change constamment. Dans cet article, nous présentons une expérience qui vise à faire apparaître de façon inductive les plans textuels émergents, qui peuvent devenir un nouveau genre ou une nouvelle typologie textuelle dans peu de temps. Il s'agit de regrouper (analyse de groupement) les pages web en utilisant des traits linguistiques et de présentation. Les résultats sont encourageants et invitent à poursuivre la recherche dans ce domaine.

The Web has triggered many adjustments in many fields. It also has had a strong impact on the genre repertoire. Novel genres have already emerged, e.g. *blog* and *FAQs*. Presumably, other new genres are still in formation, because the Web is still fluid and in constant change. In this paper we present an experiment that explores the possibility of automatically detecting the emerging textual patterns that are slowly taking shape on the Web. Emerging textual patterns can develop into novel Web genres or novel text types in the near future. The experimental set up includes a collection of unclassified web pages, two sets of features and the use of cluster analysis. Results are encouraging and deserve further investigation.

1 The Web: An Evolution still in progress

The Web has had a strong impact on the genre repertoire. Novel genres have already emerged (Crowston and Williams 1997) such as *personal home pages, hotlists, FAQs*, and more recently *blogs, ezines*, *clogs*, etc. Presumably, other new genres are still in formation, because the Web is still fluid and in constant change. Three main well-established genre categories have been identified on the Web: reproduced/replicated, adapted/variant, novel/spontaneous (Crowston and Williams 1997; Shepherd and Watters 1998). 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, such as *academic papers*. However, most genres coming from the paper world have undergone some adjustments when moving on to the Web and variants have been created. For instance, *online newspapers* and *online manuals* show the adaptation of paper genres to the functionalities provided by the Web (cf. Crowston and Williams 1999, Shepherd and Watters 1999). Some genres are novel and spontaneous. They have become fully acknowledged and genre labels have been invented for them only in recent years, for instance *home pages* (personal, academic, organizational, etc.), *FAQs, newsletters, emails, weblogs*. However, there are many web pages that do not fall into one of these three categories. It is often hard to assign a genre label to a web page. Many web pages remain "unclassified" or labelled as "mixed" (Santini 2005a). We suggest that those web pages that are unclassifiable might represent an emerging textual pattern, i.e. a new textual organization, strongly influenced by the hypertextual structure and the functionalities provided by the Web.

In this paper we present an experiment that explores the possibility of automatically detecting emerging textual patterns that are slowly taking shape on the Web. Emerging textual patterns are interesting because they can reveal novel genres or novel text types in an embryonic form. The proposed approach consists in running cluster analysis (an inductive/unsupervised statistical algorithm based on similarity measures) in order to create groups of similar pages across a corpus of 1000 unclassified English web pages. The qualitative analysis of these groupings (the clusters) would reveal whether emerging patterns could be identified. This experiment does not include any classification tasks, because it would be hard to classify something which is not fully formed. The goal here is the analysis and the interpretation of new textual patterns, if any, brought about by the dynamism of the Web.

As mentioned earlier, emerging textual patterns are embryonic forms that are likely to develop into novel Web genres or novel text types in the near future. While emerging textual patterns are related to groups of documents that show new textual traits, genres and text types refer to fully formed categories. Many definitions of genre and text types have been formulated since Aristotle. Here, by genre, we refer to the socio-cultural connotation of a document together with the linguistic/discoursal devices enacted in the document itself. For instance, a *letter*, a *manual*, an *article*, an *academic paper* are genres. Web genres are genres that are used on the Web, and they range from plain electronic versions of paper genres, to genres more tailored to take advantage of the potentials of the Web. By text types, we refer to the purpose of the text, i.e. the reason for which a text has been written. Text types are related to the producer's intention towards the receiver(s). An advert is written to *persuade* customers to buy something; a car manual might *instruct* on how to fix a component in the car. Researchers involved in automatic genre classification rely on texts which are pre-classified by genres and use discriminant analysis or supervised learning to "learn" from exemplar texts belonging to a restricted set of genres and generalize this learning over unseen/unclassified documents. The quantitative approach to text types identification, instead, is mainly linked to the multi-dimensional analysis proposed by Biber (1988, 1989, 1995, 2004). His text types cut across genres (Biber 1988, 1989) or registers (Biber 1995). More recently, he has started

sketching a typology of web registers (Biber 2004) by using two main Google topical categories incorporating multiple subcategories. Biber's approach has strongly influenced two projects for French, TyPText (Folch et al. 2000 and Illouz et al. 2000) and TyPWeb (Beaudouin et al. 2001a and 2001b) (see Santini 2004 for a state-of-the-art of genre and text type identification).

The paper is organized as follows: section 2 briefly describes previous work on the same subject and identifies references to "unclassifiable" web pages in web user studies; section 3 describes the experiment and the results; section 4 draws some conclusions.

2 Related Work

So far, only a very recent exploratory study has been carried out to address the issue of automatic detection of emerging textual patterns on the Web, more specifically the identification of genres still in formation (Santini 2005a). From this preliminary study, it appears that although automatic clustering did not return any emerging genres, traditional rhetorical/discoursal types could be identified, despite some noise. The presence of this noise was ascribed to the use of shallow features, too shallow to highlight textual novelties.

Some references to "unclassified" web pages can be derived indirectly from the few surveys carried out so far on web pages. Crowston and Williams (1997) found that some of the web pages could not be classified because they did not have a recognizable genre. In these cases, the raters agreed that there was a genre, but did not know its name, and labelled the pages as "unclassified". Interestingly, one of the conclusions was that some of these unclassified pages could be interpreted as belonging to "emerging genres". Similarly, in Roussinov et al. (2001)'s exploratory user study on Web genres, a number of pages could not be classified, but no special conclusions were drawn.

Understandably, the main interest of web page surveys is to find what can be classified. However, web pages that are "unclassified" today, might become instantiations of a new web genre or a new text type tomorrow. In this respect, unclassified web pages could be seen as anticipations or forerunners of new textual categories, not fully formed yet.

3 Experiment

3.1 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 rather than complete websites. This collection includes about 95 million web pages. 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 a few other details. It represents a genuine slice of the real Web. 1000 random English web pages were extracted from this collection and used in the experiment.

3.2 Features

Two sets of features were used, each including three subsets. The first set of features includes 28 functional cues, 29 syntactic patterns, and 33 HTML tags (90 features). We will refer to this set as *sy_pat*. The second set includes instead 28 functional cues, 52 connectives and subordinators, and 33 HTML tags (113 features). We will refer to this set as *con_sub*. Functional cues and syntactic patterns are hand-crafted and parser-dependent features (the

parser used in this experiment is Connexor by Tapanainen and Järvinen (1997); Santini (2005b) contains the description and motivation of these features). Connectives and subordinators are lexical items. They represent an easy way to capture syntactic and discoursal information, even though their semantic interpretation is often ambiguous. Finally, HTML tags account for layout and functionalities, both important elements in a web page.

Two sets of features represent two views on the same data. As cluster analysis has a somewhat subjective nature, cluster solutions must be validated. Here we use the extent of the overlap between the two cluster solutions returned by the two sets of features as a measure of the stability of the final clusters.

3.3 Methodology

Cluster analysis is said to have the potential to reveal structures within the data by grouping homogeneous objects together on the basis of similarity measures. It can be used in an exploratory or confirmatory way. Here the aim is exploratory. The clustering algorithm chosen for the experiment is K-means, as implemented in SPSS. K-means is suitable for large datasets, it is easy to understand and very fast. However, it involves two hard decisions, one concerning the number of clusters that better represent the data, the other related to the set of initial seeds. Several alternatives are available (cf. Anderberg 1973). In this experiment, random seeds were used (default in SPSS), and the number of clusters was selected on the basis of the maximum distance between clusters and the minimum distance within each cluster. This approach ensures the highest distinctiveness and compactness of a solution. A 15-cluster solution was suggested for both sets of features.

The following steps were performed:

- Extraction of 1000 random English web pages from the SPIRIT collection.
- Parsing of the text-only version of the web pages.
- Extraction and frequency counts of the two sets of features.
- Normalization of the frequencies by the number of words in a web page.
- Transformation of the normalized frequencies into z-scores (z-scores represent the number of standard deviations that a raw score is above or below the mean; they represent the deviation from a "norm", and can be used as a way of weighing features within a corpus).
- Selection of the best cluster solutions, one for each set of features.
- Measure of the overlap between the two best cluster solutions.
- Qualitative analysis of the overlapping areas between the two best cluster solutions.

The two cluster solutions were compared using a method commonly used in document clustering, i.e. the comparison by pairs. First, all the possible unique combinations of pairs of the 1000 documents were computed (499,500 pairs). Then an algorithm was built to answer the two following questions: "does the pair get classified as "same" or "different" by the *sy_pat* cluster solution" and "does the pair get classified as same or different by the *con_sub* cluster solution"? The overlap between the two clustering solutions was computed using a two-by-two contingency table. The simple matching coefficient used to measure the overlap had a value of 0.61 (this coefficient ranges from 0=no overlap to 1=full overlap). This value shows an overlap of above 60%, i.e.10% more than the random baseline, and represents an acceptable degree of stability of the clusters returned by the two sets of features. For the qualitative analysis, we selected web pages shared between the two solutions and closest to the cluster centroids (intuitively, the most representative of each cluster).

3.4 Results and Discussion

There is almost a perfect distributional overlap between the minority clusters (see poster). The type of these web pages is easily recognizable. These pages comprises very short server messages, lists of names and extensions, a glossary, bibliographies, tables, summary lists; etc. We asked a web user to manually cluster the 24 web pages in the minority clusters and assign six labels (*server message, telephone directory, bibliographic references, glossary, tabular information, summary list*) to the pages. Then we compared the user's clustering with the automatic clustering and computed the K statistic as an inter-rater measure (Carletta 1996). The value returned was above 0.90, indicating a very good level of agreement. The almost complete agreement of the two cluster solutions with human assessment on the minority clusters is an important confirmation of the validity of the approach.

As for the majority clusters (see poster), web pages in "*con_sub* cluster 6" (368 cases) fall almost entirely (more than 98%) into "*sy_pat* cluster 1". This is a sign of stability. This cluster includes web pages with highly laid-out information, little text, centered (hot)lists, a photograph with personal details, such as address, phone, email, etc. They could be seen as ***contact web pages***. Web pages in "*con_sub* cluster 2" (39 cases) mainly fall into "*sy_pat* cluster 1" (72%). These web pages all share a highly laid-out textual organization, with a large number of hyperlinks, short schematic information, block language, many images. It seems that they share the common purpose of conveying information quickly and exhaustively, leaving to the user the decision whether to display more details by following the hyperlinks. This textual pattern is cross-genres. An e-shop, headlines, an art gallery, an animation website are some of the web genres that were gathered together by the clustering algorithm. The purpose seems to be a ***quick information delivery***. "*Con_sub* cluster 1" (508 cases) is spread across the following *sy_pat* clusters: cluster 1 (35%), cluster 3 (35%) and cluster 9 (25%). This large cluster is more heterogeneous and many of its web pages look like "containers", showing a multi-purpose intent. This cluster is still too diversified to inspire a single label.

In summary, minority clusters show extremely well-defined textual profiles, corresponding to recognized categories by human assessment. As highlighted earlier, these clear-cut clusters confirm that the approach is valid and the features are robust enough to show clear similarities among web pages. Two majority clusters can be seen as emerging textual patterns that we labelled as ***contact web pages*** and ***quick information delivery***. It will be interesting to see whether such textual patterns (still a little bit loose) will develop into a actual web genres or text types in the near future. The largest cluster, instead, shows a kind of mixed textuality and appears to be still too heterogeneous. The cluster is too big to have one single profile. Many of its web pages look like "containers" showing different communicative purposes.

4 Conclusions

Results are encouraging and the approach seems to be effective in providing hints about emerging textual patterns. With unsupervised techniques, deep linguistic features appear to be much more effective than shallow features (cf. Santini 2005a). Although vagueness and elusiveness are common conditions before a novel textual pattern becomes formally and functionally established and recognized, a couple of emerging textual patterns could be identified. In our opinion, they have a good chance of becoming novel web genres or text types once their traits become more tightly woven.

Although the objective evaluation of emergent textual patterns is a new and open issue (any discussions on this subject will be fruitful), as a whole the approach proposed to identify

emerging textual patterns seems to be valid. It may represent a starting point for further investigation on the textuality of web pages.

Références

Anderberg M. (1973), *Cluster Analysis for Application*, Academic Press, New York-London.

Beaudouin V., Fleury S., Habert B., Illouz G., Licoppe C., Pasquier M. (2001a), TyPWeb: décrire la Toile pour mieux comprendre les parcours, *Proc. of CIUST 2001*, France.

Beaudouin V., Fleury S., Habert B., Illouz G., Licoppe C., Pasquier M. (2001b), *Traits textuels, structurels et présentationnels pour typer les sites web personnels et marchands*, available at http://www.atala.org/je/010428/TyPWeb.ppt

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

Biber D. (1989), A typology of English texts, *Linguistics*, Vol. 27, 3-43.

Biber D. (1995), *Dimensions of register variation*, Cambridge University Press, UK.

Biber D. (2004), *Towards a typology of web registers: A multi-dimensional analysis*. Invited lecture, Conference on Corpus Linguistics: Perspectives for the future. Heidelberg University.

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

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 Intern. ACM SIGIR Conf. on Research and Development in IR*, Finland.

Crowston K., Williams M. (1997), Reproduced and Emergent Genres of Communication on the World-Wide Web, *Proc. of the 30 Hawaii Intern. Conf. on System Sciences*, USA.

Crowston K., Williams M. (1999), The Effects of Linking on Genres of Web Documents, *Proc. of the 32 Hawaii Intern. Conf. on System Sciences*, USA.

Folch H., Heiden S., Haber B., Fleury S., Illouz G., Lafon P., Nioche J., Prévost S. (2000), TyPText: Inductive typological text classification by multivariate statistical analysis for NLP systems tuning/evaluation, presented at *LREC 2000*, Greece.

Illouz G., Habert B., Folch H., Heiden S., Fleury Serge, Lafon S., Prévost S. (2000), TyPText: Generic features for Text Profiler, *RIAO 2000*, France.

Santini M. (2004), *State-of-the-art on Automatic Genre Identification*, Tech. Report ITRI-04-03, 2004, Brighton University, UK.

Santini M. (2005a), Genres In Formation? An Exploratory Study of Web Pages using Cluster Analysis, *Proc. of the CLUK 05*, UK.

Santini M. (2005b), *Linguistic Facets for Genre and Text Type Identification: A Description of Linguistically-Motivated Features*, Tech. Report, ITRI-05-02, Brighton University, UK.

Shepherd M., Watters C. (1998), The Evolution of Cybergenre, *Proc. of the 31 Hawaii Intern. Conf. on System Sciences*, USA.

Shepherd M.,Watters C. (1999), The Functionality Attribute of Cybergenres, *Proc. of the 32 Hawaii Intern. Conf. on System Sciences*, USA.

Tapanainen P., Järvinen T. (1997), A non-projective dependency parser, *Proc. of the 5 Conf. on Applied Natural Language Processing*, USA.